Advancing structural analysis with computational methods development

Thesis by

Huanghao Mai

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy in Biochemistry and Molecular Biophysics

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended July 1st, 2024

Huanghao Mai
ORCID: 0000-0003-2278-0768

ACKNOWLEDGEMENTS

Given an unusual journey of grad school study, the possibility that I am completing on a five-year timeline would not exist without tremendous support. Nurtured by "meals from a hundred households" (a Chinese idiom), I present a long, but by no means comprehensive, acknowledgment.

I would like to first thank my advisors and thesis committee for their guidance and support. For the first half of my study, I worked with Tom Miller. Training in the Miller Group helped me learn effective scientific communication and methods development as a theorist. In the second half of my study, I worked with Hosea Nelson. Hosea has helped me mature as a scientist through his interdisciplinary perspectives, rigorous scientific standards, and career mentorship. Throughout my grad school, Bil Clemons has played several important roles, including being my committee chair, the BMB option representative, and my mentor. Bil gave me a welcoming home when Tom left and has continued rooting for me when my scientific interest shifted. I have also had inspiring discussions with Steve Mayo and Doug Rees and received their helpful advice on various projects.

Other professors have also helped my academic growth. I had an extended rotation with Grant Jensen during COVID which led to the work in Chapter 2. His BMB173 course introduced me to electron microscopy. Shu-ou Shan was a supportive member on my candidacy committee, and I appreciated her BMB178 teaching. Dave Van Valen provided helpful research and career advice. I enjoyed discussions with visiting/collaborating professors: Jon Schlebach, Clémence Corminboeuf, Michele Ceriotti, Jose Rodriguez, and Alison Narayan.

I express my gratitude to members of all labs I have been part of. Both the structural and synthesis sides of Nelson Group have been supportive as I complete the final chapter of my graduate study. Jess Burch and Lygia de Moraes went out of their ways to train me and mentored my project as described in Chapter 3. I enjoyed working with Isabel Hernandez Rodriguez and Dmitry Eremin on other computational work. Lee Joon Kim and David Delgadillo helped with my onboarding. I thank everyone for putting up with my perplexing equations in subgroup and group meetings, and I am lucky to have made many friends here. The former Miller Group has been an invaluable resource throughout my grad school. Matthew Zimmer guided me through every single step that led to my first first-author work in Chapter 4. Scientific,

professional, and personal mentorship from him remains significant even after his graduation. Continued support from Tomislav Begušić and Xuecheng Tao has also been crucial during my key transitions scientifically and professionally. I received helpful feedback from Dan Jacobson and Sherry Cheng. My "bio buds" Vignesh Bhethanabotla and Marta Gonzalvo Ulla are helpful peer support. Shyam Saladi, Ailiena Maggiolo, Victor Garcia-Ruiz, Karen Orta, Alex Barlow, Michelle Fry, and other Clemons-Rees Lab members have taught me wet lab skills during my rotation and also been a great source of help in chaotic times of my grad school.

I benefited from many other people and resources at Caltech. Chris Balzer, Uriah Israel, Linqing Peng, Verena Neufeld, and other theorist friends offered miscellaneous technical and professional help. Gio Pinton Tomaleri, Taylor Stevens, Masami Hazu, Przemek Dutka, Tom Röschinger, and other experimental friends in BMB and Biology have been a great support network. Through the Big/Little Sib program in CCE, I met Professor Susan Sharfstein who helped me navigate the difficult times. I learned about startups through OTTCP internship, and my supervisor Jay Chiang offered helpful career advice. CCE admins Annette Luymes, Rebecca Fox, Ann Mao, Courtney Oaida, and Priscilla Boon, Laura Kim at ISP, and other Caltech staff have made my academic life easier.

Stanford alumni/affiliates have also helped me in various ways. I would not be at Caltech without mentoring from Evan Feinberg and Brooke Husic. Ariana Peck has supported my professional and personal growth and mentored both projects in Chapters 2 and 3. She taught me crystallography and introduced me to methods development. I am very fortunate to have learned ML, data analysis, and many important research and professional skills from Kevin Dalton. I had a great internship at SLAC, and I am excited to work with my supervisor Frédéric Poitevin again. I thank David Mobley, Michael Shirts, and John Chodera for scientific discussions.

Finally, some personal support has been essential to my well-being. The daunting task of writing a thesis could not have been completed without the help of my partner Vignesh. For the past 5 years, Linqing has been a close friend, a running buddy, and my feminist support. Sheetal Ramsurrun has been there for me since undergrad and helped me through all difficult times. Zhuyun Zhuang, with whom I have had 15 years of friendship and shared the same household in grad school, has offered significant caring. Lastly, I thank my family for their unconditional love and support of my esoteric pursuits in a foreign country.

ABSTRACT

In this thesis, a set of computational methods is developed to extend structural techniques beyond their conventional practice. First, we build in silico simulations and image processing protocols to design a new data acquisition workflow in cryo-electron tomography. This enables in situ visualization of macromolecular complexes at sub-nanometer resolution in a micron-scale field of view. Then, we demonstrate the applicability of a novel machine-learning algorithm in processing small molecule electron diffraction data for the first time. For most molecules tested, the correct *ab initio* structures can be obtained without the common practice of manual dataset curation. Finally, molecular dynamics simulations using crystallographic structures of protein and drug molecule complexes are performed to investigate the fundamental principles of a ternary binding property. A minimal forcefield with multi-scale coarse-graining enables alchemical free energy calculations at an unconventional size of perturbation while providing physical insight into the role of the drug linker and protein shapes.

PUBLISHED CONTENT AND CONTRIBUTIONS

(1) Mai, H.; Zimmer, M. H.; Miller, T. F. Exploring PROTAC Cooperativity with Coarse-Grained Alchemical Methods. *The Journal of Physical Chemistry B* **2023**, *127*, 446–455. DOI: 10.1021/acs.jpcb.2c05795.
Contributions: Mai developed the simulation, performed data analysis and figure preparation, and wrote the manuscript.

(2) Peck, A.; Carter, S. D.; Mai, H.; Chen, S.; Burt, A.; Jensen, G. J. Montage Electron Tomography of Vitrified Specimens. *Journal of Structural Biology* **2022**, *214*, 107860. DOI: 10.1016/j.jsb.2022.107860.
Contributions: Mai participated in the methods development, performed simulations and data analysis, and participated in the preparation of figures and the manuscript.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

The 3-dimensional (3D) structure of molecules is important for understanding their biological functions, probing their relationship to other molecules, and advancing the rational design and optimization of molecular properties for therapeutic purposes. Biologically relevant molecules span a wide range of sizes, from small molecules on the scale of Ångstroms to macromolecular complexes approaching tens to hundreds of nanometers. While structural techniques targeting different spatial resolutions have been developed, emerging biological and chemical problems constantly push them to their limit, necessitating further methods development, among which include computational methods that compose an integral part of structural studies.

Computational methods have proven useful for structural studies at all stages of the workflow. Data analysis and modeling algorithms are essential to the reconstruction of an accurate and high-resolution 3D structure from raw experimental signals. After obtaining a structural model, molecular dynamics simulations and analysis can be done to calculate structure-based properties of the molecules of interest. Computer simulations can also be developed *ad hoc* to predict various experimental outcomes in a cheap and accessible way, assisting the optimization of experimental design.

In this thesis, three projects focused on computational methods development in electron microscopy (EM) and molecular dynamics (MD) are presented. To motivate the readers for our work, in this chapter, we provide the theoretical background of each project, including the mathematical formulation of the structural techniques and an overview of the data analysis algorithms involved.

1.1 Large complexes visualized in larger context

Cryo-electron tomography (cryo-ET) is a powerful approach in cryo-electron microscopy (cryo-EM) that enables the visualization of macromolecules at relatively high resolution without the need for crystallization in an X-ray diffraction experiments or purification in the single particle analysis (SPA) approach in cryo-EM. A thinly sliced specimen is flash-frozen to maximally preserve the native state of molecules within the cellular environment and reduce radiation damage during data acquisition. Then, it is placed in a transmission electron microscope (TEM) and

tilted along an axis to acquire a series of 2D projection images. An ideal projection of the specimen with electrostatic potential V at orientation $\boldsymbol{\phi}$ can be mathematically described as,

$$\varphi(x, y) = \int_z V\left(\boldsymbol{\phi} \cdot [x, y, z]^T\right) dz \tag{1.1}$$

where z-axis is parallel to the beam. A tilt series can be reconstructed into a 3D tomogram using the Fourier-slice theorem [1]:

$$\mathcal{F}[\varphi(x, y)] = \mathcal{F}[V]|_{\boldsymbol{\phi}} \tag{1.2}$$

which states that the Fourier transform of the 2D projection is a central slice of the Fourier transform of the electrostatic potential in 3D at the same orientation. For molecules of interest, the 3D structural features can be further sharpened using subtomogram averaging.

The biological context captured by cryo-ET, however, is physically limited by the magnification necessary to capture high-resolution details of molecular features. A straightforward solution would be to perform a montage data collection scheme, combining magnified images — referred as tiles — into a larger image at each tilt angle, with some overlap between adjacent tiles necessary to stitch them together. However, biological specimens are extremely sensitive to radiation damage from the electron beam even under cryogenic conditions. The overlapping regions between adjacent images receive double if not more doses in the naive way of montage, losing critical structural information. Moreover, when stitching experimental images together, additional data processing is needed to account for Fresnel fringes, uneven radial illumination, and imperfect beam shifts.

Hence, Chapter 2 investigates the optimal montage data collection scheme of cryo-ET and showcases the experimental outcome. In 2D, the problem of minimizing the overlapping regions using circular beams is mathematically solved by using a hexagonal layout [2]. In 3D, however, the location and volume of the overlapping regions also change as a function of the tilt angle of the sample. Therefore, a TEM simulator is developed to iteratively test and compare the distribution of radiation doses received by samples under different tiling and tilting strategies. The optimal strategy is implemented for experimental validation through SerialEM [3], a software that interfaces with the TEM to support programmed and automated data acquisition. A protocol for additional image processing is developed to address challenges in stitching overlapping images together. In addition to the visualization of diverse cellular features with smooth transitions at the overlap of tiles, the quality

of the stitched montage images is also assessed quantitatively by estimating the contrast transfer function (CTF).

Raw images obtained from a TEM in the real world, unfortunately, are corrupted 2D projections different from what eq. 1.1 describes. This is because the information content is transferred to the image in a sinusoidal fashion across the spatial resolution k, which is mathematically described as the CTF [4]:

$$\mathrm{CTF}(k) = \sin\left[\pi\Delta f\lambda k^2 + \frac{\pi C_s\lambda^3 k^4}{2}\right]. \tag{1.3}$$

CTF depends on the defocus Δf and the spherical aberration C_s, and is further modified by an envelope function that dampens the overall amplitude of information toward high resolution due to other imperfections of the microscope such as beam incoherence and chromatic aberration. Defocus is necessary to improve contrast in the images for alignment and particle picking, but should be carefully controlled to reduce the compromise of high-resolution information. A naive implementation of montage data collection at constant defocus would lead to a wide range of effective defocus among individual tiles due to the much larger field of view. Thus, the defocus of each tile should be adjusted based on its actual z-height in the TEM. Despite this adjustment, microscopes are not perfect and some errors are inevitable. The effective Δf is predicted by CTF estimation on collected data [5] before performing CTF correction in data processing.

The effects of CTF can be visualized as Thon rings [6] in the power spectrum generated from Fourier analysis of the image. Rings observable towards the edge of the power spectrum indicate the high-resolution limit of the data. Higher resolution Thon rings are expected at low tilt angles because of sample thinness and dose-symmetric tilting [7]. Dose-symmetric tilting prioritizes data quality at low angles, as data are collected in the order of $0°$, $\pm\alpha°$, $\pm 2\alpha°, \ldots$, where $\alpha°$ is the tilt angle increment, such that the effects from radiation damage over time are accumulated at data collected at high tilt angles.

In Chapter 2, the detection of the Thon rings limit is used to assess the quality of images before and after stitching. Before stitching, the individual tiles present Thon rings detectable at 10 to 20 Å. After stitching overlapping tiles into a larger image, Thon rings better than 10 Å can be detected at the $0°$ tilt angle for two of the three reconstructed tomograms, and Thon rings better than 15 Å can be detected at low tilt angles as expected. While the inevitable extra doses at the overlapping regions of tiles were considered too damaging for montage cryo-ET, our optimized montage

scheme and stitching protocol, evidenced by CTF analysis, successfully preserves information at a resolution typical of regular cryo-ET tomograms.

1.2 Small molecules elucidated from smaller crystals

3D electron diffraction (3D ED), also known as microcrystal electron diffraction (microED), is another emerging approach in EM that proves particularly useful for characterizing small molecules. Because electrons interact with matter more strongly than X-rays, structures can be determined *ab initio* from micro- and even nano-sized crystals without arduous efforts to obtain a large and high-quality crystal as needed by the conventional single-crystal X-ray diffraction (SCXRD) experiments. After depositing the sample on a TEM grid, diffraction images are recorded as each crystal is continuously rotated. Diffraction spots are formed at the back focal plane of the lens of a TEM by focusing plane waves scattered by the crystal. Under the kinematic approximation, each scattered plane wave is linearly related to the structure factor $\boldsymbol{F}(h,k,l)$, which is a Fourier component of the electrostatic potential of the unit cell

$$V(x,y,z) \propto \sum_{h,k,l} \boldsymbol{F}(h,k,l) \exp\left[-2\pi i\left(hx+ky+lz\right)\right]. \tag{1.4}$$

$\boldsymbol{F}(h,k,l)$ is a complex number that can be expressed in terms of its amplitude $F(h,k,l)$ and its phase $\exp\left[i\phi(h,k,l)\right]$. At image formation, only the amplitudes are recorded in the intensities of the spots,

$$I \propto \boldsymbol{F}\boldsymbol{F}^{*} = F^{2}. \tag{1.5}$$

The loss of information on the phase is known as the phase problem in crystallography. Various techniques have been developed to solve the phase problem, especially in the context of X-ray diffraction (XRD). Programs originally developed for XRD can be used to estimate the amplitudes and phases of structure factors from 3D ED data to reconstruct a 3D structure, as the phase problem is essentially the same in both methods.

Despite the availability of established software, 3D ED data processing can be more challenging than SCXRD. In particular, merging data from multiple crystals is routinely required. This is a consequence of hardware limitations in most TEMs and poor data quality from inelastic and dynamical scattering as well as radiation damage. The direct outcome of merging is estimated values of the amplitude F. An accurate estimate is important for downstream data processing — *i.e.*, solving

the phase problem *ab initio* and determining the final molecular structure after refinement. Multi-crystal merging can demand significant human input on dataset curation when a large number of datasets are collected. Chapter 3 investigates whether a new merging algorithm based on variational inference (VI) yields superior and more efficient outcomes for small molecule 3D ED data.

Conventional methods estimate $F = \sqrt{\bar{I}}$ by averaging redundant measurements of I weighted by their uncertainty σ_I after performing a scaling that corrects for systematic errors. This approach can be interpreted as the maximum likelihood estimation (MLE), where using normal distribution as the error model, the likelihood of N measurements

$$p_\theta\left(I\right) = \prod_i^N p_\theta\left(I_i\right) = \prod_i^N \frac{1}{\sigma_{I_i}\sqrt{2\pi}} \exp\left[-\frac{\left(I_i-\theta\right)^2}{2\sigma_{I_i}^2}\right] \tag{1.6}$$

is maximized when the model parameter θ is the weighted average

$$\theta = \bar{I} = \arg\max p_\theta\left(I\right) = \frac{\sum_i I_i/\sigma_{I_i}^2}{\sum_i 1/\sigma_{I_i}^2}. \tag{1.7}$$

VI, in contrast, jointly performs scaling and merging by considering F as a latent variable that generates the observed I. With the introduction of latent variables, the likelihood term generally becomes intractable:

$$p_\theta\left(I\right) = \int p_\theta\left(I,F\right) dF = \int p_\theta\left(I|F\right) p\left(F\right) dF = \mathbb{E}_{p(F)}\left[p_\theta\left(I|F\right)\right]. \tag{1.8}$$

A naive estimation by Monte Carlo sampling from the prior distribution $p\left(F\right)$ converges slowly. A statistical trick that can be useful here is importance sampling, which samples from a surrogate distribution $q\left(F\right)$ and corrects the sampling bias:

$$\begin{aligned}\int p_\theta\left(I|F\right) p\left(F\right) dF &= \int p_\theta\left(I|F\right) \frac{p\left(F\right)}{q\left(F\right)} q\left(F\right) dF \\ &= \mathbb{E}_{q(F)}\left[p_\theta\left(I|F\right) \frac{p\left(F\right)}{q\left(F\right)}\right].\end{aligned} \tag{1.9}$$

Using Jensen's inequality,

$$\begin{aligned}\log p_\theta\left(I\right) &= \log \mathbb{E}_{q(F)}\left[p_\theta\left(I|F\right) \frac{p\left(F\right)}{q\left(F\right)}\right] \\ &\geq \mathbb{E}_{q(F)}\left[\log\left[p_\theta\left(I|F\right) \frac{p\left(F\right)}{q\left(F\right)}\right]\right] \\ &= \mathbb{E}_{q(F)}\left[\log p_\theta\left(I|F\right)\right] - \mathbb{E}_{q(F)}\left[\log \frac{q\left(F\right)}{p\left(F\right)}\right] \\ &= \mathbb{E}_{q(F)}\left[\log p_\theta\left(I|F\right)\right] - D_{\mathrm{KL}}\left[q\left(F\right) \| p\left(F\right)\right]\end{aligned} \tag{1.10}$$

where the last line of eq.1.10 is also known as the **e**vidence-**l**ower **bo**und (ELBO). In other words, in VI, maximizing ELBO is guaranteed to maximize the likelihood. When $q(F) = p(F|I)$, $p_\theta(I)$ is exactly recovered using Bayes' theorem. In fact, ELBO can be alternatively derived from minimizing $D_{\mathrm{KL}}[q(F) \| p(F|I)]$. This motivates the interpretation of $q(F)$ as a good approximation of the intractable true posterior $p_\theta(F|I)$. Thus, unlike MLE or maximum a posteriori methods that yield a point estimation of F, VI enables characterizing the distribution of F with uncertainty information.

Another intuitive interpretation of maximizing ELBO as the optimization objective of VI is that the former term encourages fitting variable F to observed data I by maximizing the log-likelihood, whereas the latter Kullback–Leibler (KL) term penalizes overfitting by constraining model $q(F)$ to not deviate too far from the prior distribution $p(F)$.

The choice of the error model for the log-likelihood term is not limited to a normal distribution. An implementation of VI [8] shows that the Student's t-distribution performs better than the normal distribution on merging XRD data in a few macromolecular cases. This is likely because diffraction measurements are very noisy, and the Student's t-distribution is more tolerant of outliers. Thus, we use the same error model in Chapter 3 to test the VI model [8] on small molecule 3D ED data.

The choice of the prior distribution can be adapted to specific experimental design [9]. In crystallography, a general and reasonable prior distribution as implemented in Dalton *et al.* [8] is the Wilson distribution [10], which is the intensity distribution if atoms are uniformly distributed within the unit cell. A detailed derivation is available in Srinivasan and Parthasarathy [11], and key steps are summarized below.

In a crystal without symmetry where there are N atoms in a unit cell and the scattering factor of the j^{th} atom is f_j, $\boldsymbol{F}$ can be written as

$$\boldsymbol{F}(\boldsymbol{s}) = \sum_{j=1}^{N} f_j(\boldsymbol{s}) \exp[i2\pi \boldsymbol{s} \cdot \boldsymbol{r}] = A + iB \tag{1.11}$$

$$A = \sum_j f_j \cos(2\pi \boldsymbol{s} \cdot \boldsymbol{r}) \qquad B = \sum_j f_j \sin(2\pi \boldsymbol{s} \cdot \boldsymbol{r}) \tag{1.12}$$

where $\boldsymbol{s} \equiv (h, k, l)$ and $\boldsymbol{r} \equiv (x, y, z)$. In the limit of large N, by the central limit theorem, A and B both follow normal distributions. Furthermore, assuming a uniform distribution of the atoms in the unit cell and setting the center of mass as

the origin such that the mean and variance are 0 and $\frac{1}{2}\sum_j f_j^2$ for both A and B. The joint probability of A and B is:

$$p\left(A,B\right) = \frac{1}{\pi \sum_j f_j^2} \exp\left[-\frac{A^2+B^2}{\sum_j f_j^2}\right]. \tag{1.13}$$

Since $F^2 = A^2 + B^2$, a polar coordinate transformation can be performed where $A = F\cos\left(\alpha\right)$ and $B = F\sin\left(\alpha\right)$, and

$$p\left(F,\alpha\right) = \frac{F}{\pi \sum_j f_j^2} \exp\left[-\frac{F^2}{\sum_j f_j^2}\right]. \tag{1.14}$$

Finally, $p\left(F\right)$ is obtained by marginalizing over α.

With centrosymmetric space group, $\boldsymbol{F} \in (-\infty,\infty)$, $F \in [0,\infty)$, $B = 0$, and the variance of A becomes $\sum_j f_j^2$. These conditions simplify the derivation such that

$$p\left(\boldsymbol{F}\right) = p\left(A\right) = \frac{1}{\sqrt{2\pi \sum_j f_j^2}} \exp\left[-\frac{A^2}{2\sum_j f_j^2}\right]. \tag{1.15}$$

Together, we have the Wilson prior

$$p\left(F\right) = \begin{cases} \dfrac{2F}{\sum_j f_j^2} \exp\left[-\dfrac{F^2}{\sum_j f_j^2}\right] & \text{if } (h,k,l) \text{ acentric,} \\ \sqrt{\dfrac{2}{\pi \sum_j f_j^2}} \exp\left[-\dfrac{F^2}{2\sum_j f_j^2}\right] & \text{if } (h,k,l) \text{ centric.} \end{cases} \tag{1.16}$$

The cumulative distributions of normalized intensity for centric and acentric reflections can be derived [12] from the Wilson distributions (eq. 1.16) and are tabulated in standard data processing programs to assess twinning [13] and other non-ideal conditions [11]. Small molecule electron diffraction data do not always perfectly obey the Wilson distributions, as dynamical scattering, background noises, and a small number of atoms in the unit cell can all affect the intensity statistics. Meanwhile, VI using the Wilson prior as implemented in Dalton *et al.* [8] has only been tested on macromolecular XRD experiments.

In Chapter 3, we show for the first time that scaling and merging using VI robustly generalizes to small molecule electron diffraction data. Moreover, we look into the impact of manual dataset curation and explore an extension to the model in Dalton *et al.* [8] using machine learning principles to automate dataset curation. In our tested cases, dataset curation — whether manual or automated, is less effective

than conventionally thought. This suggests that the VI algorithm can efficiently leverage information from all available datasets and reduce human bias in data processing. Finally, limitations and practical challenges are discussed for experimental practitioners.

1.3 Protein-drug complexes simulated at multi-scale

Structure determination is often motivated by the need to understand molecular functions carried out by interactions with other molecules, referred to as binding events. Structures of the bound and unbound states of the molecules of interest unveil key information such as the binding sites and binding modes. However, obtaining complete structural information at high resolution is not always possible, especially for complexes exhibiting substantial conformational flexibility that makes sample preparation challenging and blurs the measured signals. Moreover, a static structural model is insufficient to characterize thermodynamic averages such as the binding free energies. A quantitative measurement of the binding energies requires additional techniques such as surface plasmon resonance (SPR) and isothermal titration calorimetry (ITC).

The development of proteolysis targeting chimera (PROTAC) as a novel drug modality is a concrete example where analysis of the binding properties requires structural insights and complementary techniques. A PROTAC molecule has two warheads for simultaneously binding the target protein and recruiting an E3 ligase that induces the degradation of the target protein. An interesting property of the target-PROTAC-E3 complex is that the presence of E3 ligase affects the binding affinity between the PROTAC and the target protein and vice versa. Mathematically, this phenomenon is summarized by the binding cooperativity $\alpha = K_{\mathrm{D}}^{\mathrm{binary}} / K_{\mathrm{D}}^{\mathrm{ternary}}$. With positive cooperativity ($\alpha > 1$), the E3 ligase facilitates the binding between the PROTAC and the target protein and lowers the binding free energy. However, rational optimization of the cooperativity is hindered by a poor understanding of the fundamental principles due to the scarcity and low resolution of experimental data. Structures of the ternary complexes can be difficult to obtain compared to the binary complexes. In addition, few studies measure PROTAC binding energies against systematic perturbation of structural features.

Free energy calculation methods complement experimental techniques and can be insightful for cases such as the ternary complexes of PROTAC where experimental measurements are challenging. Given the binding modes from binary complex

structures, computer simulations can be leveraged to calculate the cooperativity of the ternary complexes. Within each thermodynamic state, different configurations of the complexes are sampled by MD simulations to predict the binding energies.

The Helmholtz free energy difference (ΔA) between the start state s and the final state f is:

$$\Delta A_{f,s} = -\beta^{-1}\left(\ln Q_f - \ln Q_s\right) = -\beta^{-1}\ln\frac{Q_f}{Q_s} \tag{1.17}$$

where $\beta^{-1} = k_B T$, k_B is the Boltzmann constant, T is the temperature, and

$$Q = \int_{\Gamma} \exp\left[-\beta U\left(\mathbf{q}\right)\right] d\mathbf{q} \tag{1.18}$$

is the configurational partition function over phase space Γ. The probability of a particular configuration $\mathbf{q}$ of the molecule or molecular complex is its Boltzmann weight normalized by the partition function:

$$p\left(\mathbf{q}\right) = \frac{\exp\left[-\beta U\left(\mathbf{q}\right)\right]}{Q}. \tag{1.19}$$

A naive way to calculate binding free energy is to directly observe the binding and unbinding events in an MD simulation. However, atomistic MD simulations of the ternary complex at a biologically relevant timescale are infeasible due to the high computational cost.

In principle, using importance sampling, the free energy difference can be exactly obtained by importance sampling and only simulating configurations in the start state:

$$\begin{aligned}
\frac{Q_f}{Q_s} &= \frac{\int \exp\left[-\beta U_f\left(\mathbf{q}\right)\right] \exp\left[\beta U_s\left(\mathbf{q}\right)\right] \exp\left[-\beta U_s\left(\mathbf{q}\right)\right] d\mathbf{q}}{Q_s} \\
&= \int \exp\left[-\beta \Delta U_{f,s}\left(\mathbf{q}\right)\right] p_s\left(\mathbf{q}\right) d\mathbf{q} \\
&= \left\langle \exp\left[-\beta \Delta U_{f,s}\left(\mathbf{q}\right)\right]\right\rangle_s \\
&= \frac{1}{N_s}\sum_{i}^{N_s} \exp\left[-\beta \Delta U_{f,s}\left(\mathbf{q}_i\right)\right].
\end{aligned} \tag{1.20}$$

This is known as exponential averaging or Zwanzig relationship [14]. However, in practice, this method is challenging due to the stringent requirement on substantial phase space overlap between Γ_s and Γ_f. Practically, this means that the convergence of calculations requires prohibitively long simulation time to sample enough configurations from simulations using U_s.

One solution is to perform enhanced sampling along the binding pathway, but this information is often unavailable *a priori*. An alternative approach is to perform free energy perturbation through alchemical transformations. Readers are referred to Mey *et al.* [15], Pohorille *et al.* [16], and Klimovich *et al.* [17] for additional reviews and guidelines.

In alchemical free energy calculations, intermediate states are generically defined by a coupling parameter λ for the potentials U, such that $\lambda = 0$ corresponds to the start state s and $\lambda = 1$ corresponds to the end state f. For comparative studies such as the calculation of binding cooperativity, alchemical transformation defines the relative presence of two molecular species to be compared at both bound and unbound states.

Thermodynamic integration (TI) can be used to approximate the free energy difference by performing numerical integration along λ:

$$\begin{aligned}\frac{dA_\lambda}{d\lambda} &= -\beta^{-1}\frac{d\ln\int\exp\left[-\beta U\left(\lambda,\mathbf{q}\right)\right]d\mathbf{q}}{d\lambda}\\ &= -\beta^{-1}\frac{-\beta\int\exp\left[-\beta U\left(\lambda,\mathbf{q}\right)\right]\frac{dU(\lambda,\mathbf{q})}{d\lambda}}{Q_\lambda}d\mathbf{q}\\ &= \left\langle\frac{dU\left(\lambda,\mathbf{q}\right)}{d\lambda}\right\rangle_\lambda\end{aligned} \tag{1.21}$$

$$\Delta A_{f,s} = \int_0^1\frac{dA_\lambda}{d\lambda}d\lambda = \int_0^1\left\langle\frac{dU\left(\lambda,\mathbf{q}\right)}{d\lambda}\right\rangle_\lambda d\lambda. \tag{1.22}$$

Another way to improve convergence along the idea of importance sampling is through the use of the multistate Bennett acceptance ratio (MBAR) method [18] that reweights the samples by the mixture distributions of all states [19]. The reweighted probability of a configuration $\mathbf{q}$ is:

$$p\left(\mathbf{q}\right) = \sum_\lambda\frac{N_\lambda}{N}p_\lambda\left(\mathbf{q}\right) = \frac{1}{N}\sum_\lambda N_\lambda\frac{\exp\left[-\beta U\left(\lambda,\mathbf{q}\right)\right]}{Q_\lambda} \tag{1.23}$$

where $N = \sum_\lambda N_\lambda$ is the total number of samples drawn from all states.

Using importance sampling,

$$\begin{aligned} Q_\lambda &= \int \exp\left[-\beta U\left(\lambda, \mathbf{q}\right)\right] d\mathbf{q} \\ &= \int \frac{\exp\left[-\beta U\left(\lambda, \mathbf{q}\right)\right]}{p\left(\mathbf{q}\right)} p\left(\mathbf{q}\right) d\mathbf{q} \\ &= \left\langle \frac{\exp\left[-\beta U\left(\lambda, \mathbf{q}\right)\right]}{p\left(\mathbf{q}\right)} \right\rangle \\ &= \frac{1}{N} \sum_i^N \frac{\exp\left[-\beta U\left(\lambda, \mathbf{q}\right)\right]}{\frac{1}{N}\sum_\lambda N_\lambda \frac{\exp\left[-\beta U(\lambda, \mathbf{q})\right]}{Q_\lambda}} \\ &= \sum_i^N \frac{\exp\left[-\beta U\left(\lambda, \mathbf{q}\right)\right]}{\sum_\lambda N_\lambda \frac{\exp\left[-\beta U(\lambda, \mathbf{q})\right]}{Q_\lambda}}. \end{aligned} \tag{1.24}$$

This gives a set of equations that yields solutions up to a multiplicative constant. Since only the difference of free energies is of interest (eq. 1.17), one of the Q_λ is fixed at an arbitrary constant such as 1 to solve the rest.

Without the introduction of the intermediate states, MBAR reduces to the Bennett acceptance ratio (BAR) approach [20], which was originally derived by constructing a scaling function $\alpha\left(\mathbf{q}\right)$ that minimizes the variance of the estimated free energies from

$$\begin{aligned} \frac{Q_f}{Q_s} &= \frac{1/Q_f \int \alpha\left(\mathbf{q}\right) \exp\left[-\beta U_f\left(\mathbf{q}\right)\right] \exp\left[-\beta U_s\left(\mathbf{q}\right)\right]}{1/Q_s \int \alpha\left(\mathbf{q}\right) \exp\left[-\beta U_f\left(\mathbf{q}\right)\right] \exp\left[-\beta U_s\left(\mathbf{q}\right)\right]} \\ &= \frac{\langle \alpha\left(\mathbf{q}\right) \exp\left[-\beta U_s\left(\mathbf{q}\right)\right] \rangle_f}{\langle \alpha\left(\mathbf{q}\right) \exp\left[-\beta U_f\left(\mathbf{q}\right)\right] \rangle_s}. \end{aligned} \tag{1.25}$$

A straightforward application of these free energy calculation methods is challenging because perturbation of less than 10 heavy atoms is typically required for convergence. However, calculating PROTAC binding cooperativity requires perturbation at the scale of a protein domain. While coarse-graining (CG) can be used to reduce the effective system size and smoothen the energy landscape, the large discrepancy between the size of the proteins and the size of a PROTAC precludes an aggressive and uniform approach that loses the resolution needed for structural and chemical insights.

Therefore, Chapter 4 explores a CG approach at mixed scales and rigorously analyzes the convergence properties of the alchemical free energy calculations for PROTAC binding. A minimal forcefield U is constructed to characterize the role of the PROTAC linker length and protein shapes in cooperativity as a result of

the configurational entropy. Future inclusion of more sequence-specific and PROTAC configurational features in the CG forcefield, which may improve the method towards more quantitative predictions, are discussed.

Chapter 2

MONTAGE ELECTRON TOMOGRAPHY OF VITRIFIED SPECIMENS

Adapted from

(1) Peck, A.; Carter, S. D.; Mai, H.; Chen, S.; Burt, A.; Jensen, G. J. Montage Electron Tomography of Vitrified Specimens. *Journal of Structural Biology* **2022**, *214*, 107860. DOI: 10.1016/j.jsb.2022.107860.

Abstract

Cryo-electron tomography provides detailed views of macromolecules *in situ*. However, imaging a large field of view to provide more cellular context requires reducing magnification during data collection, which in turn restricts the resolution. To circumvent this trade-off between field of view and resolution, we have developed a montage data collection scheme that uniformly distributes the dose throughout the specimen. In this approach, sets of slightly overlapping circular tiles are collected at high magnification and stitched to form a composite projection image at each tilt angle. These montage tilt-series are then reconstructed into massive tomograms with a small pixel size but a large field of view. For proof-of-principle, we applied this method to the thin edge of HeLa cells. Thon rings to better than 10 Å were detected in the montaged tilt-series, and diverse cellular features were observed in the resulting tomograms. These results indicate that the additional dose required by this technique is not prohibitive to performing structural analysis to intermediate resolution across a large field of view. We anticipate that montage tomography will prove particularly useful for lamellae, increase the likelihood of imaging rare cellular events, and facilitate visual proteomics.

2.1 Introduction

Cryo-electron tomography (cryo-ET) is a powerful tool for studying macromolecular structures in the near-native context of frozen-hydrated cells, unperturbed by stains or fixatives [21, 22]. In this cryo-electron microscopy (cryo-EM) technique, a series of projection images is recorded as a vitrified specimen is tilted in an electron microscope. The resulting tilt-series is reconstructed into a tomogram, or volumetric map of the specimen's electrostatic potential. Cryo-ET has revealed important details of cellular ultrastructure, and subtomogram averaging algorithms enable determining the structures of macromolecular complexes at a resolution of 0.4-4 nm [21]. Recent technical advances have extended the high-resolution potential of this technique, yielding subtomogram averages of purified HIV-1 Gag particles to 3.1 Å and *in situ* ribosomes to 3.7 Å [23, 24]. However, data must be collected at high magnification to retain this high-resolution signal. The trade-off is a smaller field of view, limiting the region that can be imaged to a tiny fraction of a cell.

In principle, montage tomography could permit imaging a large field of view without sacrificing high-resolution details. For montage data collection, the beam is tiled across a specimen and the recorded images are computationally stitched together during reconstruction [25]. To date, montage tomography has only been performed on resin-embedded samples, which resist radiation damage but suffer from artifacts induced by chemical fixation [26–28]. By contrast, cryo-preserved specimens can tolerate only a limited electron dose before being destroyed [29]. Historically this dose sensitivity has been considered prohibitive to collecting montage data from vitrified samples, as portions of the sample must be exposed multiple times to facilitate stitching images during reconstruction.

Recent technical developments, however, motivate revisiting the potential of montage cryo-tomography. First, the highly stable optics of modern microscopes enable precise control over the region being exposed [30]. Such precision is critical both to prevent gaps between neighboring images and to avoid accidentally enlarging the overlap region. Second, the ability to collect data using a circular beam with fringe-free illumination allows for more efficient tiling strategies and reduces loss of information due to corruption by Fresnel fringes [31, 32]. Third, the increased sensitivity of modern detectors permits decreasing the exposure while achieving the same signal-to-noise ratio [33–36]. Fourth, focused ion beam (FIB) milling has significantly expanded the range of specimens that can be studied by cryo-ET but is highly time-consuming, motivating efforts to image as much of the lamellae as

possible [37–39]. In addition, there is growing recognition in the field that radiation damage is a progressive phenomenon [40–42]. As a result, important features of cellular biology remain observable even after receiving what was previously considered an intolerably high dose for vitrified samples. While other approaches like serial blockface scanning electron microscopy (SBF-SEM) [43], soft X-ray tomography [44, 45], and 3d focused ion beam scanning electron microscopy (FIB-SEM) [46] also permit large volume imaging of cellular ultrastructure, montage cryo-ET has the greatest potential to achieve reconstructions with subnanometer resolution.

Montage tomography of cryo-preserved specimens would further the potential of cryo-ET by increasing the likelihood of imaging transient events and providing significantly more cellular context for macromolecules of interest. Here we use simulations to optimize a montage data collection scheme and develop strategies to stitch tiles at each tilt angle into a composite projection image. We then harness modern cryo-ET algorithms to reconstruct tomograms from these montage tilt-series with a small pixel size but a large field of view. To demonstrate proof-of-principle, we applied this technique to the thin edges of HeLa cells. Thon rings to better than 15 Å were observed at both low and intermediate tilt angles in the individual tiles and detected to under 10 Å in the untilted projection images after stitching. The reconstructed tomograms spanned a 3.3 μm^2 field of view and contained diverse cellular features, including mitochondria, multilammelar vesicles, and microtubules. These results indicate that despite the additional dose required by this method, montage tomography enables capturing large fields of view for structural analysis at the intermediate resolutions typical of cryo-ET data.

2.2 Optimization of tiling strategies

Given the sensitivity of biological samples to radiation damage [40–42], the success of montage tomography depends on efficiently distributing the total exposure both at each tilt angle and across the full tilt-series. The former is readily addressed for a circular beam: the optimal strategy to pack circles in a plane uses a hexagonal tiling scheme, in which circles are centered on the vertices of a regular hexagonal grid and three neighboring circles intersect at a point (Fig. 2.1A) [2]. The question then remains how to displace these hexagonally-packed circular tiles between tilt angles to most uniformly spread the dose across the tilt-series. Applying global translations and rotations to the hexagonal array of circles between tilt angles changes which regions of the sample lie in an overlap region at each tilt angle, thereby reducing the amount of sample that receives excess dose (Fig. 2.1B-C).

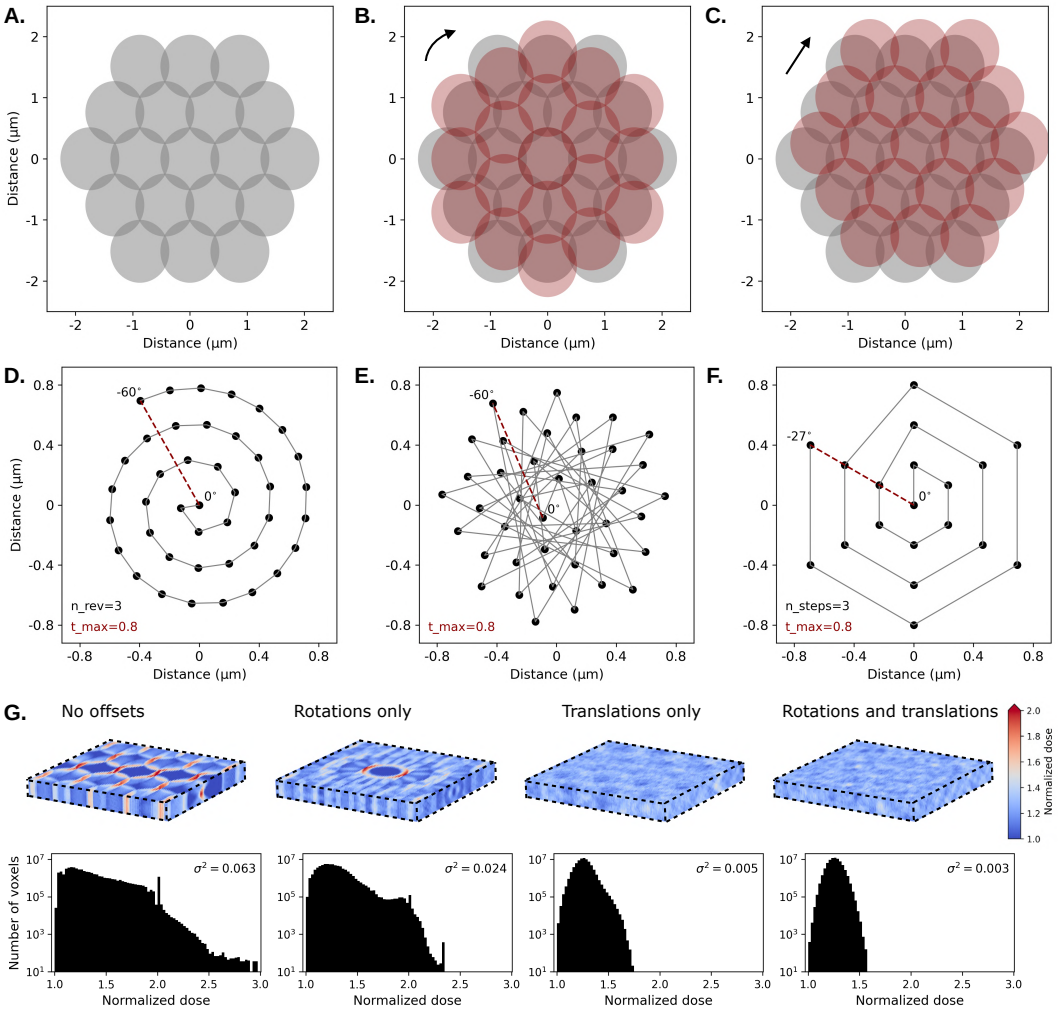

Figure 2.1: Optimization of a montage tiling strategy. (A) At each tilt angle, a hexagonally-packed set of circular tiles is imaged. Applying a global (B) rotation or (C) translation to the tiles between tilt angles changes which regions of the specimen lie in an overlap region to more uniformly spread the dose. Global translation is applied using one of three spiral patterns: (D) Archimedean, (E) sunflower, or (F) "snowflake." The positions of the central tile are indicated by black dots and spiral outwards during the tilt-series. Tunable parameters for the patterns include the maximum translation of the central tile (t_max) and the number of revolutions (n_rev or n_steps). (G) The spatial distribution of the dose is mapped on the specimen (upper). The variance (σ^2) of the dose distribution is noted at the upper right of the histograms (lower). For the "no offsets" strategy, tiles were rotated by 10° relative to the plane of the detector with no further offsets applied during the tilt-series. For the "rotations only" strategy, a 20° clockwise rotation was applied between tilt angles. For the "translations only" strategy, tiles were translated along an Archimedean spiral with three revolutions and a maximum translation of 80% of the beam radius for each tile.

Since *a priori* it is unclear which combination of offsets would most efficiently distribute dose, we used simulations to characterize hundreds of different tiling strategies. These simulations used a right-handed coordinate system with the detector oriented in the xy plane and the incoming electron beam directed along the z-axis (Fig. 2.8). A rectangular specimen with a width of 3,420 nm and depth of 400 nm was discretized into 4 nm cubic voxels and tilted about the x-axis following a standard dose-symmetric tilt-scheme [47]. At each tilt angle, a 1 μm diameter beam was used to illuminate a set of hexagonally-packed circular tiles. Even with fringe-free illumination, residual fringes were observed to affect up to 2% of the outer edge of each tile (see below). To prevent gaps in the montage after discarding this corrupted region, the overlap between tiles was increased relative to optimal hexagonal packing such that a small fraction of voxels was exposed up to three times at any tilt angle. We then computed the accumulated dose received by each voxel of the specimen under tiling strategies that differed in the translational and rotational offsets applied between tilt angles.

To systematically introduce translational offsets, we examined three basic spiral patterns: an Archimedean spiral, in which adjacent points are equidistant along the curve of the spiral (Fig. 2.1D); a sunflower or Fibonacci spiral, in which points are distributed in concentric shells of equal area, and successive points are placed in the largest angular gap between previous points [48] (Fig. 2.1E); and a "snowflake" spiral, in which points are positioned on the vertices of concentric hexagons that mirror the 6-fold symmetry of the packed circular tiles (Fig. 2.1F). The positions of all tiles were uniformly shifted between successive tilt angles to follow the path of the spiral while maintaining hexagonal packing at each tilt angle. For the snowflake spiral, the pattern is repeated starting from the center if the outermost position is reached before the final tilt angle. For all three spiral types, one of the adjustable parameters was the maximum translation permitted for each tile across the entire tilt-series (Fig. 2.9A); this was capped at a distance of one beam radius to ensure that all but the outermost tiles remained fully in the field of view. The second translational parameter was the number of revolutions or radial steps respectively for the Archimedean spiral and snowflake pattern (Fig. 2.9B). For the sunflower pattern, positions were dictated exclusively by the number of points and maximum translation. A third translational parameter was an optional scaling of the x-axis displacements by $1/\cos\alpha$, where α is the tilt angle (Fig. 2.9C). This scaling mimicked the y-axis elongation of circular tiles into ellipses at high tilt angles (Fig. 2.8), with the intent of preserving efficient circular symmetry.

In addition to translational offsets, rotational offsets were systematically introduced by varying three parameters. The first parameter was the starting angle, which dictated the initial orientation of the hexagonal array of circular tiles in the plane of the detector. The second parameter was the rotational step size, which determined the magnitude of the global rotation applied to the hexagonally-packed tiles between each tilt angle. Third, this global rotation was either applied continuously or in an alternating fashion. For the continuous scheme, the hexagonally-arranged tiles were rotated by the same amount and in the same direction between each tilt angle. For the alternating scheme, global counterclockwise and clockwise rotations of the same rotation step size were applied between successive tilt angles.

In total we simulated 576 snowflake, 546 spiral, and 286 sunflower patterns by systematically varying the parameters described above. The dose distributions received by voxels of a discretized specimen during a simulated tilt-series are compared. Each pattern was scored by the variance of the distributed dose, with superior patterns characterized by low variance (Fig. 2.1G). Across all spiral types, we found that translational offsets were more critical than rotational offsets to uniformly spread the dose throughout the specimen (Figs. 2.1G, 2.10). Although the top-ranked variant for the (Archimedean) spiral, sunflower, and snowflake patterns achieved similar scores, we found that variants of the first consistently scored well for translational offsets of the same magnitude (Fig. 2.10). All variants were also observed to perform similarly across different data collection schemes (Fig. 2.11). We thus chose the best-performing spiral variant for experimental data collection. This pattern was characterized by a maximum tile translation of 80% of the beam radius, 3 revolutions, no scaling of the x-axis translational component, a starting offset angle of 10° of the hexagonally-arranged tiles in the plane of the detector, and continuous rotations of 20° between tilt angles (Fig. 2.1G, right). These offsets reduced the variance of the distributed dose by 20-fold compared to the corresponding pattern without any offsets (σ^2=0.003 versus σ^2=0.063).

2.3 Montage data collection and processing

Montage tilt-series collection

Montage tilt-series were collected on a Titan Krios G3i (Thermo Fisher Scientific) equipped with fringe-free illumination [31], a Gatan imaging filter, and a K3 Summit direct electron detector (Gatan). Data acquisition was performed using SerialEM in electron-counting mode [3] by providing the software a custom script (see Code Availability). Each tile was acquired using a circular beam of diameter 1.08 μm,

such that the beam spanned the short edge of the detector at a pixel size of 2.65 Å. A beam-centering step was performed after collecting each tile to reduce drift without applying additional exposure. At each tilt angle, 37 tiles were acquired: one central tile, with three surrounding rings of hexagonally-packed tiles. The beam coordinates were updated using SerialEM's image shift function to follow a spiral pattern as described in Section 2.2. Once the full complement of 37 tiles was collected at a particular tilt angle, the stage was rotated to the next tilt angle following a grouped dose-symmetric tilt-scheme with a group size of 6°, 2° increments between tilt images, and a tilt-range of ±60° [49]. Applying the above data collection scheme at a constant defocus was predicted to yield a ~8 μm defocus gradient at the highest tilt angles (Fig. 2.2A, left). To avoid this spread, we adjusted the defocus with which each tile was collected based on that tile's estimated z-height in the microscope. Defocus values ranged from -5 to -11 μm for different tilt-series, and a total dose of 60-106 $e^-/Å^2$ was used. Each montage dataset took 4 hours to collect, yielding 100 GB of raw data.

Tile pre-processing

Before stitching tiles into a composite image, Fresnel fringes and non-uniform illumination must be accounted for. Fringe-free illumination (FFI) reduces but does not entirely eliminate Fresnel fringes during image formation [31, 32]. Residual fringes are particularly evident when data are collected at a defocus and magnification typical for tomography, since current FFI set-ups are optimized for single-particle cryo-EM applications instead. The number of observed fringes depends on both defocus and the signal intensity of the specimen, so removing a fixed fraction of the outer edge of each tile does not adequately eliminate the fringe-contaminated region.

Given this variability, we developed the following heuristic approach to mask residual fringes. A 2-dimensional Gaussian bandpass filter was applied to each tile, using kernel sizes of 2 and 6 nm. In the filtered tile, a high-intensity ring that spanned the Fresnel fringes was observed at the edge of the illuminated region (Fig. 2.12A). A circle was fit to this ring of pixels using least-squares optimization, and the start of the Fresnel fringes was estimated as the ring's inner radius. All pixels outside this radius were masked (Fig. 2.12B-E).

In addition to residual fringes, we observed a consistent reduction in radial intensity by ~15% between the center and edge of each tile (Fig. 2.12B). Left uncorrected,

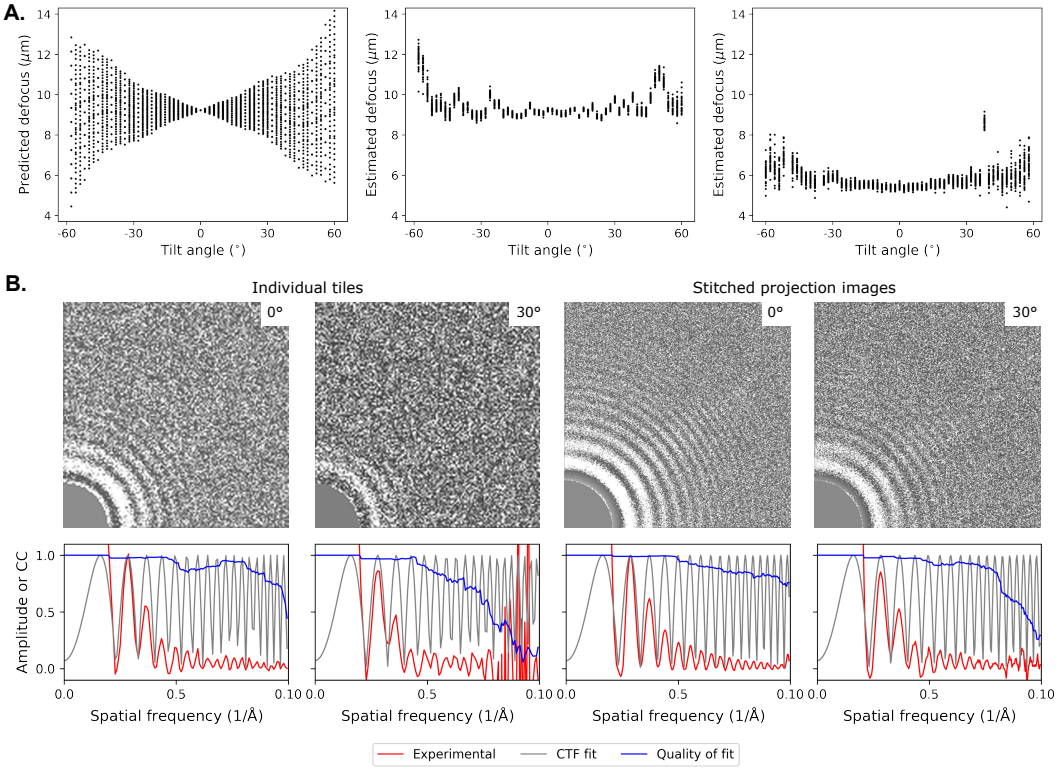

Figure 2.2: CTF estimation reveals a stable defocus throughout the tilt-series and Thon rings to better than 15 Å. (A) Defocus values were predicted based on the tiles' estimated heights in the microscope and are shown as a function of tilt angle (left). By contrast, the per-tile defocus values estimated by CTFFIND4 showed a relatively stable defocus gradient, as plotted for two representative tilt-series (middle and right). This was accomplished by performing an autofocusing step prior to collecting each tile to compensate for the predicted defocus gradient across each tilt angle. (B) The 2D experimental spectrum (upper) and rotationally-averaged 1D CTF fits (lower) are shown for representative tiles (left) or CTF-uncorrected stitched projection images (right) at the indicated tilt angle.

this would artificially depress the intensity of the overlap regions during stitching. We therefore applied a radial gain correction as follows. At each tilt angle, the tiles' radial intensity profiles were normalized to a value of 1 in the central region of the tile and merged to generate a single intensity profile for the tilt angle. The resulting radial intensity profile was median-filtered and applied as a gain reference, with linear interpolation used to compute the correction factor at each pixel (Fig. 2.12B-E).

Correcting for the contrast transfer function

Despite adjusting the focus on a per-tile basis to avoid an excessive defocus gradient across the tilt-series, some variation in defocus between tiles was expected due to microscope error. We therefore performed a contrast transfer function (CTF) correction on individual tiles. We used CTFFIND4 (version 4.1.13) to estimate each tile's defocus [5] and *ctfphaseflip* to correct for the CTF [50]. The estimated defocus values confirmed that the per-tile focus adjustment compensated for the large changes in z-height due to montage collection and yielded a relatively stable defocus throughout the tilt-series (Fig. 2.2A). The high-resolution limit of detected Thon rings for most tiles ranged from 10-20 Å, with better resolution at lower tilt angles as expected (Fig. 2.2B, left). To verify that the stitching procedure described below did not degrade resolution, CTFFIND4 was also applied to the CTF-uncorrected stitched images to assess the high-resolution limit of detectable Thon rings (Fig. 2.2B, right). In the untilted stitched images, CTFFIND4 detected Thon rings to resolutions of 8.2, 9.3, and 12.1 Å for the three montage datasets presented in this work (Fig. 2.13).

Tile registration

We developed an image registration workflow tailored for montage cryo-tomograms collected using a circular beam. Current software for processing montage data is designed for resin-embedded specimens and unsuitable for two reasons. First, resin-embedded samples are typically acquired using the full area of the detector, yielding a large and rectangular overlap region between tiles [25]. By contrast, we employ a more efficient tiling strategy of hexagonally-tiled circular beams to overcome radiation sensitivity. This scheme results in lemon-shaped overlap regions that change positions and orientations relative to the specimen at each tilt angle (Fig. 2.1B-C). Second, the algorithms used to perform automated landmark extraction and alignment rely on high-contrast features that typify resin-embedded specimens but are absent in cryo-tomograms even when collected at high defocus [51]. Although gold fiducials can be added to the sample to provide high-contrast features, a high concentration would be needed to ensure that sufficient markers are present in the overlap regions for use during tile registration. Further, these fiducials reside on the sample's surface so may experience more severe warping or doming effects than the cellular matter below.

To overcome low contrast, data collected with a pixel size of 2.65 Å were first binned to 10.6 Å. After applying a correction for uneven radial illumination, tiles were bandpass-filtered and masked to remove the unilluminated and Fresnel fringe-

corrupted regions. For bandpass-filtering, we found that kernel sizes of roughly 6 and 19 nm enhanced features such as gold beads, membranes, and grid hole boundaries that serve as useful landmarks for image registration. These features were further selected for by thresholding; specifically, only pixels with intensities in the bottom 15th percentile for each tile (belonging to high contrast features) were retained.

Despite the stable optics of modern microscopes, some drift is expected during data collection. To refine the tile positions from the beam coordinates supplied to the microscope, we computed the translational shifts that maximized the normalized cross-correlation between pairs of overlapping tiles, T and T', at each tilt angle:

$$\arg\max_{x,y} \frac{\sum_{i,j} T(i,j)\, T'(i+x, j+y)}{\sqrt{\sum_{i,j} T(i,j)^2 \sum_{i,j} T'(i+x, j+y)^2}} \tag{2.1}$$

where the sum is over all pixels at coordinates (i, j) that are in register when T' is translated by (x, y) while the position of T remains fixed. The initial search was performed in a box of length 30 nm centered on the beam coordinates used during data collection. If the translational shift that maximized the cross-correlation score was located at the edge of this box, the search box was re-centered to this position and another search was performed. This calculation yielded the relative positions for each pair of overlapping tiles.

The full mosaic for each tilt angle was then generated by fixing the central tile at the origin and determining the coordinates of the surrounding six tiles relative to this anchor tile (Fig. 2.1A). The position of each tile in the surrounding layer was estimated as the mean of all pairwise positions weighted by the cross-correlation scores between the tile of interest and each of its neighbors. Using the cross-correlation coefficients as weights ensured that overlap regions with the highest contrast features contributed most to tile positioning. The coordinates of the tiles in this first ring were then fixed. Tiles in each successive concentric ring were positioned using the same strategy, based on the consensus coordinates from pairwise registrations between neighboring tiles in the ring under consideration and anchor tiles in the previously fixed ring. Once the positions of the tiles in the outermost ring were determined, the full registration procedure was repeated using the optimized positions as the tiles' starting coordinates. Tile registration was then performed on the unbinned data, using the optimized coordinates as the starting tile positions and a smaller search space.

Tiles were then stitched based on these optimized positions to generate a mosaic (Fig. 2.3, inset). For pixels lying in the overlap regions, the intensity values were selected from the tile whose center was nearest to the pixel. Continuous cellular features were observed in the overlap regions of the montage projection images, indicating minimal radiation damage and robust stitching (Figs. 2.3 and 2.14). Occasionally the combination of beam shift error and masking resulted in small gaps between tiles; these missing pixels were filled by randomly sampling intensity values in the surrounding region to prevent holes in the stitch. Each montage projection image was then shifted to compensate for the spiraling translational offsets applied between tilt angles, cropped to the maximal region imaged at all angles, and stacked to generate a tilt-series. The size of each tilt-series was approximately 33 GB.

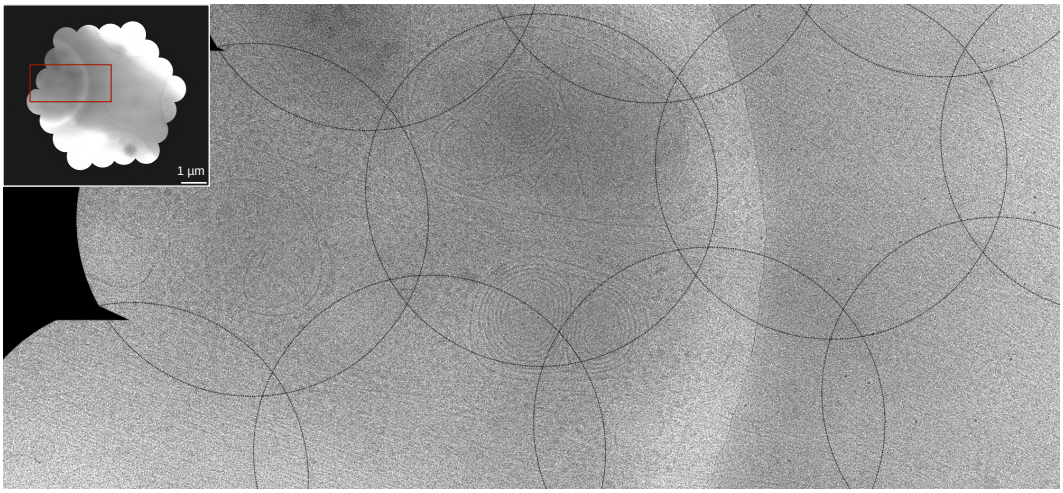

Figure 2.3: Continuity of cellular features in the overlap regions indicates successful stitching. The montaged projection image at 0° from a representative tilt-series is shown in the upper left inset. The region boxed in red is visualized at higher detail in the main image, with the boundaries of the circular tiles drawn in black. The clear and continuous membranous features visible in the overlap regions between adjacent tiles suggests both successful stitching and minimal radiation damage despite the extra dose. The diameter of each circular tile is 1.08 μm.

Methods for sample preparation and tomogram reconstruction are available in the original publication from which this chapter is adapted.

2.4 Example montage cryotomograms

For proof-of-principle, we collected montage tilt-series from the thin edge of HeLa cells and reconstructed them into tomograms. During data processing, Thon rings to better than 15 Å were frequently observed on individual tiles (Fig. 2.2B) and detected to better than 10 Å in the untilted stitched projection images for two of the

three tomograms (Fig. 2.13). Distinct cellular features that spanned the overlap regions were visible throughout the tilt-series (Figs. 2.3 and 2.14). These observations suggest that radiation damage was not severe despite the additional dose required for stitching. Consistent with this, there was no evidence of bubbling, a hallmark of radiation damage, or unevenly distributed dose in the reconstructed volumes, which were visually comparable to cryotomograms collected by standard data collection protocols. Inspection of the montage tomograms revealed rich and diverse cellular structures, including microtubules, multilamellar vesicles, mitochondria with calcium granules [52], actin bundles, and ribosomes (Figs. 2.4-2.6 and 2.15). Movies are available in the original publication.

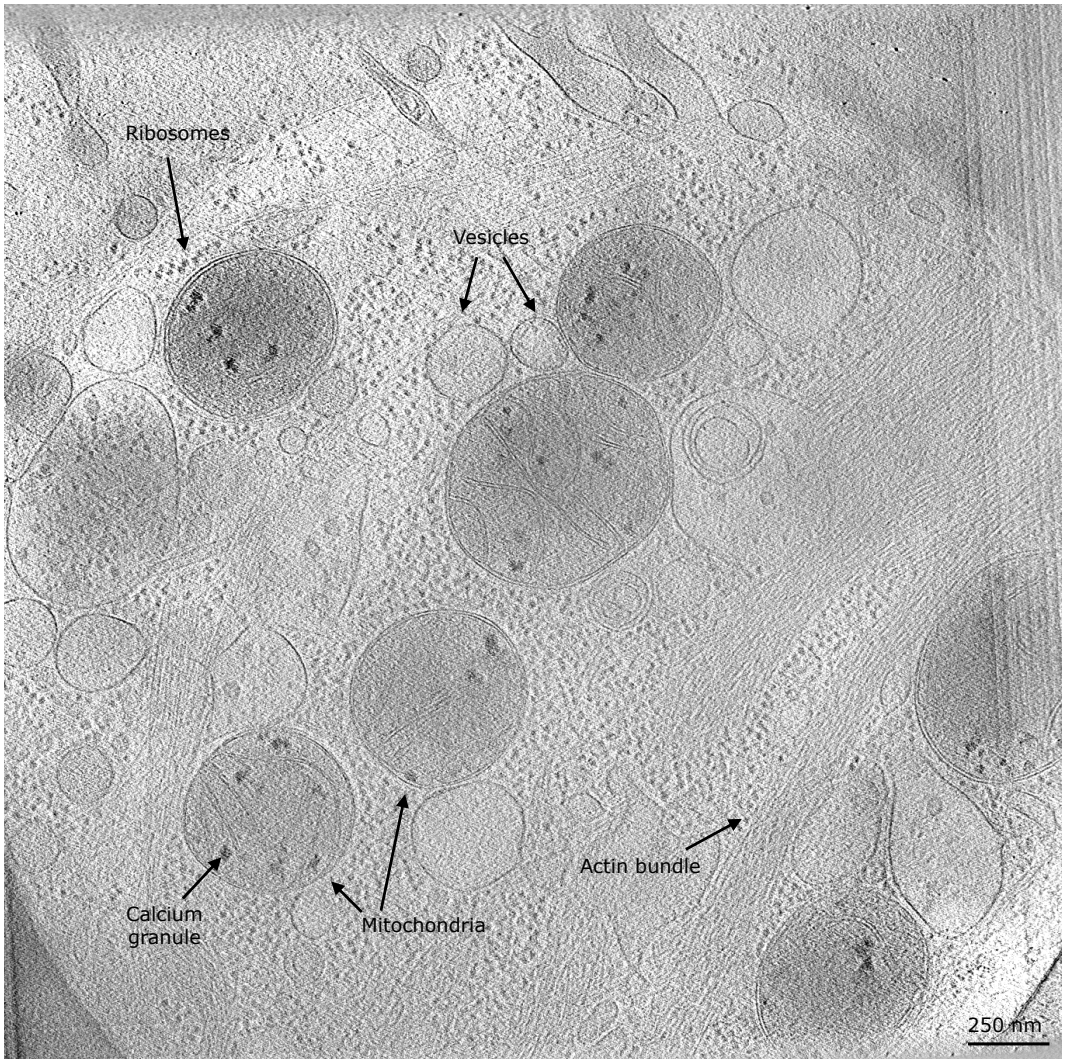

Figure 2.4: Diverse cellular features are observed in an example cryo-tomogram reconstructed from montage tilt-series. A slice is shown from a representative montage tomogram that spans a 3.3 μm^2 field of view.

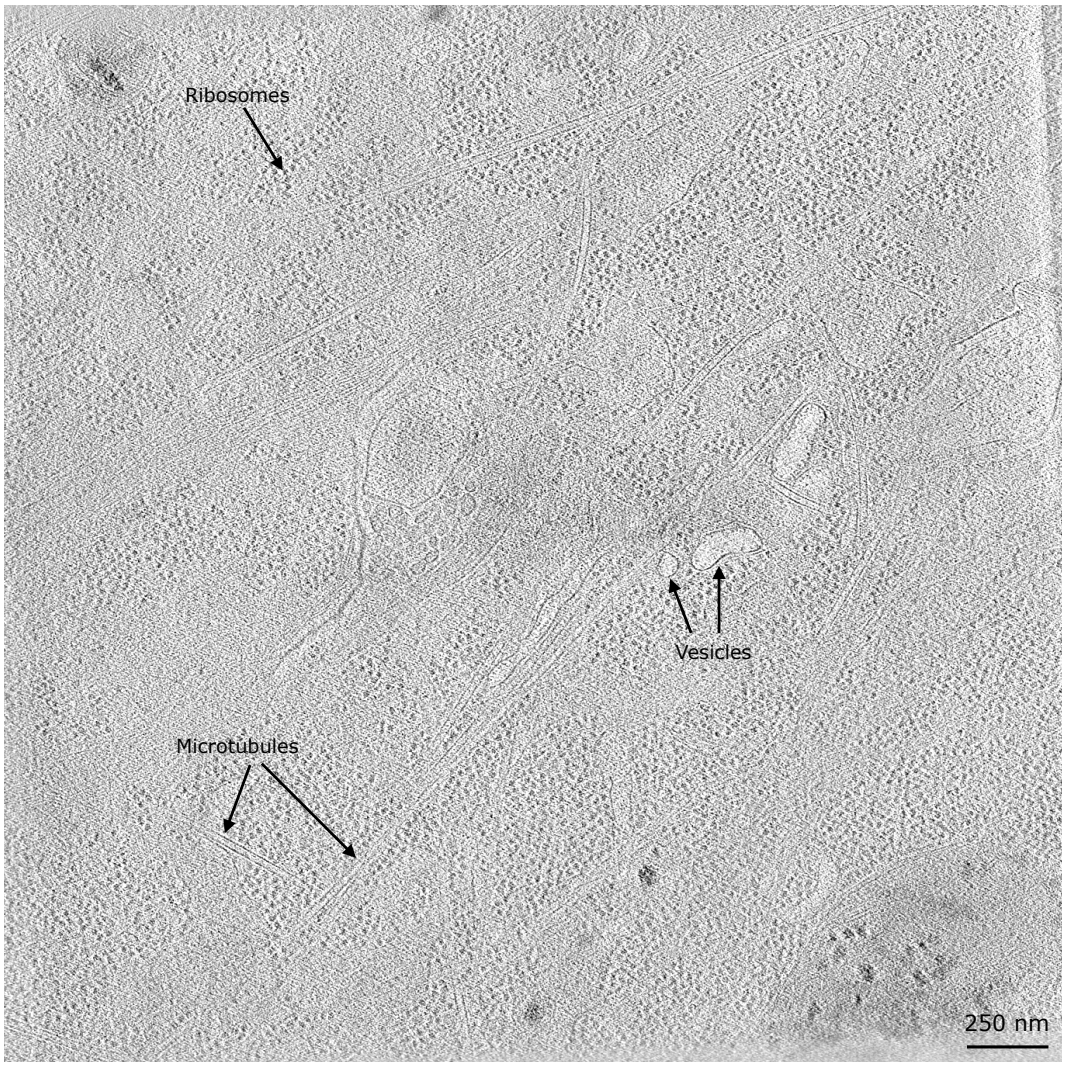

Figure 2.5: Representative montage cryo-tomogram from the thin edge of a HeLa cell. A tomographic slice spanning a 3.3 μm^2 field of view is visualized, with cellular features of interest annotated.

To quantitatively assess data quality, we performed subtomogram averaging of ribosomes picked from three tomograms collected at different defocus values. The subtomogram average grossly resembled the reference structure, with a resolution of 27 Å based on the Fourier shell correlation (FSC) between random half-sets (Fig. 2.7A-B). However, the observation of Thon rings to higher resolution — in some cases to 10 Å in the stitched projection images — suggested the retention of higher resolution signal in the montage tilt-series (Fig. 2.2B, right). It is possible that the resolution of the subtomogram average could be improved by the inclusion of more particles across a more finely sampled defocus range. Mapping the ribosomes back onto the tomograms revealed that the particles used to generate the subtomogram

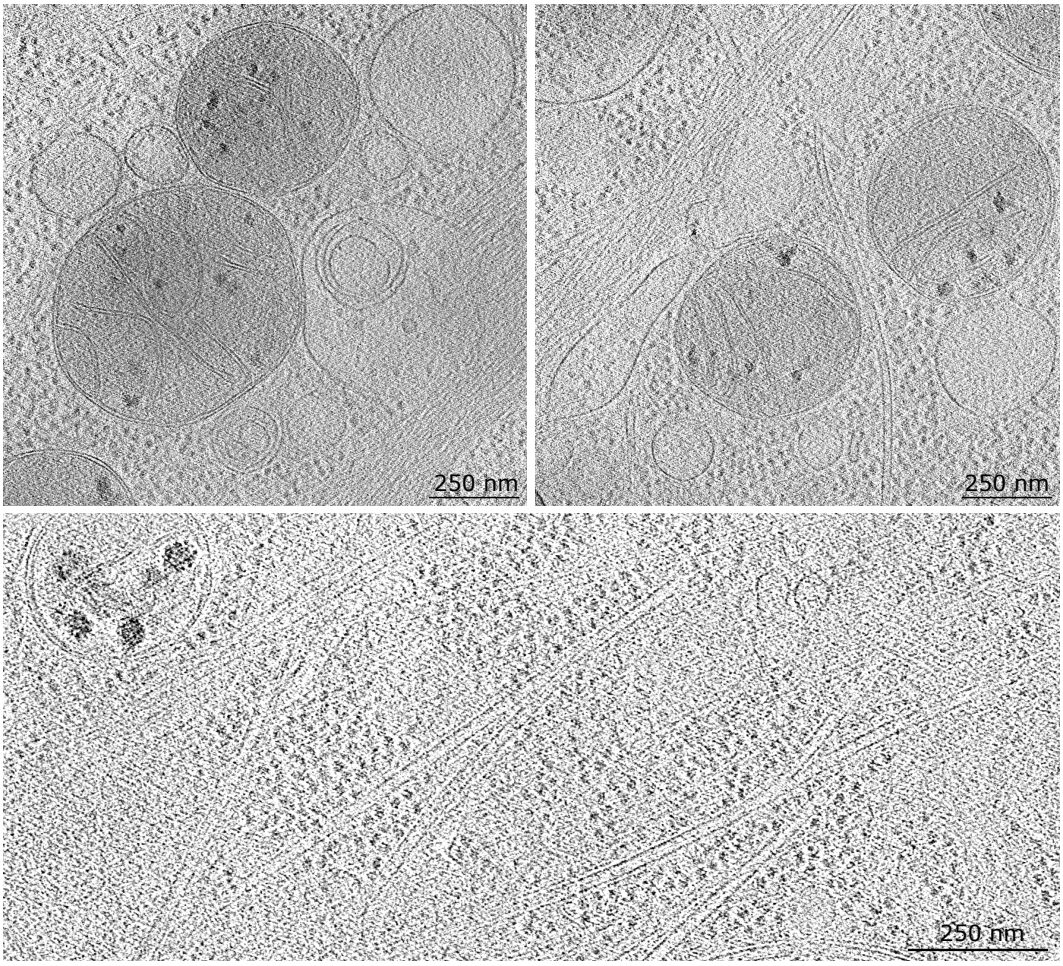

Figure 2.6: Insets from representative montage cryo-tomograms. Close-up views from the cryo-tomograms presented in Figs. 2.4 and 2.5 visualize the mitochondria (upper) and microtubules (lower) in greater detail.

average were spread throughout the sample rather than localized to particular regions (Figs. 2.7C-D), in keeping with the aim of evenly distributing the dose during data collection.

2.5 Discussion

Here we present a tomographic data collection and processing workflow to acquire montage tilt-series with a small pixel size but a large field of view. We used simulations to determine an acquisition strategy that efficiently distributed additional dose throughout the specimen and developed algorithms to stitch the recorded data into seamless projection images. We then assessed the efficacy of this pipeline by applying it to the thin edge of HeLa cells, yielding tomograms that spanned a 3.3 μm^2 field of view with a pixel size of 7.95 Å. The pixel size of the montage tilt-

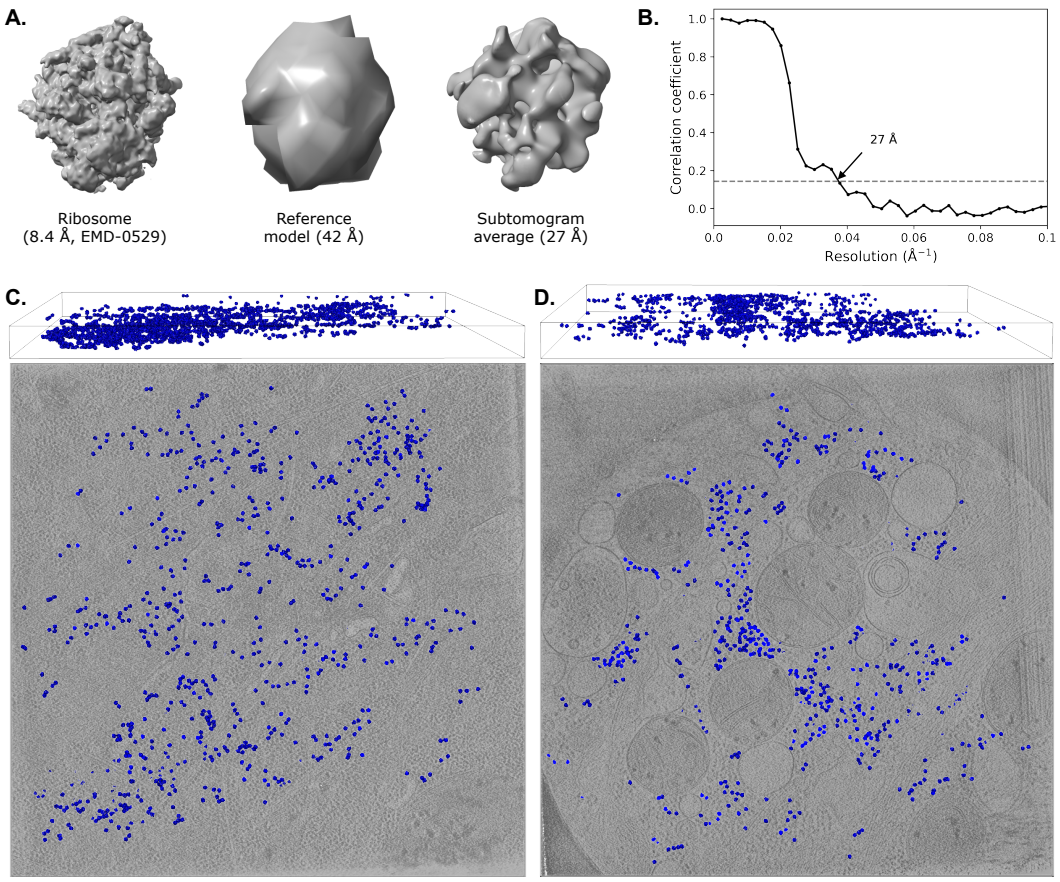

Figure 2.7: Ribosome subtomogram average and spatial distribution. (A) Isosurfaces are shown from an 8.4 Å structure of the eukaryotic ribosome (left), the 42 Å reference map used during averaging (center), and the subtomogram average from particles picked from three montage tomograms (right). (B) The Fourier shell correlation between random half-sets indicated a resolution of 27 Å for this reconstruction. (C-D) Particles (blue) used to generate the subtomogram average are mapped to their positions in two tomograms (upper) and overlaid on representative slices through these volumes (lower).

series, which would determine the Nyquist frequency of a subtomogram average, was 2.65 Å. The observation of Thon rings to better than 10 Å in the untilted stitched images indicated that the montage tilt-series retained signal to a resolution typical of cryo-ET, despite the extra dose required for stitching that historically has been considered prohibitive to this method's success.

These results are a pioneering demonstration of montage tomography that we hope will motivate further development of this technique. On the data acquisition side, we anticipate that the next phase will involve expanding the spiral pattern to encompass hundreds rather than just tens of tiles. This would dramatically expand the field

of view, which in principle is only limited by aberrations associated with extreme beam shifts and, for FIB-milled specimens, the accessible areas of lamellae. Such massive montages would in turn require the development of new software to optimize alignment and reconstruction. We anticipate that the optimal strategy will involve cropping out subvolumes from the initial reconstruction and then using these as fiducials to refine the tile alignments and estimated defocus, while also modeling the global warping of the sample that occurs during data collection. This approach would be an extension of the per-particle, per-tilt refinement employed by some subtomogram averaging packages like emClarity and EMAN2 [23, 53, 54], and similar to the multi-particle refinement scheme used by the software tool M to model spatial deformations of the specimen [24]. Extending these algorithms to montage data would not only maximize the signal extracted from the small montages shown in this work, but also be invaluable for much larger montages, where warping and doming are more severe.

Once such software is available to maximally exploit the montage technique, a pressing question will be whether montage tomograms offer more information than the corresponding single-exposure tomogram acquired at lower magnification. We anticipate that the montage technique will retain higher resolution information, in part because one of the principal resolution-limiting factors in cryoET is the precision to which the defocus can be estimated for CTF correction. In principle, montage tomography should permit more precise defocus determination because the geometrical relationship between all tiles is known *a priori*, such that CTF corrections can leverage all the available data, which far exceeds the amount of data acquired for a single-exposure tomogram. The disadvantage of montage cryoET is the additional exposure received by the overlap regions, but these are minimized by an efficient tiling scheme and can still be integrated during reconstruction with appropriate dose-weighting. We observed that only a small fraction (~4%) of the imaged sample must be discarded due to corruption by residual Fresnel fringes, but fine-tuning of the FFI set-up may further reduce this. We expect that the benefits provided by a more precise defocus estimation will outweigh the cost of this small amount of wasted dose, though further software development is needed.

We also anticipate that montage tomography could prove particularly useful for cellular lamellae. The FIB-milling required to produce these samples is challenging and time-consuming, so maximizing the yield during data collection is a critical concern [37–39]. By imaging a large field of view, this technique also increases the

chances of capturing transient or infrequent cellular events. Finally, data collected by this method will be a valuable resource for visual proteomics, which seeks to build atlases of cellular structure at molecular resolution [55, 56]. As technical advances continue to improve the high-resolution limit of cryo-ET, montage tomography offers a way to provide more cellular context while retaining detailed views of macromolecules *in situ*.

Data and code availability

Raw data are available at the Caltech Data Repository (`https://data.caltech.edu`) under accession IDs 2096, 2099, and 2103. The processed tilt-series binned to 5.3 Å can be found in the Caltech Electron Tomography Database (`https://etdb.caltech.edu/`). The code developed for data collection and processing is available at `https://github.com/apeck12/montage`.

Acknowledgments

Data collection and analysis were respectively performed at the Beckman Institute Resource Center for Transmission Electron Microscopy and the Resnick High Performance Computing Center at Caltech. We thank Wei Zhao for preliminary samples, David Mastronarde for valuable discussions, and Tom Morrell for help with the Caltech Data Repository. A.P. is The Mark Foundation for Cancer Research Fellow of the Damon Runyon Cancer Research Foundation (DRG 2361-19). This work was supported by NIH grant AI150464 to G.J.J.

2.6 Appendix

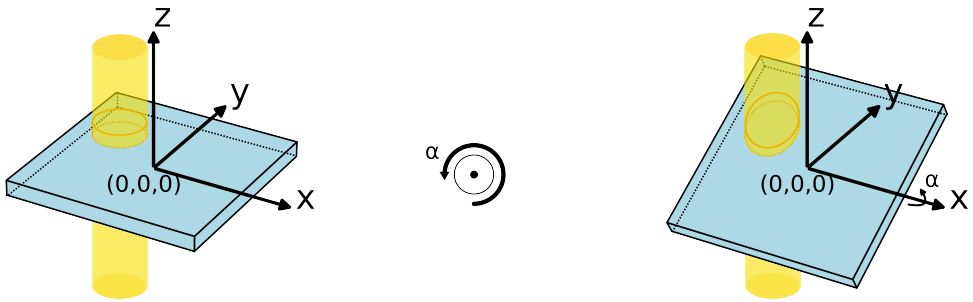

Figure 2.8: Coordinate system and axes conventions for simulating montage data collection. All simulations use the standard right-handed Cartesian coordinate system, with the origin at the center of the specimen (blue rectangular prism). The specimen is tilted around the x-axis. A tilt angle α is positive if the specimen is tilted counterclockwise when viewed from the positive side of the x-axis. An electron beam (yellow cylinder) is delivered to the specimen along the z-axis from the positive side. The illuminated volume of the specimen is the cylinder outlined by the orange ellipses. In projection the beam is circular at the first tilt angle of 0° and becomes increasingly elliptical at higher tilt angles, with the long axis of the ellipse aligned with the y-axis of the detector.

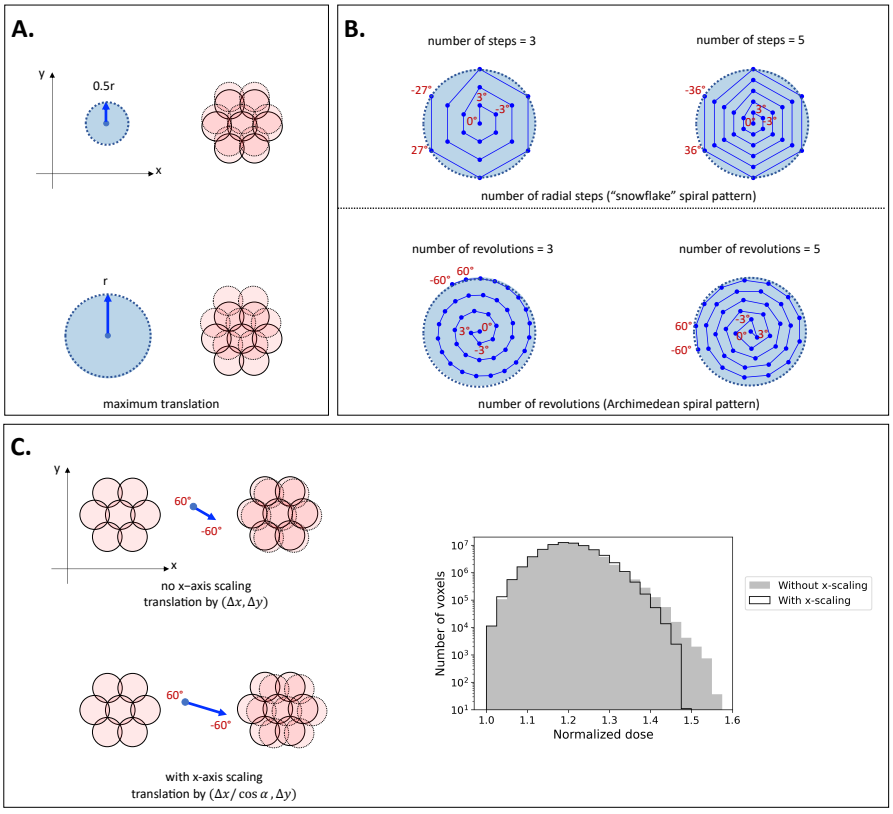

Figure 2.9: Global translational offsets of the tiles change which regions of the specimen are imaged more than once during the tilt-series. For visual clarity, only the central seven tiles (pink circles) are shown for before (solid) and (dashed) after the indicated translation between successive tilt angles. Four parameters were tuned to optimize the offsets. (A) The maximum translation for any tile during the tilt-series was varied from 0 to r (one beam radius). Examples are shown for maximum translation of $0.5r$ (top) and r (bottom). (B) The number of radial steps / revolutions specifies how tightly packed the spiral pattern is. Each blue dot represents the position of the central tile at each tilt angle. The large light blue circles represent the region of allowed displacement as described in (A). For the snowflake pattern (top), this parameter is the number of points between the center and edge of the light blue circle. The pattern is repeated from the center if the outermost position is reached before the final tilt angle. Archimedean spirals avoid this issue by fitting all tilt angles into the light blue circles regardless of the number of revolutions (bottom). (C) The x-axis component of all translations is optionally scaled by the cosine of the current tilt angle. The difference is most evident at high-tilt angles. This scaling is motivated by the elliptical projection of the beam at high tilt angles, with the long-axis of the ellipse aligned with the y-axis (Fig. 2.8). Regions of the specimen that were previously in the overlap region are more likely to be in the overlap region again along the y-axis than x-axis. To compensate for this, an additional x-axis translation was tested. As shown by the overlaid histograms on the right, this strategy was effective for a basic spiral pattern.

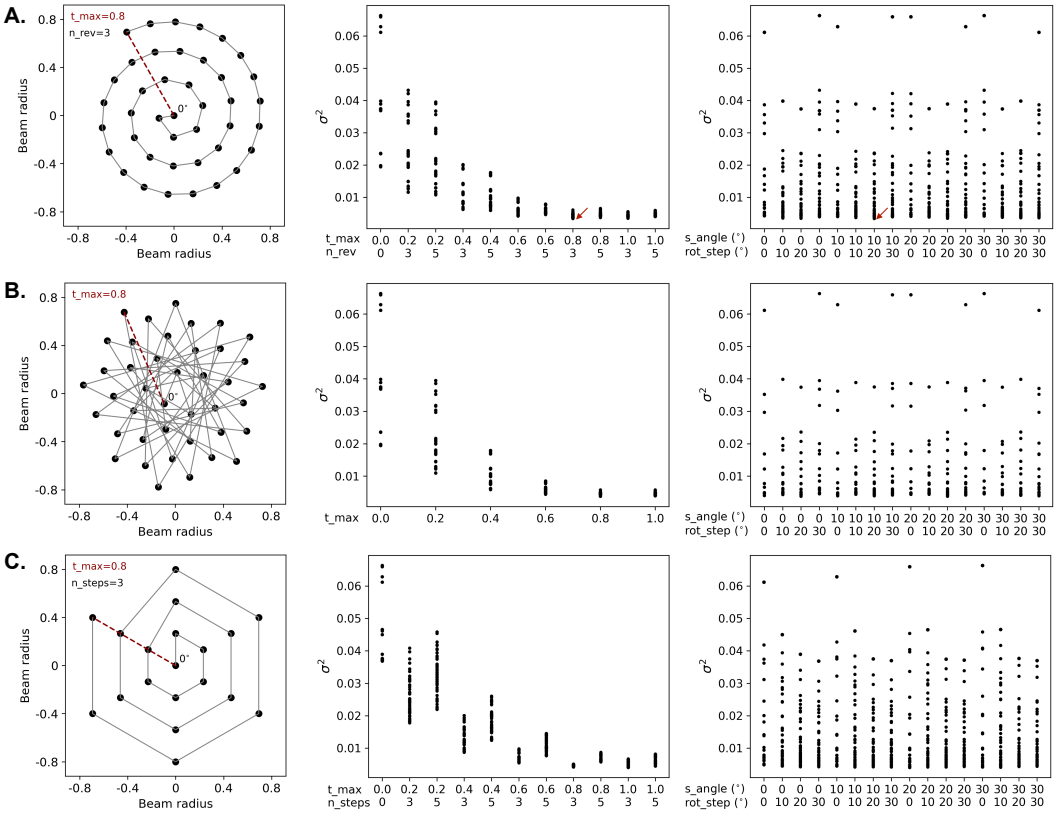

Figure 2.10: Tiling pattern efficiency is more sensitive to translational than rotational offsets between tilt angles. The changes in the position of the central tile between tilt angles for the (A) Archimedean spiral, (B) sunflower, and (C) snowflake patterns are shown at left. For each pattern, the position of the central tile at the first tilt angle of 0° is noted, and its position follows an outward spiral (indicated by the grey line) during the tilt-series. For the snowflake pattern, the original pattern is repeated after the outermost point is reached. Translational parameters are noted at the upper left. For each, t_max (red line) indicates the maximum displacement of the central tile during the tilt-series. The parameters n_rev and n_steps correspond to the number of revolutions and radial steps for the spiral and snowflake patterns, respectively. Variants of each spiral pattern were simulated and ranked based on the variance (σ^2) of the normalized dose distribution. Variance scores are shown as a function of the indicated translational parameters (center) and rotational parameters (right). For the latter, parameters s_angle and rot_step respectively refer to the initial angular offset of the hexagonally-packed tiles in the plane of the detector and the global rotation applied between tilt angles. Simulations were performed using a tilt-range of ±60° with 3° increments between tilts, applying a 1/cos exposure scheme, and assuming that Fresnel fringes contaminated 2% of the beam radius. The red arrows in (A) indicate the top-ranked pattern.

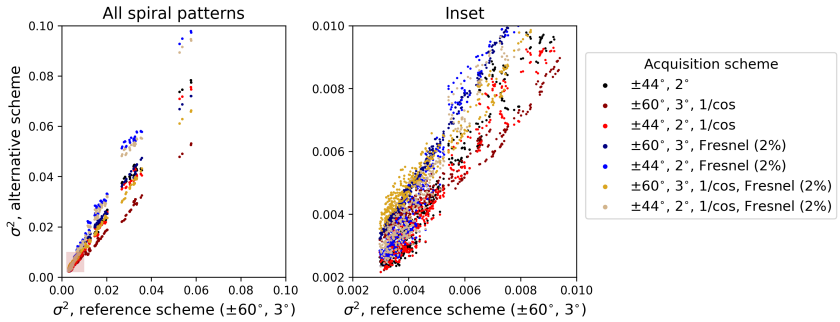

Figure 2.11: Tiling strategy efficiency is similar across different data collection schemes. 546 unique variants of the spiral pattern were tested and scored by the variance (σ^2) of the accumulated dose distribution. The score of each pattern for the indicated data acquisition scheme is compared to its score for a reference data collection strategy of ±60° with 3° increments between tilt angles. Alternative acquisition schemes used a tilt-range of ±44° with 2° increments between tilt angles, increased the dose as 1/cos of the tilt-angle, and/or increased the overlap between neighboring tiles to account for Fresnel fringes spanning 2% of the beam's radius. The transparent pink box overlaid on the highest-ranked patterns (left) is shown in the inset (right).

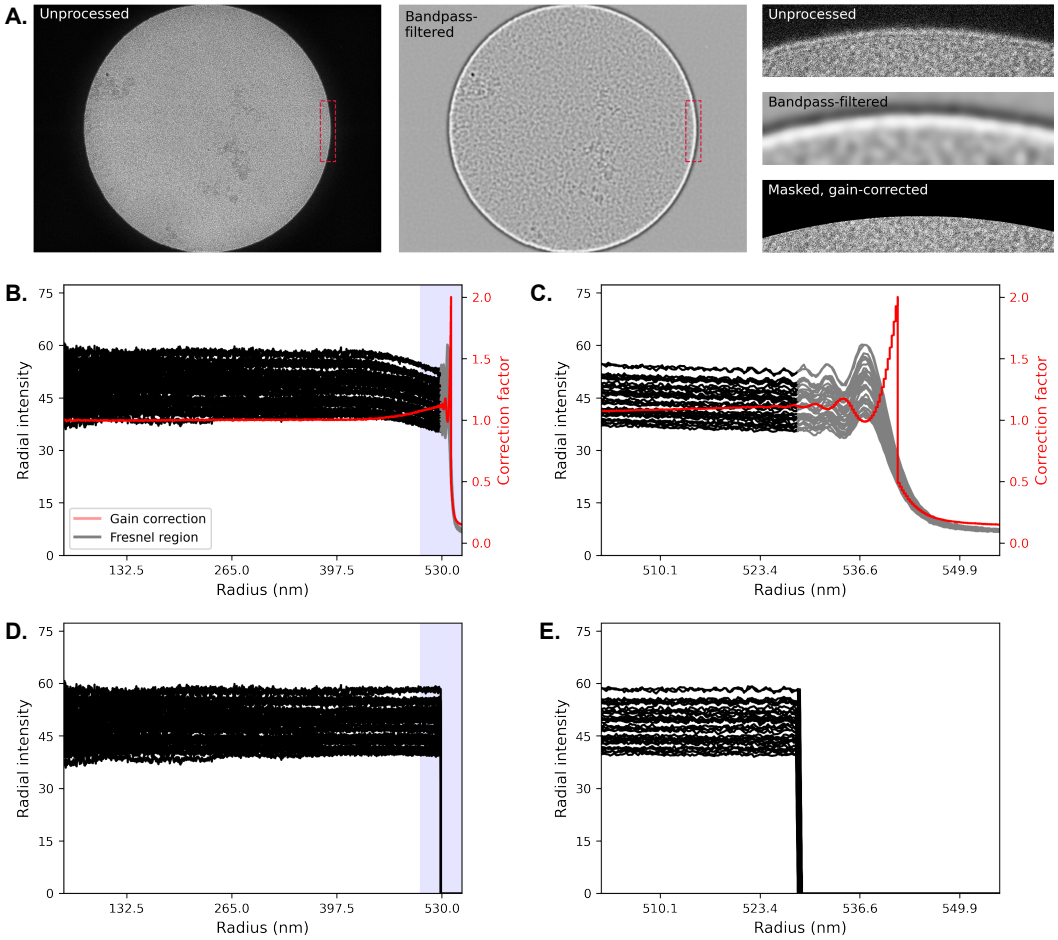

Figure 2.12: Removal of the fringe-corrupted region and correction for uneven radial illumination. (A) A representative tile is shown before (left) and after (center) applying a bandpass filter. In the filtered tile, a high intensity ring (white) coincides with the Fresnel fringes at the tile's edge. (Right) The region boxed in red from the unprocessed, bandpass-filtered, and masked tiles visualizes elimination of the Fresnel fringes. (B) Radial intensity profiles for all 37 tiles from a representative tilt angle are plotted in black, with the region judged to be corrupted by Fresnel fringes plotted in grey. The radial gain factor used to correct uneven illumination is plotted in red. (C) Inset of the shaded blue region in (B) that focuses on the edge of the tiles affected by Fresnel fringes. (D) Radial intensity profiles of these tiles after correcting for uneven radial illumination and masking the fringes and (E) the corresponding inset.

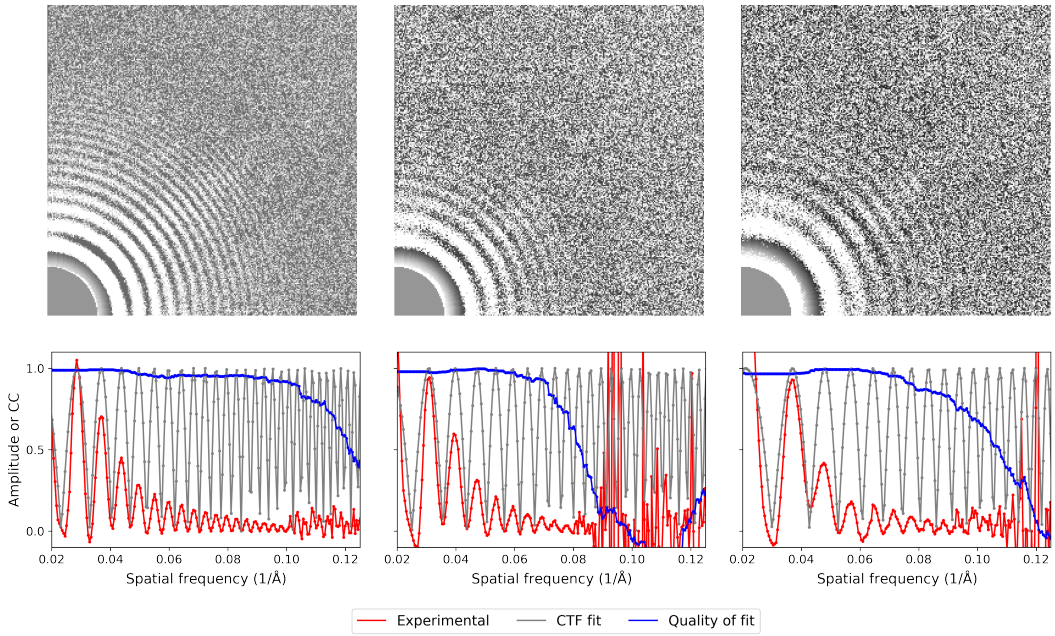

Figure 2.13: CTFFIND4 detects Thon rings to 8-12 Å in the untilted stitched projection images. The 2D experimental spectrum (upper) and rotationally-averaged 1D CTF fits (lower) are shown for the CTF-uncorrected stitched projection images collected at 0° from three different tomograms. The leftmost plots are also shown in Fig. 2.2B but are reproduced here for comparison.

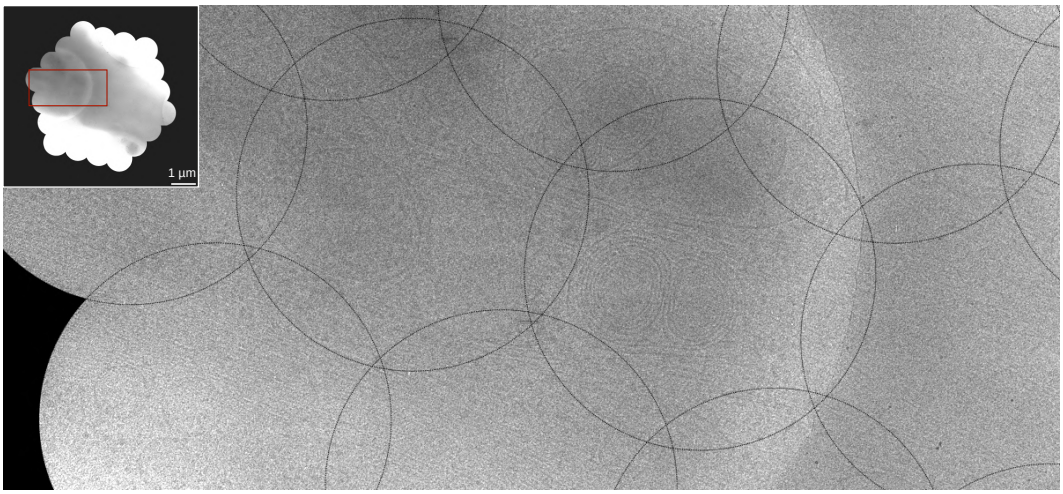

Figure 2.14: Continuous cellular features are observed in the overlap regions at intermediate tilt angles. As in Fig. 2.3, except the stitched tiles imaged at 30° are visualized. The inset displays the full mosaic prior to cropping for reconstruction. The region boxed in red is enlarged in the main image. The boundaries of the 1.08 μm diameter circular tiles are outlined in black. The continuity of membrane features in the overlap regions between adjacent tiles suggests minimal radiation damage, even at an intermediate stage of data collection.

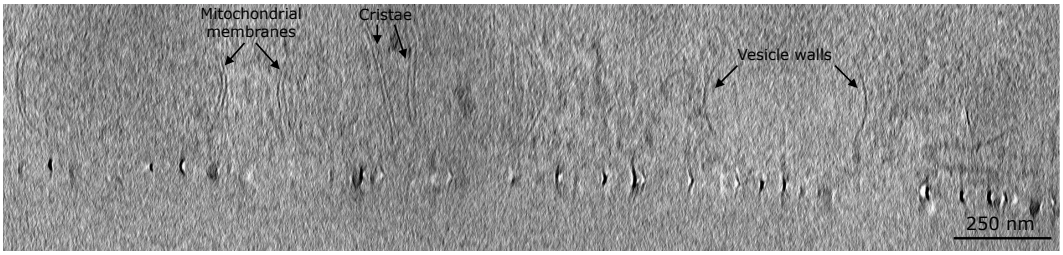

Figure 2.15: Continuous cellular features in an orthoslice from a montage tomogram. An orthoslice from the tomogram presented in Fig. 2.7 is shown, with cellular features annotated.

Chapter 3

ASSESSING THE APPLICABILITY OF BAYESIAN INFERENCE FOR MERGING SMALL MOLECULE MICROED DATA

Abstract

Microcrystal electron diffraction (MicroED) is an emerging technique for characterizing small molecule structures from nanoscale crystals. Merging data from multiple crystals is a particularly challenging step in the microED workflow. A common practice is to manually curate datasets and apply scaling programs conventionally utilized in rotational X-ray diffraction (XRD), but this could be time-consuming and risks introducing human bias in data analysis. Recently, a Bayesian inference program named Careless [8] has demonstrated excellent performance in merging macromolecular XRD data. Here, the applicability of Careless to small molecule microED data is evaluated and an investigation of the impact of dataset curation is performed. Benchmarking against XDS/XSCALE shows that Careless is an effective complementary approach that merges data to a higher $CC_{1/2}$ value at high resolution. Furthermore, merging outcomes are not significantly improved by curating datasets either manually or with an automated extension to Careless, cautioning against the common practice of manual dataset curation.

3.1 Introduction

Structural characterization is critical for understanding small molecule properties and advancing research in chemistry fields, including organic chemistry, natural product chemistry, and drug discovery. For decades, single-crystal X-ray diffraction (SCXRD) has been the gold standard for determining the bond connectivity of molecules with high precision. However, the need to obtain large (10–100 μm) single crystals has severely limited the applicability of SCXRD. To overcome this limitation, new methods have been developed to solve structures from smaller crystals [57, 58]. Synchrotrons now offer micro-focus beamlines that can reduce the beam width to match micron-sized crystals [59]. X-ray free electron laser (XFEL) facilities enable studying even smaller crystals using serial crystallography by delivering extremely bright, femtosecond-long pulses that have the additional benefit of outrunning radiation damage [60]. However, few structures of small molecules have been elucidated using these resources, and limited XFEL sources and micro-focus beamlines render these techniques unsuitable for routine analysis compared to in-house instruments, which offer rapid turnaround times.

Microcrystal electron diffraction (microED) [61], also known as continuous rotational electron diffraction (cRED) [62] and a sub-method of 3D electron diffraction (3D ED), provides a powerful alternative for small molecule structure determination [62–68]. Compared to photons, electrons interact more strongly with matter, enabling this technique to measure diffraction signals from nanoscale crystals at sub-Ångstrom resolution. For small molecules, such nanocrystals are generally far easier to obtain than micron-sized or larger crystals required by other techniques, and they can be found in seemingly amorphous powders as well as crude natural products extracts [63, 69]. MicroED is also advantageous due to the broad availability of transmission electron microscopes (TEM) and the potential to use this technique with increasing throughput [69–73]. However, the widespread adoption of microED calls for improvements in data processing [74]. Current microED workflows typically leverage software such as XDS which was originally developed for rotational XRD experiments [75], but several steps of microED data processing still require time-consuming manual intervention.

A particularly challenging step of microED data processing is data merging. In SCXRD, merging refers to the process where measured symmetry-equivalent reflection intensities are reduced to a set of unique values after being scaled to correct systematic errors, a process also known as data reduction. The scaled and merged

reflections can then be phased to solve the structure. In small molecule crystallography, *ab initio* phasing is a standard practice, which requires accurate estimates of the structure factor amplitudes at 1.2 Å or higher resolution [76] from the merging output. Compared to SCXRD, merging in microED is inherently more challenging. The background noise is higher due to non-negligible diffuse and inelastic scattering [77]. Dynamical scattering events increase the variation among intensities that should be theoretically equivalent, effectively contributing to the errors in conventional merging [78, 79]. Moreover, most TEMs restrict the accessible tilt range to less than ±70°, resulting in a missing wedge of information where data cannot be measured. Although crystallographic symmetry should in principle overcome this low completeness, in practice, data from multiple crystals often need to be merged due to radiation damage that compromises intensities in a way that is reflection-dependent and non-monotonic in dose [80].

Multi-crystal merging is often a trial-and-error process conducted under the assumption that few crystals are sufficiently isomorphous and of high enough quality to yield correct *ab initio* phasing solutions and acceptable refinement statistics. Thus, even though high multiplicity from redundant measurements is considered helpful in macromolecular XRD [81], empirically, many microED small molecule structures have been solved by merging only several crystals out of the tens to hundreds collected. In other emerging fields of diffraction experiments such as serial femtosecond crystallography (SFX) from multiple crystals, specialized software has been developed [82, 83] to merge data that conventional methods find challenging. Nevertheless, this is a computationally expensive approach. Small-wedge serial crystallography has prompted iterative approaches [84, 85] that combine preliminary clustering and then outlier rejection [86]. In the microED field, clustering-based heuristics [71, 87–89] and brute-force enumeration of dataset combinations [73] have been deployed, but the manual curation of datasets remains a dominant practice.

A recent machine learning (ML)-based merging program, Careless, promises a unifying framework through Bayesian inference for any type of diffraction experiment [8]. It has been successfully applied to a variety of macromolecular XRD experiments but has not been validated in small molecule or microED studies. Compared to conventional programs, Careless has the potential to be adapted for multi-crystal merging with minimal human bias in selecting datasets or setting cutoff values for clustering and filtering datasets. Here, we reprocess 17 molecules from previous

studies on natural products and pharmaceutical compounds to test the applicability of Careless in comparison to a conventional scaling and merging program XDS/XSCALE [75]. In addition, the popular practice of manual dataset curation motivates us to explore an extension to the ML algorithm that automates this process. We compare the merging outcomes of dataset curation, whether manual or automated, with a naive merging of all datasets. Finally, we show how well results from different merging protocols are translated to *ab initio* structures using standard phasing programs, and present recommendations to microED practitioners.

3.2 Results

Careless as an alternative merging tool for microED data

To assess whether Bayesian inference generalizes well to microED data processing, we compile existing microED datasets comprised of a diverse set of 17 small molecules (Fig. 3.5), from simple cases such as calcium oxalate (Fig. 3.5-**15**) [69] to challenging cases such as fischerin (Fig. 3.5-**5**) [65]. They span a range of crystallographic complexity, including 7 space groups and unit cell volumes from 10^3 to 10^4 Å^3 (Table 3.5). 6 of the molecules were solved from single-crystal datasets processed by XDS, and the rest were solved after individually processing all datasets in XDS and then merging a manually curated subset in XSCALE [63, 65, 69, 90–92].

Careless is first evaluated using the single-crystal datasets and manually curated datasets in multi-crystal merging. Previously published structures (Fig. 3.5 and Table 3.5) are used as reference structures to assess merging performance. For consistency, in multi-crystal merging, we adhere to the same dataset curation manually done by the authors of the reference structures, and the effect of manual dataset curation is examined in the next section (3.2). Merging outcomes are evaluated by examining the internal consistency of the intensities measured by $CC_{1/2}$ as well as $CC_{F_oF_c}$, which indicates the accuracy relative to calculated structure factors from the reference structure (Methods section 3.4). To compare the average performance of different merging protocols, we use a stringent criterion of whether the 95% bootstrap confidence intervals (CI) for the mean overlap or not to assess statistically significant differences.

Compared to XDS/XSCALE [75, 93], a conventional merging software, Careless merging yields lower overall $CC_{1/2}$ but comparable overall $CC_{F_oF_c}$ for single-crystal merging (Table 3.1) and multi-crystal merging (Table 3.2). This suggests that even

though the Careless merging output is less precise among symmetry-equivalent measurements, it is not necessarily less accurate. In the highest resolution bin of each dataset, Careless performance is on average similar to XDS/XSCALE as indicated by the 95% CIs of the $CC_{1/2}$ and $CC_{F_oF_C}$ of all 17 molecules (Table 3.3).

Table 3.1: Single-crystal* merging results.

	molecule	AMG10 (**13**)	mannitol (**14**)	calcium oxalate (**15**)	chryso-phanol (**16**)	6β-hydroxy-eremoph-ilenolide (**17**)
XDS	highres† bin (Å)	0.9-0.85	1.01-0.95	1.01-0.95	0.9-0.85	0.9-0.85
	Completeness (%)	98.5 (98.6)	82.2 (87.3)	87.7 (88.9)	91.8 (92.4)	89.1 (95.1)
	$CC_{1/2}$ (%)	98.4 (57)	99.8 (21.9)	98.7 (74.8)	98.1 (88)	98.9 (38.9)
	$CC_{F_oF_c}$ (%)	76.6 (71.9)	95.4 (61.2)	86.8 (60.9)	88.1 (85)	89.4 (30.5)
	phasing	✓‡	✓‡	✓‡	✓‡	✓‡
Careless	highres bin (Å)	0.89-0.85	0.98-0.95	1-0.95	0.88-0.85	0.88-0.85
	Completeness (%)	99.7 (98)	84.6 (86.2)	89.8 (94.7)	92.2 (92.6)	89.7 (94.4)
	$CC_{1/2}$ (%)	87 (61)	97.9 (27.5)	92.9 (85.4)	92.2 (78.1)	94.7 (49.5)
	$CC_{F_oF_c}$ (%)	75.1 (80.9)	92.3 (50.2)	92.1 (55.7)	88.2 (76.5)	85 (38.6)
	phasing	✓‡	✓	✓‡	✓	✓

* For biotin single-crystal results, see Table 3.2.
† Statistics in the highest resolution (highres) bin are shown in parentheses.
‡ Phased by SHELXT. Otherwise phasing is performed using SHELXD.

Table 3.2: Multi-crystal merging results using manually selected datasets.

	molecule	demeth-oxyvir-idin (**1**)	calliter-penone acetate (**2**)	pachy-basin (**3**)	Py-469 (**4**)	fischerin (**5**)	peysso-barican-oside B (**6**)
	# crystals	2	3	3	2	4	3
XDS/XSCALE	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	90 (95.9)	82.4 (85.8)	83.6 (81.9)	84.5 (84.5)	89.9 (91.1)	94.0 (93.8)
	$CC_{1/2}$ (%)	97.1 (54)	97.8 (84.2)	98.7 (47.8)	98.8 (84.7)	99.2 (29.9)	99.1 (31.1)
	CCF_oF_c (%)	71.7 (48.8)	94.2 (88.4)	92.7 (70.0)	77.9 (85.1)	90.4 (49.1)	92.5 (72.9)
	phasing	✓	✓	✓‡	✓	✓	✓
Careless	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	90.3 (96.7)	83.3 (83.9)	83.8 (81.9)	85 (84.8)	90 (90.6)	94.6 (92.9)
	$CC_{1/2}$ (%)	92.7 (77.6)	96.2 (90.5)	96.1 (64.2)	91.9 (88.9)	83.6 (75.8)	93.8 (77.4)
	$CC_{F_oF_c}$ (%)	81.6 (54.1)	92.4 (79.1)	91.4 (37.4)	79.8 (79.3)	73.3 (82.5)	92.7 (83.4)
	phasing	✓	✓	✓	✓	✓	✓

Table 3.2 continued

molecule		YT-348 (**7**)	AMG3 (**8**)	AMG4 (**9**)	AMG7 (**10**)	AMG11 (**11**)	biotin (**12**)
# crystals		9	2	2	3	2	1
XDS/XSCALE	highres bin (Å)	1.04-1	1.04-1	1.04-1	1.03-1	0.93-0.9	0.9-0.85
	Completeness (%)	87.7 (87.6)	82.7 (85.0)	83.1 (85.3)	84.5 (86.0)	86.5 (91)	97.5 (96.4)
	$CC_{1/2}$ (%)	98.7 (26.7)	98.4 (74.6)	99.1 (85.3)	95.2 (85.6)	97.4 (29.8)	96.8 (80.3)
	CCF_oF_c (%)	89.4 (47.4)	94.1 (83.4)	95.9 (91.9)	76.4 (88.6)	90.5 (69)	86.9 (70.7)
	phasing	✓	✓	✓	✓	✓‡	✓‡
Careless	highres bin (Å)	1.04-1	1.04-1	1.04-1	1.03-1	0.93-0.9	0.89-0.86
	Completeness (%)	88.1 (88.8)	83.4 (86.9)	83.2 (83.7)	86.4 (89.1)	88.5 (87.5)	98.3 (97.4)
	$CC_{1/2}$ (%)	94.3 (57.0)	91.8 (78.7)	95.1 (93.3)	89 (77.7)	90.5 (71.2)	92.4 (62.3)
	CCF_oF_c (%)	87.8 (57.3)	92 (78.1)	95.3 (90.4)	87.2 (80.8)	89.9 (53.6)	87.6 (43.3)
	phasing	✓	✓	✓	✓	✓	✓‡

Table 3.3: 95% CI of bootstrapped mean of single-crystal and manually curated merging results.

metric (%)	XDS	Careless
$CC_{1/2}$ (overall)	97.58 - 98.66	90.56 - 93.87
$CC_{1/2}$ (highres)	46.52 - 69.46	62.10 - 77.94
$CC_{F_oF_c}$ (overall)	83.57 - 90.48	83.86 - 89.80
$CC_{F_oF_c}$ (highres)	60.35 - 76.30	57.40 - 73.72

Effects of dataset curation on multi-crystal merging

After demonstrating that Careless achieves comparable accuracy given curated datasets, we next sought to automate dataset curation within Careless. This was motivated by current practices in the microED field, where multi-crystal merging is commonly performed on datasets selected by manual inspection of the completeness, $CC_{1/2}$, R_{merge}, and other summary statistics. Manual curation of datasets could be time-consuming with a large number of datasets collected and risk introducing human bias to data analysis. Nevertheless, a naive merge of all datasets may compromise data quality. For effective comparisons, we first evaluate the impact of manual curation relative to the baseline that omits dataset curation.

Naively merging all datasets that could be indexed to the expected space group and unit cell maximizes the completeness (Fig. 3.1a) and multiplicity of the data. Using the naive merging as a baseline for comparison, manual dataset curation has the opposite effect on $CC_{1/2}$ for the two merging regimes studied: it is beneficial for XDS/XSCALE but harmful for Careless (Fig. 3.1b). Within each merging program, the effect of dataset curation is statistically significant for the overall $CC_{1/2}$ but less pronounced in the highest resolution bin. With uncurated merging, Careless achieves similar overall $CC_{1/2}$ with XDS/XSCALE but has the additional benefit of significantly better $CC_{1/2}$ in the highest resolution bin (Fig. 3.1b and Table 3.4).

Among the four merging protocols — XDS/XSCALE vs. Careless using manually curated vs. all datasets, the common practice of merging manually curated datasets by XDS/XSCALE still gives the best overall $CC_{1/2}$, while merging all datasets by Careless gives the best $CC_{1/2}$ in the highest resolution bin. Despite the different performances according to $CC_{1/2}$, on average, $CC_{F_oF_c}$ is minimally affected both overall and at high resolution regardless of the merging protocol used (Fig. 3.1c). This suggests that the accuracy of merging is not significantly improved by manual curation.

Although we find that manual dataset curation does not benefit Careless, it is still of interest to investigate whether a fully automated curation would improve the outcome. This approach, referred to as MC-Careless for Multi-Crystal Careless, uses ML principles to learn an optimal weighting among datasets to account for the variability of data quality across datasets collected from different crystals (section 3.4). This weight modulates the effective uncertainty of the intensities during model training for multi-crystal merging and is optimized jointly with the structure factor amplitudes (Fig. 3.4). It provides an alternative to the manual curation of datasets and reduces human bias in evaluating summary statistics and filtering datasets. Nevertheless, MC-Careless achieves similar performance on both $CC_{1/2}$ and $CC_{F_oF_c}$ to the original Careless that naively merges all datasets (Fig. 3.1b and c). This result is consistent with the observation above that manual curation of datasets does not improve Careless merging outcomes.

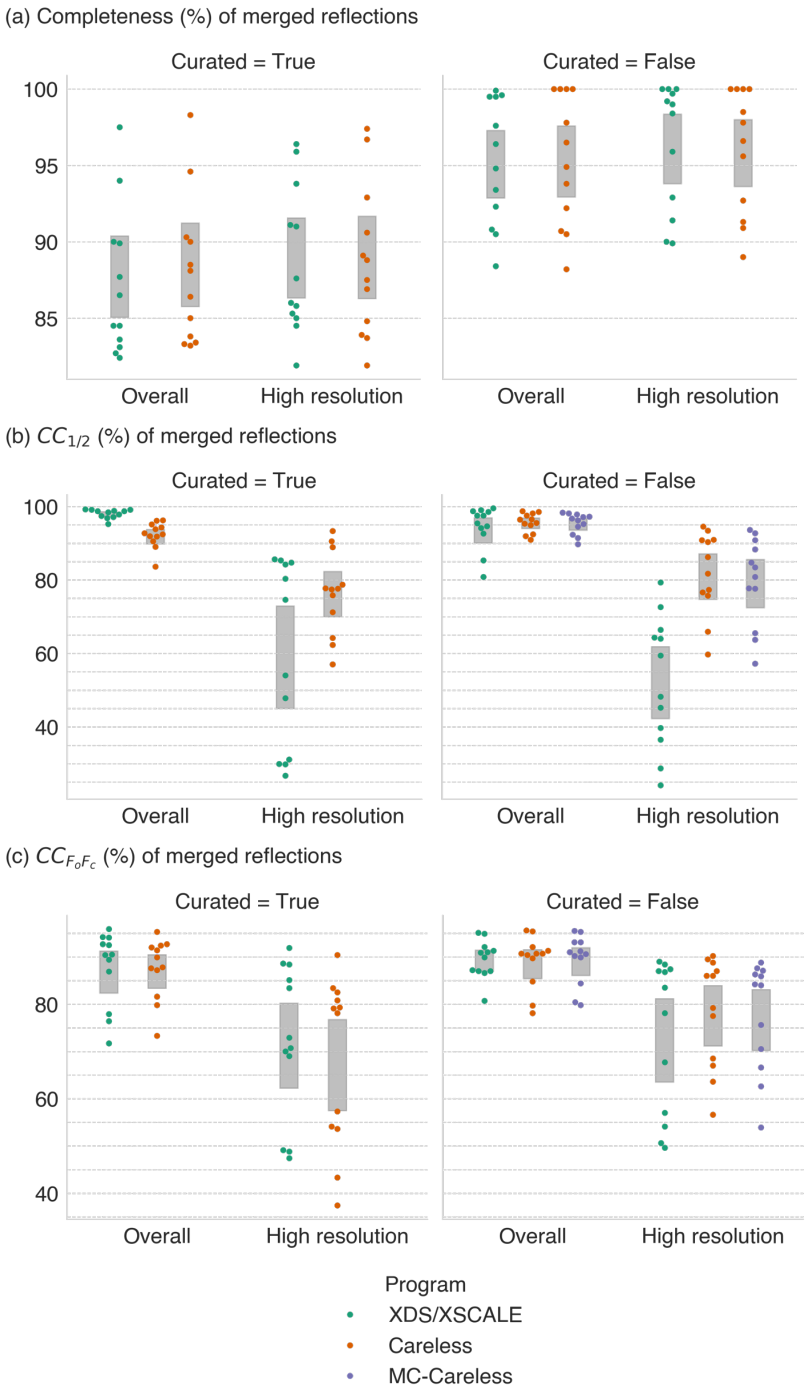

Figure 3.1: Statistics of multi-crystal merging with (Curated=True) and without (Curated=False) manual dataset curation. Grey-shaded regions represent 95% confidence intervals for the mean from bootstrapping. (a) Data completeness is maximized when using all datasets. (b) Manual dataset curation has the opposite effects on $CC_{1/2}$ in XDS/XSCALE and Careless. Automated curation by MC-Careless does not improve results. Careless consistently achieves higher $CC_{1/2}$ than XDS/XSCALE in the highest resolution bin. (c) All merging protocols achieve similar accuracy as indicated by $CC_{F_oF_c}$.

Table 3.4: Multi-crystal merging results using all datasets.

	molecule	demeth-oxyvir-idin (**1**)	calliter-penone acetate (**2**)	pachy-basin (**3**)	Py-469 (**4**)	fischerin (**5**)	peysso-barican-oside B (**6**)
	# crystals	6	9	6	21	89	16
XDS/XSCALE	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	92.3 (99)	96.4 (99.2)	99.5 (100)	94.8 (95.9)	90.8 (91.4)	97.6 (98.4)
	$CC_{1/2}$ (%)	92.6 (64)	80.8 (66.4)	97.5 (48.2)	85.3 (64.3)	99.5 (28.7)	95.4 (45.2)
	$CC_{F_oF_c}$ (%)	80.7 (78.1)	87.2 (86.8)	92.1 (67.7)	86.6 (89)	90.9 (50.6)	89.9 (57)
	phasing	✓	✓	✓‡	✓	✓	✓
Careless	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	92.9 (97.8)	96.5 (96.6)	100 (100)	94.9 (95.6)	90.7 (91.3)	97.8 (98.5)
	$CC_{1/2}$ (%)	91.9 (77.3)	94.9 (90.8)	96.5 (59.7)	98.1 (93.4)	97.5 (90.9)	95.5 (75.7)
	$CC_{F_oF_c}$ (%)	78.1 (67)	90.7 (87)	89.7 (63.6)	92.1 (89.5)	90.7 (86)	91.3 (77.5)
	phasing	✓	✓	✓	✓	×	✓
MC-Careless	$CC_{1/2}$ (%)	91.4 (80.8)	96.7 (88.3)	96.2 (57.2)	98.1 (93.6)	97.8 (90.8)	94.5 (63.7)
	$CC_{F_oF_c}$ (%)	79.8 (66.6)	91 (86.3)	90.2 (62.6)	93.1 (88.8)	89.9 (84)	93.1 (75.6)
	phasing	✓	✓‡	✓	✓	×	✓

Table 3.4 continued

molecule		YT-348 (**7**)	AMG3 (**8**)	AMG4 (**9**)	AMG7 (**10**)	AMG11 (**11**)	biotin (**12**)
# crystals		17	18	21	10	8	3
XDS/XSCALE	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	90.5 (90)	93.4 (92.9)	88.4 (89.9)	99.5 (99.7)	99.9 (100)	99.6 (100)
	$CC_{1/2}$ (%)	98.5 (39.7)	99 (79.3)	98.7 (59.4)	94.1 (72.6)	97.5 (24.1)	94.6 (36.5)
	$CC_{F_oF_c}$ (%)	91 (49.6)	94.9 (87)	95.1 (88.4)	87 (87.4)	91.3 (83.5)	87 (54.1)
	phasing	✓	✓	✓	✓	✓‡	✓‡
Careless	highres bin (Å)	1.04-1	1.04-1	0.93-0.9	0.94-0.9	1.09-1.05	1.14-1.1
	Completeness (%)	90.5 (90.9)	93.8 (92.7)	88.2 (89)	100 (100)	100 (100)	100 (100)
	$CC_{1/2}$ (%)	92.4 (65.9)	98.7 (94.5)	98.5 (90.3)	90.9 (81.7)	95.3 (76.6)	96.4 (86.2)
	$CC_{F_oF_c}$ (%)	84.8 (56.6)	95.4 (86)	95.6 (90.2)	90.4 (88.8)	79.7 (79.2)	90.7 (68.5)
	phasing	✓	✓	✓	✓	✓	✓‡
MC-Careless	$CC_{1/2}$ (%)	89.7 (65.5)	98.3 (92.7)	97.2 (84.7)	92.3 (83.4)	95.2 (77.7)	97.1 (77.6)
	$CC_{F_oF_c}$ (%)	80.4 (53.9)	95.5 (87.1)	95.3 (87.6)	90.6 (85.9)	91.2 (84.2)	84.4 (70.5)
	phasing	✓	✓	✓	✓	✓	✓‡

Phasing and initial maps

Finally, we perform *ab initio* phasing on the outputs from all five merging protocols using SHELXT or SHELXD with the same phasing parameters that previously led to the preliminary solutions of the reference structures. In the microED field, *ab initio* phasing could be challenging, especially for large organic molecules lacking heavy atoms [66]. The preliminary structure from *ab initio* phasing is often corrupted by missing or extra atoms as well as mis-assignment of elements due to the difference between X-ray and electron scattering [94, 95]. Consequently, naive structural alignment with the reference structures for quantitative comparisons is difficult. Here, we manually classify phasing as successful or not by inspecting the overall connectivity or recognizable fragments for cases with disorder, and present several visual examples of the raw phasing structures.

As expected from the analysis of $CC_{F_oF_c}$ in the section above, XDS/XSCALE merging outcomes could be successfully phased for all 17 molecules regardless of dataset curation (Fig. 3.2). Even though Careless merges data with similar $CC_{F_oF_c}$ to XDS/XSCALE, the outputs could be more challenging for conventional phasing programs. In Careless, scaling and merging are jointly performed, which requires estimating structure factors independent of scaling by physical factors [8]. A consequence of this modeling approach is that structure factors are outputted on an arbitrary scale that is flat across resolution bins [8], whereas conventional programs output intensities that decay over increasing resolution. Nevertheless, correct structural information could still be retrieved from Careless outputs in most cases (Fig. 3.2). At identical contour levels, the maps from phasing are often sharper than those from XDS/XSCALE merging (Fig. 3.3), which is likely because the parallel inference of scaling and structure factors in Careless has an analogous effect to B-factor sharpening [96].

The only phasing solution from Careless outputs that contains almost no recognizable fragments is the naive merging of fischerin (Fig. 3.5-**5**) datasets [65]. Despite achieving comparable overall $CC_{F_oF_c}$ and further improvement in the highest resolution shell compared to the curated merging by Careless or either protocol of XDS/XSCALE merging (Fig. 3.6), the phasing result is visually worse (Fig. 3.7). This was a particularly difficult case where flexible molecular conformations and preferred orientation of the crystals demanded recrystallization and more than 6 months of manual processing in previous work [65].

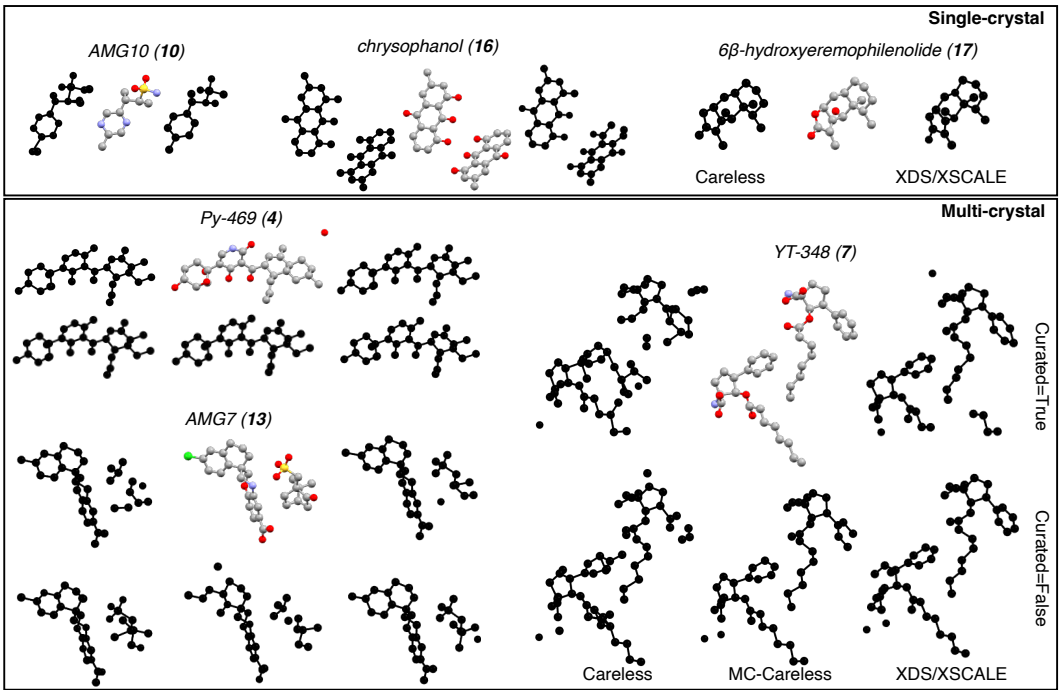

Figure 3.2: Representative examples of *ab initio* phasing outcomes (black) show that correct structural information is extracted from XDS/XSCALE and Careless merging outcomes by standard phasing programs regardless of dataset curation. Reference structures (colored by elements) are presented for comparison.

3.3 Discussion and conclusion

The successful generalization from macromolecular XRD studies [8] to small molecule microED data in this work highlights the flexibility and impact of Bayesian inference in emerging structural studies. Through benchmarking against XDS/XSCALE, we also show that this approach has some benefits in merging small molecule microED data. Careless merges reflections to higher $CC_{1/2}$ at high resolution, and comparable accuracy with respect to the reference structure is achieved both overall and at high resolution. Moreover, for the examples presented here, dataset curation, whether manual or automated, is not necessary for Careless, as the naive merging of all datasets achieves the best $CC_{1/2}$, comparable $CC_{F_oF_c}$, and the highest completeness. Thus, merging by Careless eliminates an opportunity for human bias in data processing and maximally leverages information from all datasets. Even though automated curation by MC-Careless does not further improve merging outcomes, it shows that Careless could be easily extended for future methods development.

For microED practitioners, we caution against the common practice of manually curating datasets. We find that $CC_{1/2}$ is elevated in XDS/XSCALE merging using dataset curation inherited from previous work, yet we see no significant differences

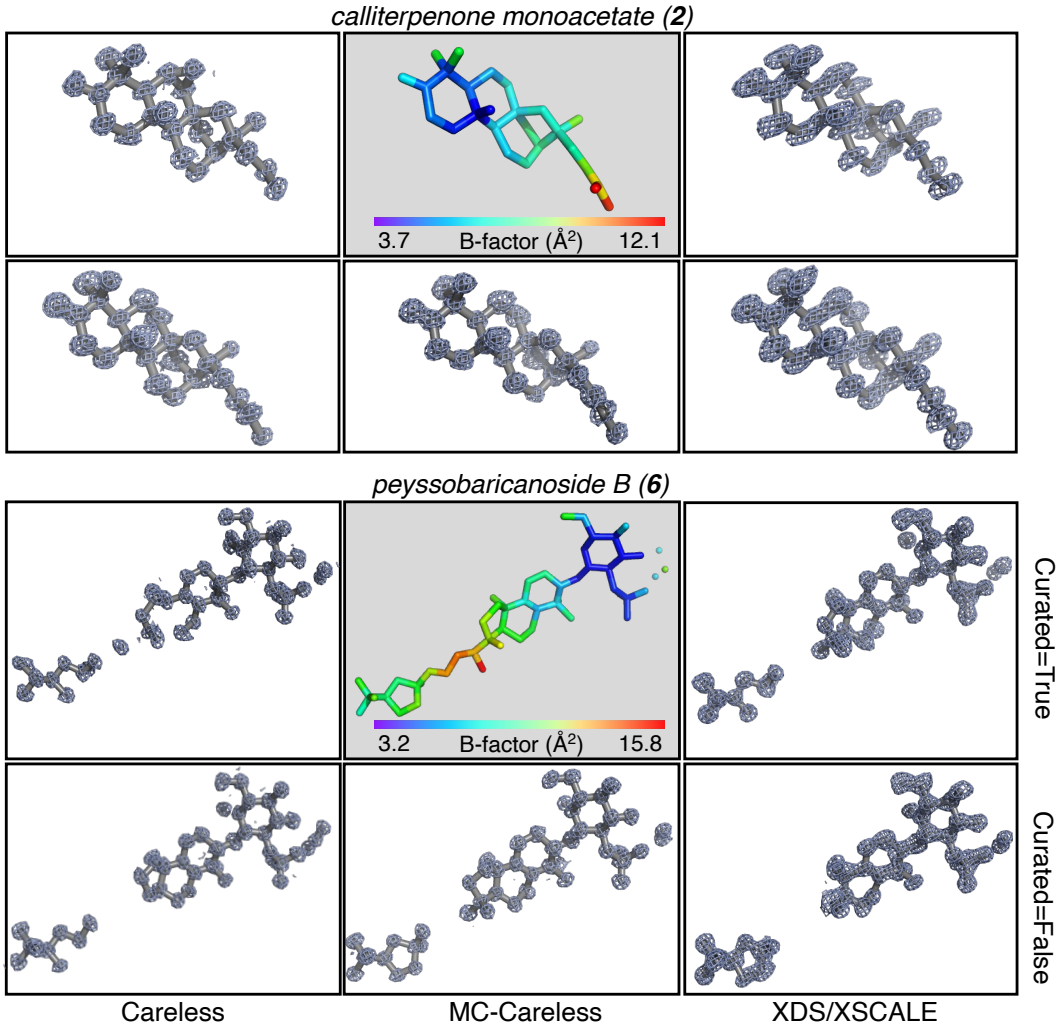

Figure 3.3: Data merged by Careless and MC-Careless yield sharper $2F_o - F_c$ maps after *ab initio* phasing. Example *ab initio* structures and maps from all 5 merging protocols are shown for calliterpenone acetate (left) and peyssobaricanoside B (right). Reference structures are colored by atomic B-factors to show flexible parts where phasing might be challenging.

in $CC_{F_oF_c}$ or the *ab initio* phasing structures. Our findings suggest that data quality is not always compromised when naively merging all datasets, indicating that useful signal can be missed in manual curation of datasets. For example, the three lowest overall $CC_{F_oF_c}$ from molecules Py-469 (Fig. 3.5-**4**) [65], demethoxyviridin (Fig. 3.5-**1**) [69], and AMG7 (Fig. 3.5-**10**) [90] are improved by more than 8% when including all datasets in XDS/XSCALE merging (Tables 3.2 and 3.4). From a practical perspective, the preliminary structures from *ab initio* phasing seem robust against dataset curation, although the impact on refinement statistics is beyond the scope of this work.

In conclusion, using experimental datasets from previous studies, we demonstrate that Careless could robustly merge microED and small molecule crystallography data. Careless could improve multi-crystal merging outcomes with reduced human bias and is a flexible framework for methods development in diffraction data processing. In most cases examined here, Careless outputs lead to similar preliminary structural solutions with sharpened initial maps compared to XDS/XSCALE outputs. For challenging cases, additional optimization of merging and phasing parameters might be necessary to obtain the correct phasing solutions. As Bayesian inference and other ML approaches have only recently been introduced to crystallography, continuing method developments are warranted. Additional case studies and future investigation in refinement outcomes may help improve the integration of Careless into existing data processing pipelines.

3.4 Methods

Merging algorithms

To contextualize the ML approach in Dalton *et al.* [8] and our extension to it to automate dataset curation in Careless, we briefly describe the formalism of scaling and merging in crystallography. Readers are referred to Aldama *et al.* [9] for a more detailed review.

Merging by weighted average

Conventionally, the true intensity $I_{\mathbf{h}}$ at Miller index $\mathbf{h}$ is estimated by computing the weighted average of redundant measurements across all images after correcting for systematic errors. This corresponds to the maximum likelihood estimate of the mean intensity. The functional form of the weights w determines the error model with normally-distributed being the most common choice.

Each measurement $\hat{I}_{\mathbf{h},i}$ on image i is corrected by estimating a scaling factor $K_{\mathbf{h},i}$:

$$\hat{I}_{\mathbf{h},i} = K_{\mathbf{h},i} I_{\mathbf{h},i}. \tag{3.1}$$

Established programs in XRD data processing such as XDS [75, 93] and AIMLESS [97] use sophisticated models to parameterize $K_{h,i}$ and minimize the least-squares loss for scaling and merging:

$$\Phi = \sum_{\mathbf{h}} \sum_{i} w_{\mathbf{h},i} (I_{\mathbf{h}} - \hat{I}_{\mathbf{h}} / K_{\mathbf{h},i})^2. \tag{3.2}$$

Merging by Bayesian inference

Careless works on the same premise in eq. (3.1) that systematic errors can be corrected by scaling $I_{\mathbf{h},i}$, but uses an alternative inference approach. Under the kinematical approximation, the true intensity is $I_{\mathbf{h},i} = F_{\mathbf{h}}^2$, where F denotes the structure factor amplitude. Careless uses variational Bayesian inference [98, 99] to reformulate merging as estimating $p(F, K|I)$, the posterior distribution of the scaling and structure factors conditioned on the observed intensities. As $p(F, K|I)$ is generally intractable, it is approximated by a parametric surrogate function q. The standard modeling objective in variational Bayesian inference is to minimize the difference between q and $p(F, K|I)$ by maximizing the Evidence Lower BOund (ELBO) [100]. The ELBO typically consists of an expected log-likelihood term that encourages fitting q to the data and a Kullback–Leibler (KL) term as a regularization to penalize deviations from a prior distribution. The exact form used in Dalton *et al.* [8] is:

$$\text{ELBO} = \mathbb{E}_q \left[\log p(I|F, K)\right] - D_{\text{KL}} \left[q_F \| p(F)\right] \tag{3.3}$$

where q is assumed to be factorizable,

$$p(F, K|I) \simeq \prod_{\mathbf{h}} \left[q_{F_{\mathbf{h}}} \prod_{i} q_{K_{\mathbf{h},i}} \right], \tag{3.4}$$

and q_K is further parameterized by a multi-layer perceptron (MLP) that takes the metadata of observed reflections as the input. Parameters of q_K are optimized without regularization from a prior distribution. A modified version of the Wilson distribution — the intensity distribution if atoms are uniformly distributed within the unit cell [10] — is used as the prior $p(F)$ for estimating the structure factor amplitudes independent of the scale [8].

Extending Careless for multi-crystal merging

The uncertainty of the observed intensity, σ_I, is important for estimating data quality [97, 101] and directly affects merging in the conventional approach as described in eq. (3.2). In the variational Bayesian inference approach, σ_I also modulates the contribution of each measurement to the training loss. Specifically, the log-likelihood term in eq. (3.3) is a parametric distribution where the mean is $\hat{K}\hat{F}^2$ obtained from drawing Monte Carlo samples $\hat{F} \sim q_F$ and $\hat{K} \sim q_K$, and the standard deviation or scale in the case of Student's t-distribution, is the σ_I estimated by integration programs.

In MC-Careless, to account for the different quality of each dataset in multi-crystal merging, we adjust the uncertainty of observed intensities σ_I inversely by a weight w that is sparsely parameterized as a categorical distribution q_w over the N crystals to merge (Fig. 3.4). The distribution is normalized such that w averages to 1 across all unmerged reflections. The contribution of a crystal to merging is decreased as w becomes smaller than 1, consequently increasing the uncertainty of observed intensities from that crystal. The modified training objective is:

$$\text{wELBO} = \mathbb{E}_q \left[\log p \left(I | F, K; \frac{\sigma_I}{w} \right) \right] - D_{\text{KL}} \left[q_F \| p(F) \right] - D_{\text{KL}} \left[q_w \| p(w) \right] \quad (3.5)$$

where w is learned jointly with F and K, and is regularized by a prior distribution $p(w)$. The prior distribution is the discrete uniform distribution that represents no adjustment to σ_I and equal weights among crystals as treated in the naive merging of all input datasets. Code that implements MC-Careless is available at `https://github.com/DorisMai/careless/tree/multi_xtal_sig`.

Data processing workflow with XDS/XSCALE

Each rotational diffraction movie was collected in SER format and converted to SMV as previously described [102]. Spot finding, indexing, integration, and correcting/scaling are performed using XDS [75]. XDS is a standard crystallography program that has proven effective for small molecule microED data [63], although other programs such as DIALS [103, 104], Jana2020 [105], and CrysAlis$^{\text{Pro}}$ [106] are also applicable. The instruction file for initial processing by XDS is generated using an in-house Python script (`https://github.com/jess-burch/microed`) for greater automation as previously described [69, 90]. To benchmark merging performance with minimal confounding errors from other processing steps, here all datasets are reprocessed by XDS using previously reported space group and unit cell parameters. XSCALE [75, 93] is used to merge data from multiple crystals. The

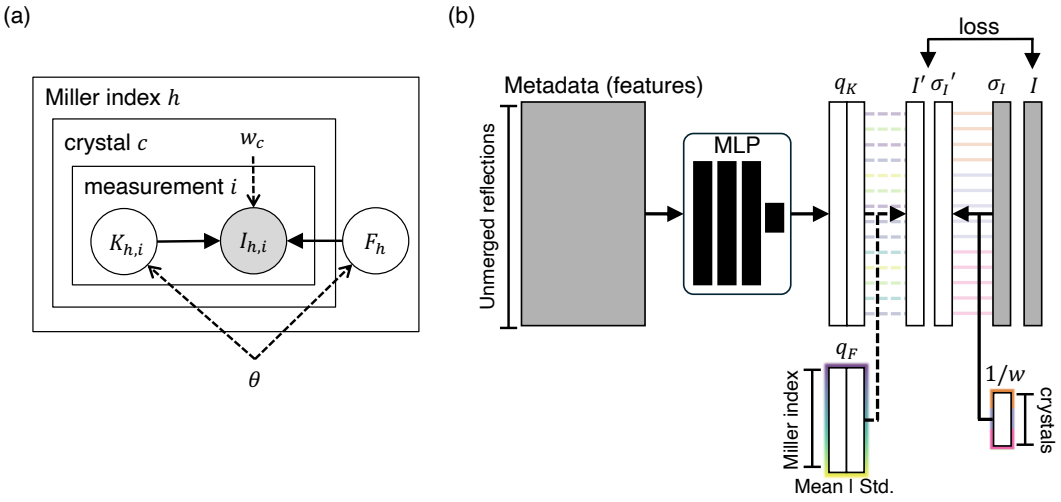

Figure 3.4: Schematic of multi-crystal extension to Careless. (a) Probabilistic graphical model of merging diffraction data using variational Bayesian inference algorithm. Solid lines denote the generative process of the observed intensity I from the scaling factor K and structure factor amplitude F. Dashed lines represent variational Bayesian inference of K and F parameterized by model parameters θ, where the uncertainty of I is adjusted by a per-crystal weight w. (b) VI model architecture. Posterior estimations of K and F are variationally approximated as q_K and q_F, and q_K is further parameterized by an MLP transformation from the metadata of unmerged reflections. Dashed lines denote the reparameterization process, where samples $\hat{K}$ and $\hat{F}$ are drawn from q_K and q_F to compute the loss between observed intensity I and predicted intensity $\hat{I} = \hat{K}\hat{F}^2$ with adjusted uncertainty σ_I/w.

resolution cutoff from previous work is used whenever possible in the reprocessing but relaxed by 0.05 Å for 6β-hydroxyeremophilenolide (Fig. 3.5-**17**), calcium oxalate (Fig. 3.5-**15**), and peyssobaricanoside B (Fig. 3.5-**6**), by 0.1 Å for AMG3 (Fig. 3.5-**8**) and 4 (Fig. 3.5-**9**), and by 0.15 Å for AMG7 (Fig. 3.5-**10**) to reproduce phasing outcomes.

Data processing workflow with Careless

Data preprocessing

XDS_ASCII.HKL files from reprocessing as described above are converted to .mtz format using `careless.xds2mtz` before merging. Each dataset is then standardized such that the intensity I has unit variance. This is achieved by scaling the observed intensities $I'_{\mathbf{h},i}$ and uncertainty $\sigma'_{I_{\mathbf{h},i}}$ by k, where

$$k = \frac{1}{\sqrt{\frac{\sum_{\mathbf{h},i}\left(I'_{\mathbf{h},i} - \bar{I}'\right)^2}{N_{\text{unmerged}}}}}. \tag{3.6}$$

This standardization supports stable training. Unit cell parameters are averaged across all crystals to be merged.

Model training

Metadata features used for model training include the image number, resolution, and X/Y positions of each observed intensity on the image. For multi-dataset merging, intensities from all datasets are concatenated, and the index of the source dataset is supplied as an additional feature. A non-negative scaling factor is enforced during model training. Training steps are 30,000 and 50,000 for single-crystal and multi-crystal merging, respectively. Training for all cases can run on an NVIDIA Tesla P100 GPU in under 1.5 hours, with the exception of merging all 89 fischerin (Fig. 3.5-**5**) datasets which could take 7.5 hours.

Hyperparameter selection

The original training objective of Careless described by eq. (3.3) is approximated using Monte Carlo sampling with 1 sample per training step as the default:

$$\text{ELBO} \approx \sum_{s=1}^{S}\left[\sum_{i}\sum_{\mathbf{h}}\log p\left(I_{i,\mathbf{h}}|F_{\mathbf{h}},K_{i,\mathbf{h}},\sigma_I;\nu\right)-\sum_{\mathbf{h}}\left(\log q_F-\log p\left(F\right)\right)\right]. \tag{3.7}$$

We increase to $S = 20$ Monte-Carlo samples to improve convergence with a minimal increase in total training time. The log-likelihood term in eq. (3.7) is modeled as Student's t-distribution with the degree of freedom ν as a hyperparameter to adjust the sensitivity to outliers. This error model becomes a normal distribution as ν approaches infinity and becomes a Cauchy distribution when $\nu = 1$. We keep $\nu = 16$, which was found to be optimal by cross-validation in Dalton *et al.* [8] and empirically robust by other users of Careless.

The relative weight between the log-likelihood term and the KL term in eq. (3.7) defaults to the average multiplicity of the datasets $m = N_{\text{unmerged}}/N_{\text{merged}}$. As of Careless version 0.3.4, this weight is adjustable through the hyperparameter λ_F. Empirically, we find that a small value of λ_F generally works well, possibly because observed intensities in small molecule 3D ED do not always obey ideal statistics described by the Wilson distribution (Fig. 3.8) which is used as a prior distribution in Careless as described in eq. (3.3). In this work, $\lambda_F = 0.01$ is used for all cases, except for fischerin (Fig. 3.5-**5**) datasets where the optimal value between 0.01 and 0.001 is chosen by cross-validation.

In MC-Careless, we also introduce λ_w to adjust the relative weight of the second KL term in eq. (3.5) such that:

$$\begin{aligned} \text{wELBO} \approx \sum_s \Bigg[& \sum_i \sum_{\mathbf{h}} \log p \left(I_{i,\mathbf{h}} | F_{\mathbf{h}}, K_{i,\mathbf{h}}, \frac{\sigma_I}{w_{i,\mathbf{h}}}; \nu \right) \\ & - m\lambda_F \sum_{\mathbf{h}} \left(\log q_F - \log p\left(F\right) \right) \\ & - m\lambda_w \sum_i \sum_{i,\mathbf{h}} \left(\log q_w - \log p\left(w\right) \right) \Bigg]. \end{aligned} \tag{3.8}$$

The optimal value of λ_w is found over 0.001, 0.01, 0.1, 1, and 10 by cross-validation.

Evaluation of merging outcomes

Merging quality is assessed based on the following metrics: completeness, $CC_{1/2}$, and $CC_{F_oF_c}$. Both the overall statistic and the statistic in the highest resolution bin are reported. Resolution bins are determined by default in XDS/XSCALE and in Careless to distribute reflections evenly across 10 bins. The completeness and $CC_{1/2}$ are extracted from CORRECT.LP (single-crystal) or XSCALE.LP (multi-crystal) for XDS/XSCALE outputs and from `careless.completeness` and `careless.cchalf` for Careless outputs.

The $CC_{F_oF_c}$ metric is calculated as the uncertainty-weighted Pearson correlation coefficient between the estimated structure factor amplitude F_o from merging and the F_c calculated from the reference structures. The CCDC numbers of reference structures are available in Table 3.5 of the supplementary information. F_c is calculated using `gemmi`, accounting for electron form factors and anisotropic atomic displacement parameters. An additional non-negative global B-factor is fit when calculating $CC_{F_oF_c}$ because Careless outputs F_o on the same scale across resolution bins, unlike XDS/XSCALE and other conventional data reduction programs.

Preliminary structures from *ab initio* phasing

Intensities merged by XDS/XSCALE are converted to SHELX format using the XDSCONV program. Intensities merged by Careless are scaled and reformatted by a separate Python script that uses `reciprocalspaceship` [107] to parse the .mtz output file from Careless. *Ab initio* phasing is then performed using SHELXT [108] or SHELXD [109], using parameters from previous work that led to the reference structures. We run SHELXD on 32 CPUs for 15 minutes for all cases except for fischerin (Fig. 3.5-**5**) where the run time is extended to 1 hour. $2F_o - F_c$ maps from

ab initio phasing are generated using the `shelx2map` program, and visualized in Pymol [110] with contouring at 1.5σ and carving at 1.2Å.

3.5 Data availability

MicroED data used in this work are available at `10.5281/zenodo.12775590`, `10.5281/zenodo.12797270`, `10.5281/zenodo.8206533`, `10.5281/zenodo.10059796`, `10.5281/zenodo.10059842`, and `10.5281/zenodo.10059864`, except for AMG3, AMG4, AMG7, AMG10, and AMG11, which are available upon request. Specific datasets used in single-crystal merging and manually curated multi-crystal merging are described in curated_movie_id.csv in `10.5281/zenodo.12775590`.

Acknowledgements

H.M. thanks Douglas C. Rees, Michael R. Sawaya, Jose A. Rodriguez, William M. Clemons, Stephen L. Mayo, and Vignesh C. Bhethanabotla for helpful discussions, Dmitry B. Eremin, Kunal K. Jha, and David A. Delgadillo for feedback on the manuscript, and David A. Delgadillo, Lee Joon Kim, and Christopher G. Jones for sharing raw microED data. This work is sponsored by the NSF Center for Computer-Assisted Synthesis, an NSF Center for Chemical Innovation (CHE-2202693). This work also used computational resources from the Resnick High Performance Computing Center, a facility supported by Resnick Sustainability Institute at the California Institute of Technology.

3.6 Appendix

Table 3.5: Space group, unit cell information, and CCDC numbers.

molecule	space group	volume* (Å^3)	unit cell lengths (Å) and angles (°)† curated	all	CCDC
1	18	1514.2	a = 21.34 ± 0.03 b = 6.60 ± 0.06 c = 10.83 ± 0.06	a = 21.57 ± 0.24 b = 6.56 ± 0.05 c = 10.94 ± 0.24	2246158
2	5	2018.5	a = 13.84 ± 0.04 b = 6.37 ± 0.04 c = 23.70 ± 0.26 β = 105.70 ± 0.29	a = 13.85 ± 0.13 b = 6.39 ± 0.05 c = 23.73 ± 0.23 β = 106.07 ± 0.63	2246153
3	14	1084.9	a = 3.83 ± 0.02 b = 12.74 ± 0.01 c = 22.22 ± 0.03 β = 92.08 ± 0.54	a = 3.83 ± 0.02 b = 12.75 ± 0.02 c = 22.25 ± 0.04 β = 92.01 ± 0.50	2246159
4	20	4838.6	a = 5.24 ± 0.01 b = 26.81 ± 0.16 c = 34.59 ± 0.04	a = 5.24 ± 0.03 b = 26.92 ± 0.18 c = 34.61 ± 0.35	2038723
5	5	14266	a = 35.52 ± 0.63 b = 20.34 ± 0.29 c = 18.59 ± 0.16 β = 96.37 ± 0.21	a = 35.55 ± 0.58 b = 20.37 ± 0.31 c = 18.63 ± 0.51 β = 95.96 ± 0.90	2020516
6	19	4100.8	a = 7.05 ± 0.07 b = 10.82 ± 0.04 c = 53.60 ± 0.27	a = 7.13 ± 0.15 b = 10.96 ± 0.29 c = 55.21 ± 3.47	2251540
7	4	2069.8	a = 17.79 ± 0.12 b = 5.73 ± 0.04 c = 21.89 ± 0.30 β = 113.31 ± 0.62	a = 17.79 ± 0.11 b = 5.73 ± 0.04 c = 21.77 ± 0.31 β = 113.09 ± 0.59	2332145
8	18	4029.9	a = 23.04 ± 0.03 b = 37.78 ± 0.42 c = 4.61 ± 0.01	a = 23.11 ± 0.13 b = 38.25 ± 0.45 c = 4.62 ± 0.02	2116696
9	18	3972.5	a = 23.05 ± 0.00 b = 37.99 ± 0.27 c = 4.63 ± 0.00	a = 23.03 ± 0.10 b = 38.00 ± 0.45 c = 4.63 ± 0.02	2116691
10	4	1291.4	a = 10.55 ± 0.10 b = 10.19 ± 0.02 c = 12.46 ± 0.18 β = 110.21 ± 0.72	a = 10.54 ± 0.07 b = 10.19 ± 0.05 c = 12.56 ± 0.15 β = 110.30 ± 0.47	2116689
11	5	995.4	a= 21.91 ± 0.04 b = 6.44 ± 0.02 c = 7.01 ± 0.00 β = 90.95 ± 0.01	a = 21.83 ± 0.13 b = 6.42 ± 0.02 c = 7.05 ± 0.03 β = 90.88 ± 0.19	2116692

molecule	space group	volume* (Å^3)	unit cell lengths (Å) and angles (°)* curated	all	CCDC
12	19	1121	a = 5.25 b = 10.40 c = 20.98	a = 5.29 ± 0.03 b = 10.47 ± 0.04 c = 21.23 ± 0.18	1876036
13	19	983.6	a = 7.46 b = 8.13 c = 16.25	N/A	2116687
14	19	864.2	a = 4.94 b = 9.03 c = 19.2	N/A	2246165
15	87	1130.5	a = 12.36 b = 12.36 c = 7.40	N/A	2246152
16	19	2253.9	a = 3.92 b = 23.25 c = 24.73	N/A	2246157
17	20	2711	a = 7.12 b = 13.36 c = 28.50	N/A	2246154

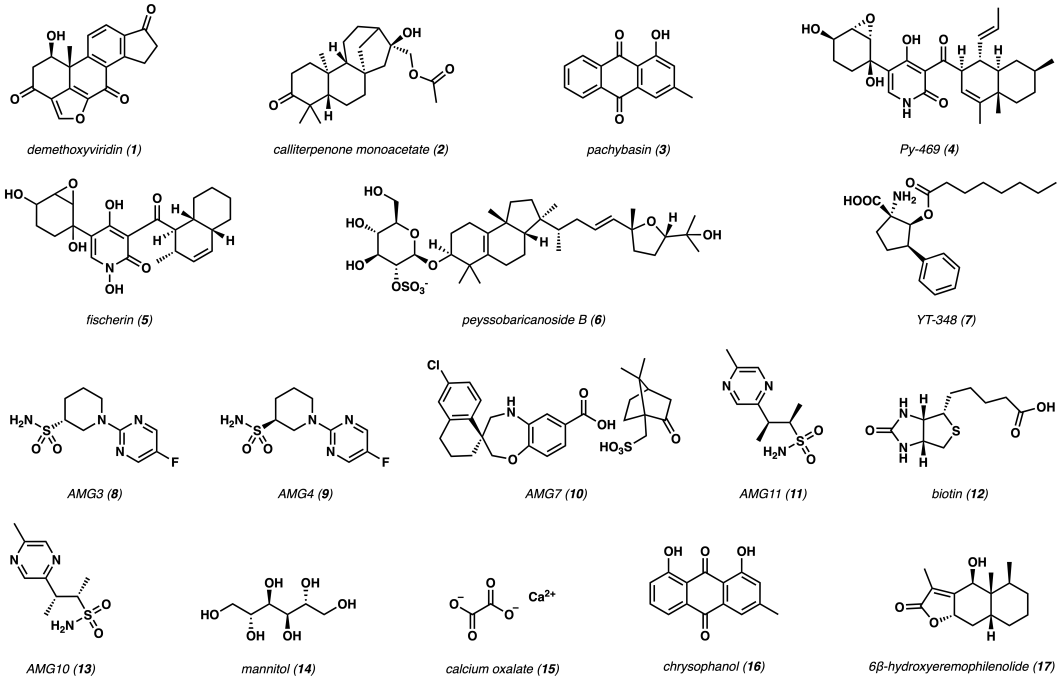

Figure 3.5: Chemical structures of all 17 molecules.

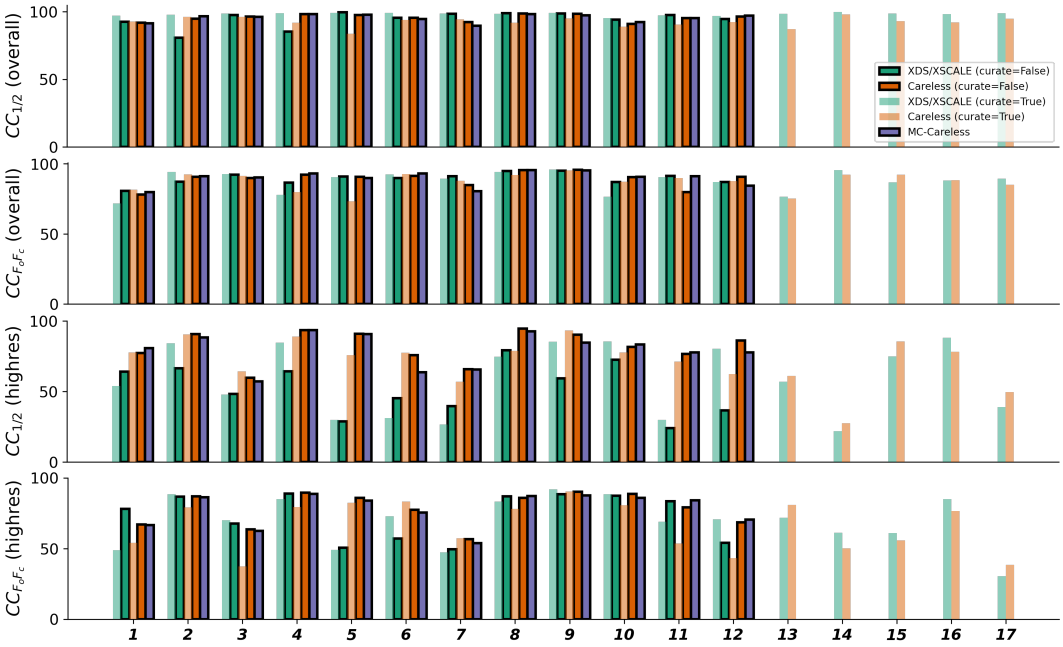

Figure 3.6: Comparison of all merging protocols on each molecule. Single-crystal cases are plotted as curated merging.

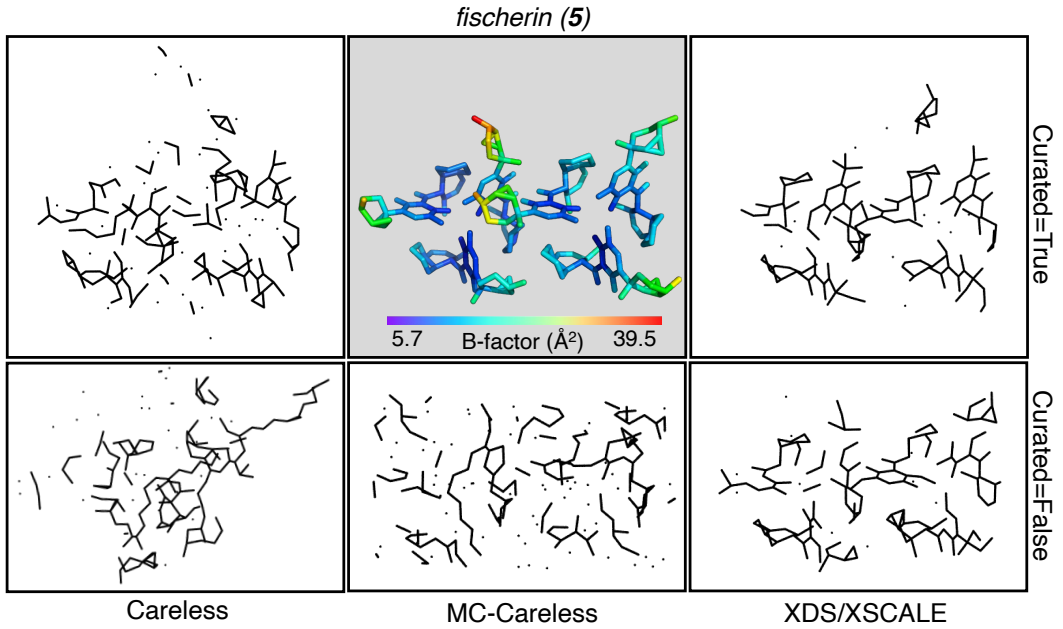

Figure 3.7: *Ab initio* phasing outcomes for fischerin datasets. The preliminary structures are rotated and shifted manually to align with the reference structure (bottom right) for visual comparison, except for the naive merge using Careless, where the phasing quality is too poor to perform structural alignment.

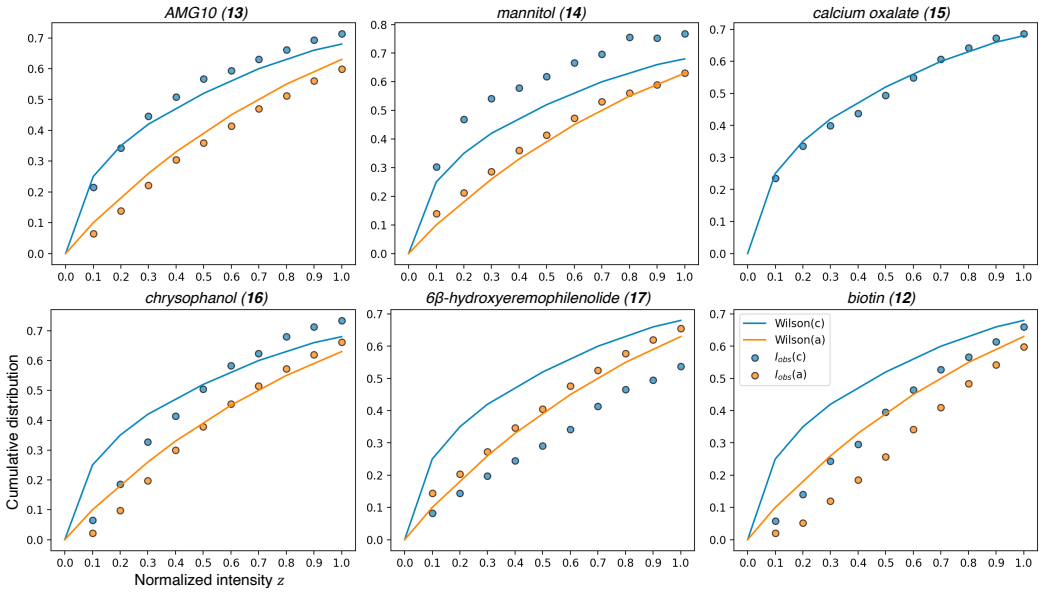

Figure 3.8: Cumulative distributions of normalized intensities (I_{obs}) for centric (c) and acentric (a) reflections in single-crystal datasets. Values are extracted from CORRECT.LP in XDS and plotted against the ideal values derived from the Wilson distribution.

Chapter 4

EXPLORING PROTAC COOPERATIVITY WITH COARSE-GRAINED ALCHEMICAL METHODS

Adapted from

(1) Mai, H.; Zimmer, M. H.; Miller, T. F. Exploring PROTAC Cooperativity with Coarse-Grained Alchemical Methods. *The Journal of Physical Chemistry B* **2023**, *127*, 446–455. DOI: 10.1021/acs.jpcb.2c05795.

Abstract

Proteolysis targeting chimera (PROTAC) is a novel drug modality that facilitates the degradation of a target protein by inducing proximity with an E3 ligase. In this work, we present a new computational framework to model the cooperativity between PROTAC-E3 binding and PROTAC-target binding principally through protein-protein interactions (PPIs) induced by the PROTAC. Due to the scarcity and low resolution of experimental measurements, the physical and chemical drivers of these non-native PPIs remain to be elucidated. We develop a coarse-grained (CG) approach to model interactions in the target-PROTAC-E3 complexes, which enables converged thermodynamic estimations using alchemical free energy calculation methods despite an unconventional scale of perturbations. With minimal parameterization, we successfully capture fundamental principles of cooperativity, including the optimality of intermediate PROTAC linker lengths that originates from configurational entropy. We qualitatively characterize the dependency of cooperativity on PROTAC linker lengths and protein charges and shapes. Minimal inclusion of sequence- and conformation-specific features in our current forcefield, however, limits quantitative modeling to reproduce experimental measurements, but further development of the CG model may allow for efficient computational screening to optimize PROTAC cooperativity.

4.1 Introduction

Proteolysis targeting chimera (PROTAC) has emerged as a promising drug modality that elicits protein degradation by hijacking the ubiquitin-proteasome system (UPS), a major regulatory component of cells. In the UPS pathway, E3 ligases transfer ubiquitins onto aberrant proteins to mark them for degradation by proteasomes. A PROTAC molecule exploits this pathway with two binding moieties that tether the target protein and an E3 ligase together. The tethered target protein thus becomes a neo-substrate of the E3 ligase and is subsequently ubiquitinated for proteasomal degradation. PROTACs require a lower dose than conventional small-molecule inhibitors because of their catalytic nature and they have the potential to target the undruggable proteome [111, 112]. Since the first proof-of-concept in 2001 [113], the number of proteins successfully degraded by PROTACs has grown rapidly, and examples of such proteins include kinases and gene regulators that are implicated in cancer. As of 2021, at least 13 PROTACs are in or approaching clinical trials [114].

Despite increasing applications, there is a lack of guidance on designing PROTACs due to the unique mode of action [115–117]. In particular, a critical step in the degradation process is the formation of the ternary complex of target-PROTAC-E3. The ternary complex involves molecular interactions beyond the binary bindings between the two warheads of a PROTAC and the two proteins. The selectivity [118–120] and stability [121–124] of the ternary complex can both be improved through favorable protein-protein interactions (PPIs) between the target protein and the E3 ligase. For certain targets, the degradation outcome can be very different depending on whether cereblon (CRBN) or von Hippel-Lindau (VHL), the two most heavily used E3 ligases, more efficiently and selectively form a productive complex with the target [121, 125–127]. As more warheads for E3 ligases are designed [128–131], choosing which of the more than 600 E3 ligases in humans [132] optimally interact with the target protein will become important [133, 134]. While PPIs depend on the sequences and the structures of the proteins, PROTACs can also modulate the PPIs by restricting the distance and relative orientation between the target and the E3 ligase, effectively changing the entropic component of PPIs.

Because of this three-body interplay and the transient nature of the ternary complex, a complete characterization of the PPIs as a function of the PROTAC, the target protein, and the E3 ligase is intractable. A few proteomics studies [126, 127, 135] on kinase degradation have used PROTACs with promiscuous warheads such that the PROTAC-induced PPIs differentially affect the degradation outcome of hundreds

of proteins. These studies reported the fold change of protein abundance due to PROTAC treatment, but analysis can be complicated by secondary interactions [134] and numerous other factors such as the permeability of the PROTAC, half-lives of the target proteins, cellular localization, and reactions downstream of ternary complex formation [136]. Other studies [118, 119, 137–140] have focused on specific target-E3 pairs and examined the effect of changing PROTAC properties such as the linker length. They measured the difference in the strength of PROTACs binding to the target or the E3 ligase due to the presence of the other protein. This difference, termed binding cooperativity, reflects the strength of PROTAC-mediated PPIs. However, few generalizable patterns have emerged and systematic experimental characterizations remain scarce.

Computational modeling based on docking or atomistic molecular dynamics (MD) has complemented experimental work [119, 139] and displayed promising prospects, but there are several limitations to current methodologies. Although standard docking protocols don't handle three-body problems, several workflows have been adapted ad hoc for PROTAC [141–145]. Docking studies rank ternary complex conformations by scoring functions biased for naturally evolved PPIs and benchmark against the few crystal structures of PROTAC-induced ternary complexes [146–148]. The results can be inaccurate as PROTAC-induced PPIs are non-native and exhibit plasticity [119, 149]. In contrast, atomistic MD is physically grounded to capture non-native PPIs. However, the size of the ternary complex modeled at an atomistic resolution significantly limits the timescale of simulations, such that naively simulating PPIs can be prohibitively slow. Sophisticated enhanced sampling techniques and distributed computing are needed to sample an ensemble of low-energy conformations that are consistent with experimental data [150]. Due to the difficulties in modeling the ternary complex, direct calculation of the binding cooperativities was not attempted until two recent studies [151, 152] that explored the molecular mechanics with the generalized Born and surface area continuum solvation (MM/GBSA).

Here, we seek an orthogonal approach that combines coarse-grained MD (CGMD) and alchemical free energy calculation methods to study PROTAC cooperativities. On the spectrum of computational tools, docking and atomistic MD are positioned at the empirical and first-principle ends, respectively, and finding a compromise in the middle of this spectrum is a promising direction. Compared to atomistic modeling, coarse-graining reduces the effective size of the model and smoothens the energy

surface, enabling simulations at a much longer timescale necessary for the PROTAC-mediated complexes. While CGMD may struggle to recapitulate the molecular basis of lock-and-key bindings, such strong and specific interactions are less imperative in non-native PPIs induced by PROTACs. Moreover, PROTAC binding reduces the ways proteins can interact with each other, differentiating and simplifying the problem studied here from the formidable task of modeling general protein-protein binding. In docking, such constraints are challenging to incorporate into the scoring functions and are approximated through separate steps to filter compatible PPI poses and PROTAC geometries. While CGMD excludes many degrees of freedom from the PROTAC, proteins, and solvent entropy, this effect of configurational entropy on PPIs from PROTAC mediation can be directly captured. Finally, we calculate binding energies using alchemical methods, which circumvents the computational challenge of directly sampling binding and unbinding events between the PROTAC and proteins. We demonstrate the computational amenity of an unconventional application of alchemical methods motivated by the PROTAC systems, and take advantage of the physical interpretability of the CGMD + alchemical approach to explore the principles of PROTAC binding cooperativity.

4.2 Methods

CGMD setup of PROTAC-protein complexes

The binary and ternary PROTAC-protein complexes are coarse-grained at two resolutions to efficiently sample complex conformational changes while retaining sufficient details for structural insight. Specifically, a major focus of this work is to characterize the entropic effect of the length of PROTACs on the strength of induced PPIs, necessitating modeling the PROTAC linker at a higher resolution than the rest of the system. Proteins are coarse-grained by mapping every three amino acids onto a large bead of $\sigma = 0.8$ nm diameter, which is approximately the Kuhn length of polypeptides [153–156]. Binding moieties at the two ends of a PROTAC are each represented by a large bead, whereas the linker region is modeled as a Gaussian chain at the resolution of a PEG unit ($\sigma_s = 0.35$ nm [157]) or three heavy atoms. Several experimental works that used flexible linear linkers motivate our modeling approach for the PROTAC linker, including Chan *et al.* [138] where an alkane linker was varied in step sizes of our linker beads and Zorba *et al.* [139] where a PEG linker is modified at smaller length steps such that linker lengths ranging from 1 to 6 σ_s in our modeling correspond to the PROTAC (1), (3), (5), (6), (8), and (10).

A minimal forcefield is used to describe the internal and interactive forces, and a full description can be found in the Supporting Information (Section 4.5). The three-dimensional structure of a protein is maintained by a bottom-up fitted elastic network model (Fig. 4.6), which allows conformational flexibility [158, 159]. Protein beads can have additional properties to describe PPIs beyond volume exclusion (Fig. 4.5). When modeling electrostatic interactions, for example, a protein bead has the net charges of the triplet of residues that it is coarse-grained from. PROTACs are modeled as Gaussian polymers with volume exclusion, and the warhead beads are attached to the binding pockets of proteins through harmonic springs. Modeling PROTAC interactions beyond warhead binding is out of the scope of this work. Thus, under current setup, PROTAC beads have 0 charge and no affinity to any other beads.

The orientation between the E3 ligase and the target protein is initialized such that the two binding pockets face each other, with a fully extended PROTAC tethering in between (Fig. 4.1a). The binding moiety beads of PROTAC are placed at the center of each binding pocket, which is defined by the residues within 4 or 5 Å from the PROTAC warhead in experimental structures. Thus, setting up the initial coordinates of a ternary complex requires the following inputs: structures of each protein, residues at the two PROTAC binding pockets, and the length of the PROTAC linker. To calculate the difference in PROTAC binding energies due to PPIs, simulations of binary target/E3-PROTAC complexes are also needed. Binary complexes are prepared by removing a protein from the initialized ternary complex.

Thermodynamic framework of alchemical perturbation

The binding cooperativity of a PROTAC is mathematically defined as $\exp\left(\frac{\Delta\Delta G}{RT}\right)$, where R is the gas constant, T here refers to the temperature in the context of an energetic scale and refers to the target protein elsewhere, $\Delta\Delta G = \Delta G_{TP}^{\text{binary}} - \Delta G_{TP}^{\text{ternary}}$, and $\Delta G_{TP}^{\text{ternary}}$ and $\Delta G_{TP}^{\text{binary}}$ are the free energies of the PROTAC (P) binding to the target protein (T) with and without the presence of the E3 ligase (E). Because of the thermodynamic cycle (Fig. 4.1b), the same $\Delta\Delta G$ can be obtained from $\Delta G_{EP}^{\text{binary}} - \Delta G_{EP}^{\text{ternary}}$. Favorable PPIs stabilize the ternary complex and facilitate PROTAC binding to both proteins. Thus, they lower $\Delta G_{TP}^{\text{ternary}}$ and $\Delta G_{EP}^{\text{ternary}}$, which leads to larger $\Delta\Delta G$ and more positive cooperativity.

Alchemical free energy calculation methods exploit alternative thermodynamic cycles to obtain $\Delta\Delta G$ without simulating binding and unbinding processes. For

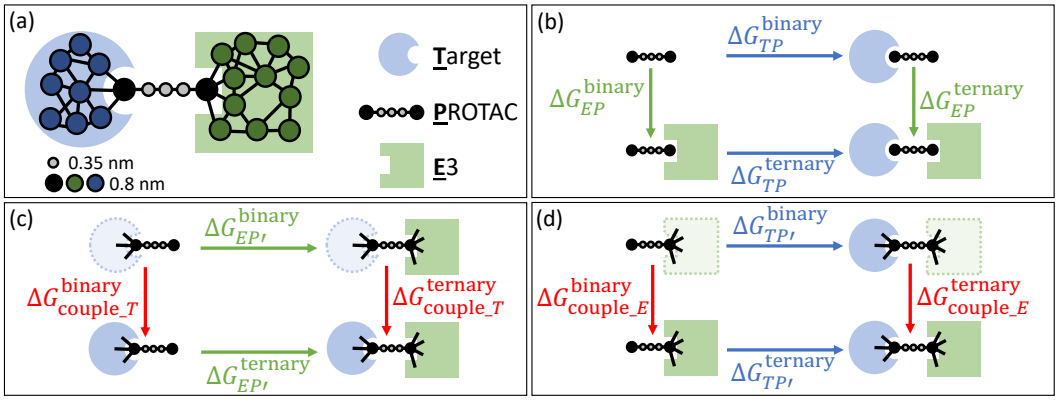

Figure 4.1: Schematic of the simulation setup for PROTAC-mediated complexes. **(a)** The target-PROTAC-E3 ternary complex is initialized with a fully extended PROTAC as drawn. The proteins are coarse-grained at the resolution of three amino acids per bead, approximately 0.8 nm. PROTAC warhead beads are represented by beads of the same size, whereas the linker is coarse-grained at a higher resolution. **(b)** PPIs affect how cooperative target-PROTAC and PROTAC-E3 bindings are and are reflected in the free energy difference between PROTAC-E3 binding with and without the target ($\Delta G_{EP}^{\text{binary}} - \Delta G_{EP}^{\text{ternary}}$). This free energy difference, $\Delta\Delta G$, can also be obtained by comparing target-PROTAC binding with and without the E3 ($\Delta G_{TP}^{\text{binary}} - \Delta G_{TP}^{\text{ternary}}$) as shown by the thermodynamic cycle. Under the alchemical setup, $\Delta\Delta G$ can be alternatively obtained by the free energy difference between the red vertical processes, which represent coupling the target ($\Delta G_{\text{couple\_T}}^{\text{binary}} - \Delta G_{\text{couple\_T}}^{\text{ternary}}$ in **(c)**) or the E3 ($\Delta G_{\text{couple\_E}}^{\text{binary}} - \Delta G_{\text{couple\_E}}^{\text{ternary}}$ in **(d)**) to the PROTAC and the PROTAC pre-bound to the other protein. In the initial states in **(c)** and **(d)**, the dotted lines represent the target or the E3 whose interactions with the rest of the system are turned off except for the harmonic constraints (black lines) to the PROTAC warhead.

simplicity, in this work, all $\Delta\Delta G$s are calculated using the cycle in Fig. 4.1c, which we describe in detail here, but one should arrive at the same result using the mirroring cycle in Fig. 4.1d. By the definition of a thermodynamic cycle, we have $\Delta G_{EP'}^{\text{binary}} - \Delta G_{EP'}^{\text{ternary}} = \Delta G_{\text{couple\_T}}^{\text{binary}} - \Delta G_{\text{couple\_T}}^{\text{ternary}}$, where $\Delta G_{\text{couple\_T}}^{\text{binary}}$ and $\Delta G_{\text{couple\_T}}^{\text{ternary}}$ represent the free energies of coupling T to P and to the target-PROTAC bound complex EP. In the initial states of both coupling processes (vertical processes in red in Fig. 4.1c), T is bound to P but is a dummy molecule at an ideal state. Specifically, multiple harmonic springs connect the binding pocket beads in T to the warhead bead of P, and T itself is an elastic network model consisting of only harmonic springs. All other interactions between T and the rest of the system — whether P or EP — are turned off. Coupling T simply means turning on these inter-molecular interactions, while the binding pocket springs remain unperturbed.

Attaching a dummy T instead of having T dissociated results in a systematic error in the horizontal free energies of EP binding ($\Delta G^{\text{binary}}_{EP'}$ and $\Delta G^{\text{ternary}}_{EP'}$ in Fig. 4.1c) such that the $\Delta\Delta G$ is unaffected. This is because the attachment of dummy T occurs via only one bead on P, except which there are no other forcefield terms involving both physically present beads and dummy beads. In the configurational partition function, energy terms describing the geometries of the physically present part of the system can therefore be separated from the term involving the dummy T and the attachment junction. The latter term is the same whether the physically present part is P or EP, such that the unphysical contribution from attaching dummy T cancels out in $\Delta\Delta G$.

Free energy calculations

Alchemically changing a protein from a dummy state to full coupling involves turning on the interaction potentials between the protein and the rest of the system in the forcefield. The interactions are turned on in stages by sequentially scaling each kind of interaction potential using a coupling parameter λ. Intramolecular potentials (e.g., the elastic network model of each protein) and intermolecular potentials not perturbed at the current stage are unaffected by the λ scaling. For the electrostatic potential, the start state (no electrostatics) and the end state (full electrostatics) correspond to $\lambda_{\text{elec}} = 0$ and 1, respectively. Intermediate states are interpolated such that the potential is defined as $U_{\lambda_{\text{elec}}} = (1-\lambda_{\text{elec}})U_{\text{no\_elec}} + \lambda_{\text{elec}}U_{\text{elec}} = \lambda_{\text{elec}}U_{\text{elec}}$. For numerical stability, the electrostatic potential is only perturbed in the presence of volume exclusion [16, 17], which is modeled by Weeks-Chandler-Andersen (WCA) potential. To turn on Lennard-Jones (LJ) or variants of LJ potentials (e.g., WCA), a soft-core scaling [15] with λ_{LJ} is used for numerical stability:

$$U_{\lambda_{\text{LJ}}}\left(r_{ij}\right) = 4\epsilon\lambda_{\text{LJ}}\left(\frac{1}{\left(\alpha\left(1-\lambda_{\text{LJ}}\right)+\left(\frac{r_{ij}}{\sigma_{ij}}\right)^6\right)^2} - \frac{1}{\alpha\left(1-\lambda_{\text{LJ}}\right)+\left(\frac{r_{ij}}{\sigma_{ij}}\right)^6}\right),$$

where $\alpha = 0.5$, r_{ij} is the distance between beads i and j, and σ_{ij} is the sum of the radii of beads i and j. The number of intermediate states and the spacing of the coupling parameter values depend on the difficulty to obtain converged free energy calculations. For the electrostatic potential, a linear pathway where λ_{elec} ranges from 0 to 1 with a step size of 0.125 is a simple and effective approach. For LJ and related potentials, because most of the free energy changes occur near

the start state of $\lambda_{\mathrm{LJ}} = 0$ (Fig. 4.2b,c), we introduce intermediate states at $\lambda_{\mathrm{LJ}} = 0.005, 0.01, 0.015, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.3, 0.5, 0.7$, and 0.9.

The ΔG of turning on each kind of interaction is calculated using thermodynamic integration (TI) [160], Bennett acceptance ratio (BAR) method [20] and the multi-state BAR (MBAR) method [18]. TI and BAR/MBAR are distinct formulations for free energy calculations, and we verify that these methods converge to similar values. The system in CGMD is evolved using overdamped Langevin dynamics with a diffusion coefficient of 253 nm^2/s and a timestep of 30 ns for stable integration. At each state, at least 64 trajectories of 6 s long are generated to sample the conformations of the complexes. After collecting the samples from trajectories, post-processing involves calculating $\frac{\partial U}{\partial \lambda}$ and ΔU_{ij} for all $i, j = 1, 2, ..., K$ states as inputs for TI, BAR, and MBAR.

4.3 Results and discussion

Alchemical perturbation of protein domains is feasible with CGMD

The binding cooperativity of PROTAC due to PPIs is a unique challenge that calls for an unconventional application of alchemical free energy calculation methods. Alchemical methods are mainly used to determine the binding energies between small-molecule ligands and proteins, and typically no more than 10 heavy atoms are perturbed for efficient and accurate calculations. In protein-protein binding, recent applications and development focus on quantifying the relative free energy changes from small-scale perturbations such as mutations of single residues [161–165]. To our knowledge, the only case that alchemically calculates PPIs in a three-body setting compares how analogs of inhibitors change aberrant multimerization of the HIV-1 integrase [166]. Their proposed thermodynamic framework involves calculating the relative free energy difference by perturbing small molecules that directly participate at a fixed PPI interface. This framework is more readily extendable to molecular glues that modulate PPIs in a similar way. PROTACs, however, due to a more modular design, are typically larger linear molecules. The flexibility of the linker is often nontrivial, such that the two proteins cannot be kept bound at a fixed interface. This configurational entropic concern necessitates an unusually large perturbation at the scale of a protein rather than a small molecule to calculate the binding cooperativity, testing the computational limit of alchemical methods.

To explore the feasibility of the CG alchemical approach, we calculate the free energy of turning on the steric repulsions between a target protein and a PROTAC-E3

complex ($\Delta G^{\text{ternary(sterics)}}$) in the absence of other inter-molecular potentials. We choose Bruton's tyrosine kinase (BTK) as the target (only the kinase domain modeled), CRBN as the E3, and the PROTAC (10) from [139], which are respectively modeled by 87, 124, and 8 beads in the CG model. Together they form the largest target-PROTAC-E3 complex simulated in this work. We compare the calculations using different percentages of the simulation data collected in the time-forward and time-reversed directions. The calculated values of $\Delta G^{\text{ternary(sterics)}}$ plateau starting around the midpoint of the simulation time, indicating numerical convergence (Fig. 4.2a). The time-forward and -reversed estimations are within 1 standard deviation (std) at the midpoint, and the time-reversed estimations remain stable after the midpoint. The observed behavior of the estimates over time suggests that unequilibrated samples at the beginning of the trajectories have been removed, and the remaining frames sample from similar distributions rather than distinct metastable states with slow transition rates [17].

Three methods, TI, BAR, and MBAR are used to separately estimate the free energies. The accuracy of all three methods depends on the number and the spacing of alchemical states. BAR and MBAR reweight conformations sampled from one state by their probability in another state to estimate the free energy differences. Having similar probability distributions between states, i.e., phase space overlap, is therefore critical to the estimation. Unlike BAR/MBAR, TI estimates the free energies by numerically integrating $\langle \frac{\partial U}{\partial \lambda} \rangle$, the ensemble average of the derivative of the potential energy U along the alchemical pathway defined by λ. Depending on the curvature of $\langle \frac{\partial U}{\partial \lambda} \rangle$, choices of intermediate states specified by λ and the integration scheme together introduce integration errors in addition to the statistical errors in estimating the ensemble average per state.

We choose an alchemical pathway that involves 12 intermediate states in addition to the start and end states, such that $\Delta G^{\text{ternary(sterics)}} = \sum_{i=1}^{13} \Delta G_{\lambda_i,\lambda_{i+1}}$, where $\Delta G_{\lambda_i,\lambda_{i+1}}$ is the free energy of changing the WCA potential between neighboring states λ_i and λ_{i+1}. With a total of 14 states unevenly spaced, the phase space overlap between neighboring states is sufficient (Fig. 4.7) for efficient reweighting-based estimations. For TI, the trapezoid rule of numerical integration is used for its simplicity and robustness. Although the quadrature errors result in a slight overestimation of $\Delta G^{\text{ternary(sterics)}}$, the $\partial U/\partial \lambda$ curve is sufficiently smooth such that TI and MBAR largely agree. In addition to the global agreement on $\Delta G^{\text{ternary(sterics)}}$, TI, BAR, and MBAR also locally agree with each other on all $\Delta G_{\lambda_i,\lambda_{i+1}}$ along the alchemical

pathway (Fig. 4.2c). We emphasize that TI and BAR/MBAR rely on distinct types of input data and processing procedures, and their consistency even at the most granular level of calculations further validate our CG alchemical approach.

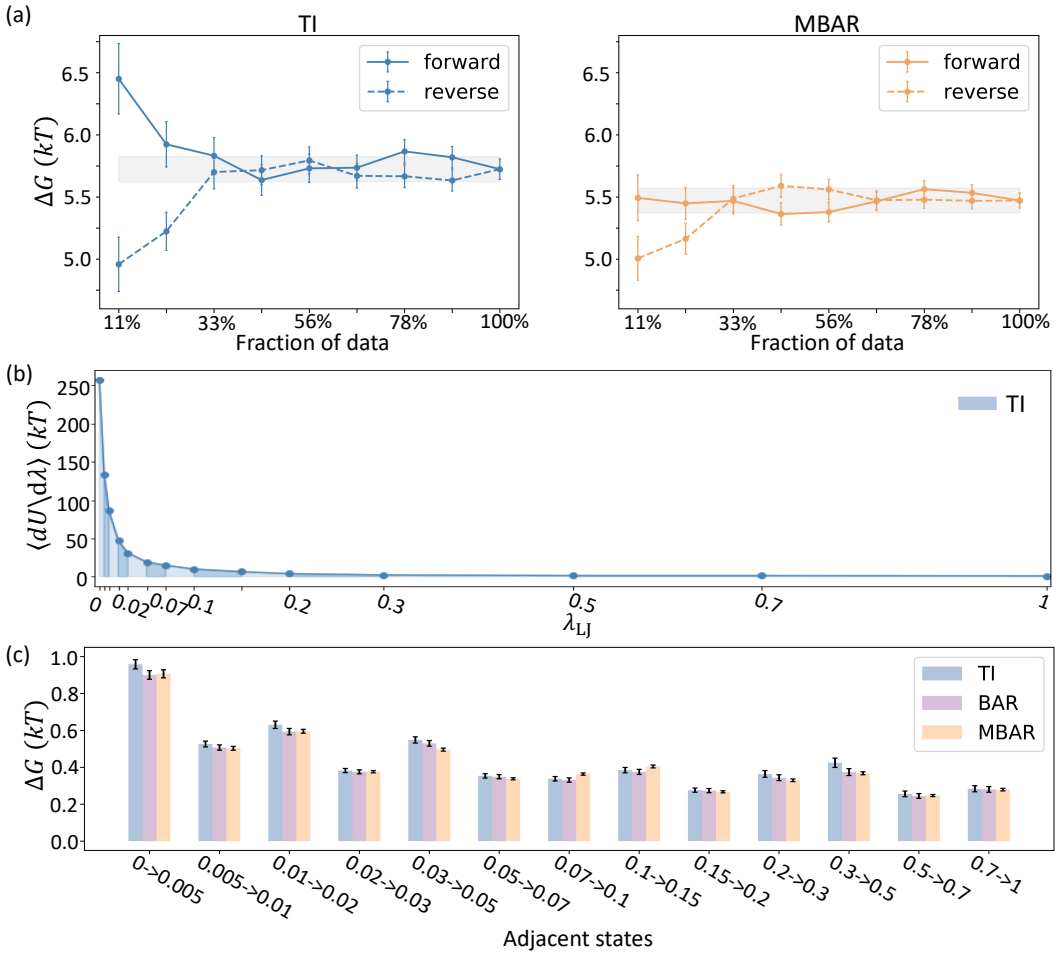

Figure 4.2: Calculation of $\Delta G^{\text{ternary(sterics)}}$ by alchemical perturbation of BTK in the ternary complex of BTK-PROTAC (10)-CRBN. **(a)** TI and MBAR both reach apparent convergence in the time-forward and time-reversed directions with no pathological signs. The grey band in each panel represents the final estimation using 100% data ±0.1 kT as a threshold for error tolerance, where k is the Boltzmann constant. **(b)** TI estimation is shown as the blue area under the curve of $\langle \partial U/\partial \lambda \rangle$. **(c)** TI, BAR, and MBAR agree for all intermediate ΔGs between adjacent states. All error bars of computational results here and in subsequent figures represent ±1 std. Color coding for TI, BAR, and MBAR results are the same in subsequent figures unless otherwise stated.

Analyses of estimations over simulation time and using different free energy calculation methods indicate that convergence of perturbing a protein can be achieved within reasonable computation time, significantly pushing the boundaries of apply-

ing alchemical methods. As parallelization can be done over the alchemical states and over trajectories for each state, the time to run one trajectory is the main limiting factor in the wall-clock computation time of applying our method. Criteria to determine how long a trajectory should be run are described in the Supporting Information (Section 4.5). For this work, depending on the size of the system, 3–14 CPU hours per trajectory of ternary complexes are sufficient.

Minimal forcefield captures entropic effects in PROTAC-mediated PPIs

Encouraged by the proof-of-concept calculations above for $\Delta G^{\text{ternary}}$, we also calculate ΔG^{binary} and complete our calculations for the $\Delta\Delta G$ of the thermodynamic cycle. We follow the sign convention of $\Delta\Delta G$ such that a positive value represents positive cooperativity. The BTK-CRBN system modeled here has been experimentally shown to lack large cooperativity, and introducing PROTACs in Hydrogen/Deuterium Exchange experiments didn't reveal significant profile changes that would indicate the presence of stable PPIs. As the starting point for our method development, we focus on this system due to its apparent simplicity and the availability of experimental characterization over a large range of PROTAC linker lengths. We characterize $\Delta\Delta G$ changes over PROTAC lengths because this relies on capturing the fundamental physics of the tertiary interactions (Fig. 4.3a-c) rather than sequence- or conformation-specific properties.

Two forcefield setups are used to describe PPIs and the resulted $\Delta\Delta G$ trends over PROTAC linker lengths are compared. In the first setup, we calculate the baseline $\Delta\Delta G$ in the absence of PPIs other than volume exclusion. In the second setup, nonspecific attractions between BTK and CRBN beads are added and explored at two strengths. The intrinsic PPIs without PROTAC mediation should be weak such that in the limit of infinite linker length the $\Delta\Delta G$ is negligible. The attenuation of weak PPIs with increasing PROTAC linker lengths originates from configurational entropy. As the PROTAC becomes longer, it experiences a greater loss of configurational freedom upon binding to proteins to induce PPIs (Fig. 4.3b and c), incurring an entropic cost. We examine this configurational entropic effect by modeling $\Delta\Delta G$ at linkers ranging from 1 to 6 beads (σ_s) long, which correspond to approximately 3.5 Å to 21 Å.

In the first setup, the steric cores of the proteins should penalize PROTAC binding and result in negative cooperativities. This is because some conformations that are accessible to the PROTAC in a binary PROTAC-protein complex become inacces-

sible in the ternary complex due to steric clashes (Fig. 4.3a). As the linker length increases and steric clashes are attenuated, the cooperativity should become less negative. We verify that such a monotonically increasing trend of negative $\Delta\Delta G$ is obtained in our model (Fig. 4.3d). Steric penalties on $\Delta\Delta G$ are most obvious at the region of short linker lengths (1–3 beads), after which the benefit from extending the linker length becomes increasingly marginal, and we expect that beyond the simulated window of linker lengths, $\Delta\Delta G$ will eventually plateau near 0. This $\Delta\Delta G$ trend is consistent with a recent effort to tabulate PROTAC linker length structure-activity relationships (SAR), which suggests that steric clashes at short linker lengths often result in a steep decrease in activity [148].

After validating the baseline trend, we next examine how the cooperativity trend is changed by the addition of favorable PPIs through LJ potentials. Increasing the well depth of LJ (ϵ_{LJ}) increases the strength of this nonspecific attraction, which is kept weak (Fig. 4.5) to approximate van der Waals forces. At the attraction strength of $\epsilon_{\mathrm{LJ}} = 0.125\ kT$, the $\Delta\Delta G$ curve is elevated compared to the previous curve without attraction (Fig. 4.3d), as favorable PPIs are expected to enhance cooperativity. Nevertheless, at this attraction strength, steric penalties still dominate and $\Delta\Delta G$s remains negative. Even though adding an LJ potential brings an additional penalty when beads overlap, shorter PROTACs still benefit more from the attractive part of LJ than longer PROTACs, resulting in a flatter $\Delta\Delta G$ trend as compared with the purely repulsive interactions.

An appropriate combination of repulsive and attractive forces may generate a non-monotonic $\Delta\Delta G$ trend, such that intermediate linker lengths promote optimal cooperativity by minimizing steric clashes while maximally sampling attractive PPIs[148]. As the attraction strength increases to $\epsilon_{\mathrm{LJ}} = 0.2\ kT$, intermediate-length PROTACs exhibit not only positive $\Delta\Delta G$s but the values can be comparable and even slightly higher than that of the longest PROTAC (Fig. 4.3d). Within the limited window of linker lengths, only the initial part of the decaying tail of a non-monotonic $\Delta\Delta G$ trend is observed. We expect that beyond the simulated window of linker lengths, configurational entropic penalties will continue driving $\Delta\Delta G$ down towards 0.

Experimentally, the linker length at 3 beads uniquely enables weak positive cooperativity for BTK-CRBN, whereas our results at $\epsilon_{\mathrm{LJ}} = 0.2\ kT$ remain biased towards favoring longer linkers and are not as sensitive to linker length changes. To see whether these characteristics are specific to the choice of the system, we then examine the $\Delta\Delta G$ trends for a different system (Fig. 4.3e), BRD4$^{\mathrm{BD2}}$-VHL, where

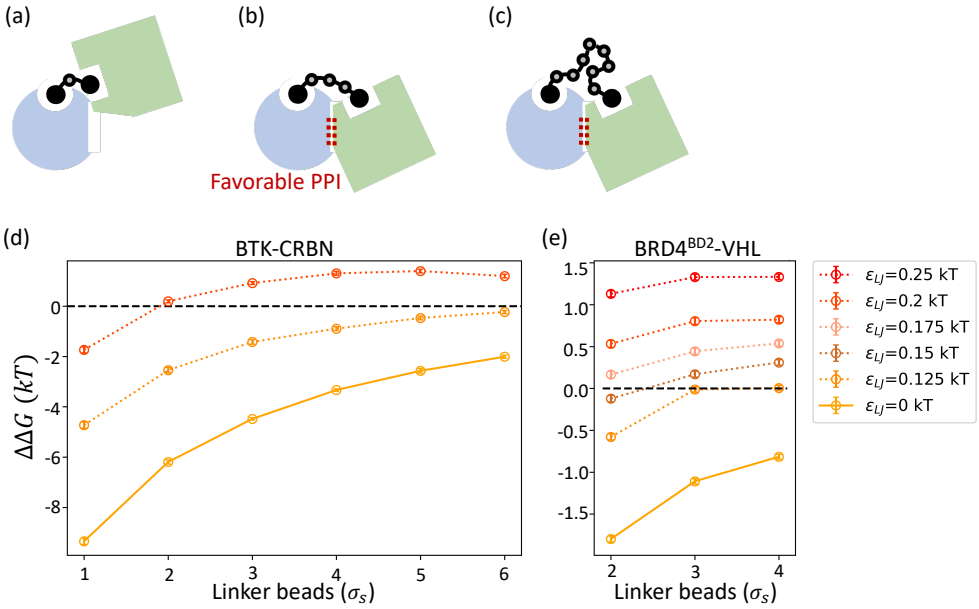

Figure 4.3: PROTAC linker length changes $\Delta\Delta G$ through modulating the effective strength of PPIs. The top three schematics illustrate the scenarios where a PROTAC linker is **(a)** too short to enable favorable contacts between the target (blue) and the E3 (green), **(b)** at an optimal length, and **(c)** sufficiently long but less frequently in a configuration that induces weak favorable PPIs (red dots). The $\Delta\Delta G$ trends over PROTAC linker lengths are calculated for two target-E3 pairs, **(d)** BTK-CRBN and **(e)** BRD4^{BD2}-VHL, under varying strengths of non-specific attractions between proteins. The solid lines represent the baseline $\Delta\Delta G$ trends where only volume exclusion is modeled between the two proteins, and the dotted lines show the trends where nonspecific attractions are added. The strengths (ϵ_{LJ}) of attractions are indicated by different colors. Higher ϵ_{LJ} represents stronger attractions, and the baselines can also be considered as results at $\epsilon_{LJ} = 0$. Results at $\epsilon_{LJ} = 0.125$ and 0.2 kT are plotted for BTK-CRBN and results at ϵ_{LJ} =0.125, 0.15, 0.175, 0.2, and 0.25 kT are plotted for BRD4^{BD2}-VHL. All calculations shown are obtained using MBAR, and results using TI and BAR are superimposed in Figure 4.9.

experimentally, the linker length at 3 beads can also optimize the cooperativity [138]. Due to the smaller size of the system, we can afford to calculate $\Delta\Delta G$s at three more attraction strengths. Similar to BTK-CRBN, in the absence of attractions, negative $\Delta\Delta G$ monotonically increases over the linker length, and adding nonspecific attractions results in flatter and higher $\Delta\Delta G$ curves. Within the narrow window of short linker lengths, scanning the attractive strength ϵ_{LJ} from 0.125 to 0.25 kT, however, does not recapitulate the optimal linker length at 3 beads. This result suggests that enhancing nonspecific attractions in the minimal model is insufficient to compensate for the steric penalties while remaining sensitive to entropic penalties from the linker length.

We demonstrate that the minimal CG model directly captures configurational entropic effects on weak nonspecific PPIs through analyzing $\Delta\Delta G$ trends over PROTAC linker lengths. Beyond this entropic effect, combining repulsive and attractive interactions at various strengths changes the behaviors of cooperativity trends and can shift the optimal linker length, as shown in BTK-CRBN. Nevertheless, chemically specific interactions or specific sampling of certain PPIs is needed to model optimal positive cooperativity at an experimentally relevant range and resolution of PROTAC linker lengths.

Electrostatics in PROTAC-mediated PPIs exhibit plasticity

As a step towards more realistic modeling of cooperativity, we seek chemically specific PPIs to include and further explore the BRD4^{BD2}-VHL system due to the availability of experimental structural information. Crystal structures of the ternary complexes have revealed specific interactions that are proposed as the molecular basis for the observed positive cooperativity and selectivity against other structural homologs [118, 167]. As shown in the previous subsection, these interactions between proteins cannot be approximated by nonspecific attractions that contribute to the cooperativity with low sensitivity to linker length and no protein sequence dependence.

The structural findings such as salt bridges at the PPI interface and the mutational studies involving charged residues on BRD4^{BD2} and homologs [118] motivate us to approach chemical specificity through modeling electrostatic interactions. As CGMD uses an implicit solvent, we choose the Debye-Hückel (DH) potential to describe electrostatics in consideration of screening effects under physiological conditions. Within the BRD4^{BD2}-VHL system, incorporating charges of protein beads results in a monotonic trend of negative $\Delta\Delta G$s with increasing linker length, (Fig. 4.4a) similar to the baseline obtained using steric repulsions only (Fig. 4.3e). Since charges are perturbed separately in $\Delta\Delta G$ calculations for numeric stability, in the following discussions, we further investigate our $\Delta\Delta G$ results by isolating the final stage ($\Delta G^{\text{ternary(charges)}}$) in which charges are turned on in the presence of sterics.

Breaking down the $\Delta\Delta G$s by each energy component shows that at all three linker lengths, $\Delta G^{\text{ternary(charges)}}$ is slightly negative, indicating a mildly favorable process, but the penalty from steric repulsions overwhelmingly dominates electrostatic contributions by an order of magnitude (Fig. 4.4c). As PROTAC linker length increases

from 2 to 4 beads, the contribution from $\Delta G^{\text{ternary(charges)}}$ monotonically diminishes. We consider the possibility that the screening of charges is too strong to model more favorable PPIs and tune the screening parameter in the DH potential at the linker length of 3 beads. Nevertheless, significantly weakening the screening strength leads to a much more unfavorable $\Delta G^{\text{ternary(charges)}}$ (Fig. 4.4c) because both the target protein and the E3 ligase have net positive charges. It is also possible that our level of coarse-graining loses the spatial resolution required for this system to capture detailed interactions like salt bridge formation as observed in the crystal structures [118, 167].

In addition to the small contribution to $\Delta\Delta G$, $\Delta G^{\text{ternary(charges)}}$ itself exhibits plasticity because conformational sampling at the stage of charge perturbation in alchemical free energy calculations is biased by the potentials turned on in previous stages. The presence of steric repulsions combined with nonspecific attractions at the strength of $\epsilon_{\text{LJ}} = 0.2\ kT$, for example, has doubled the $\Delta G^{\text{ternary(charges)}}$ obtained at the linker length of 3 beads without nonspecific attractions (Fig. 4.4c). Interestingly, this change in $\Delta G^{\text{ternary(charges)}}$ is on top of the favorable contribution from nonspecific attractions in the previous calculation stage ($\Delta G^{\text{ternary(other)}}$) before the inclusion of protein charges. For this particular ternary complex, nonspecific attractions and electrostatic interactions work synergistically.

Our dissection of the electrostatic component in $\Delta\Delta G$ under different simulation setups suggests that a more holistic parameterization is needed to accurately evaluate chemically specific PPIs. For BRD4$^{\text{BD2}}$-VHL, incorporating hydrophobic interactions will be of particular interest as there is stacking of hydrophobic residues at the PPI interface observed in the crystal structures [118, 167]. Hydrophobic interactions may also introduce non-additive free-energy contributions with electrostatics in a similar manner seen with the nonspecific attractions. It is also worth noting that the favorable PPIs revealed by crystal structures are enabled by PROTACs using a JQ1 warhead, which imposes a different linker attachment angle (i.e., exit vector) from an I-BET726 warhead (Fig. 4.11).[138] Our current forcefield does not model the PROTAC linker with angular terms to specify the exit vectors, which leads to a $\Delta\Delta G$ trend that matches well with the worse-performing I-BET726 set of PROTACs (Fig. 4.4a). As rigidifying PROTACs is a common strategy to optimize the cooperativity by entropically enhancing certain PPIs[140, 167], parameterizing linker conformations will improve modeling the specificity in PROTAC-mediated PPIs.

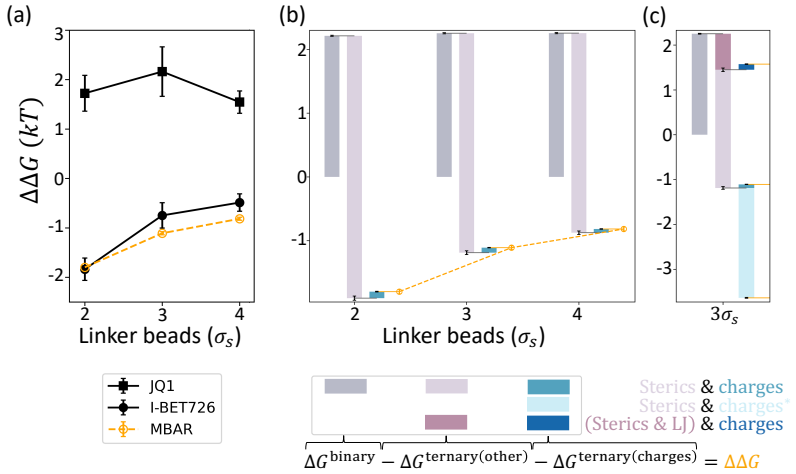

Figure 4.4: Electrostatic contributions to the cooperativity in the BRD4$^{\text{BD2}}$-VHL system are small and context-dependent. All calculations shown are obtained using MBAR, and results using TI and BAR are shown in Figure 4.10. **(a)** Calculations of $\Delta\Delta G$s over PROTAC linker lengths are shown with the experimental measurements [138] (black) converted to our units. Experimental results at 2, 3, and 4 linker beads correspond to MZ4, MZ1, and MZ2 for PROTACs using JQ1 warhead and MZP-61, MZP-54, MZP-55 for PROTACs using I-BET726 warhead. **(b)** Waterfall plot breakdown of $\Delta\Delta G$ calculations. At each linker length, bars in each triplet correspond to ΔG^{binary} (grey), $-\Delta G^{\text{ternary(other)}}$ (light purple), and $-\Delta G^{\text{ternary(charges)}}$ (turquoise), and are arranged in a cumulative manner such that the end position marks the resulted $\Delta\Delta G$ (orange). $\Delta G^{\text{ternary(other)}}$ denotes the free energy change of turning on interaction energy components other than the electrostatics, which only include steric repulsions in this panel. **(c)** $\Delta\Delta G$ breakdowns at linker length 3 under different forcefield parameterizations are superimposed for comparison. Reducing the screening effect by ten-fold (charges*) significantly increases $\Delta G^{\text{ternary(charges)}}$ (cyan), which leads to a very negative $\Delta\Delta G$. Introducing non-specific attractions ($\epsilon_{\text{LJ}} = 0.2\ kT$) not only reduces $\Delta G^{\text{ternary(other)}}$ (dark purple) but also doubles $\Delta G^{\text{ternary(charges)}}$ (steel blue), resulting in a positive $\Delta\Delta G$.

4.4 Conclusions

We explore a novel computational approach to model the binding cooperativity of PROTACs by combining CGMD and alchemical free energy calculations. The plasticity of PROTAC-mediated PPIs motivates an unconventional application of alchemical methods at a perturbation scale that is rarely attempted. We show that with coarse-graining, converged estimates from various free energy calculation methods are attainable within a reasonable amount of computation time. Our results expand the possibility of more creative use of alchemical methods. The feasibility and efficiency of the CG alchemical approach enable us to probe multiple energy components under the alchemical framework and characterize how PROTAC linker

lengths modulate PPIs under different setups to produce distinct cooperativity trends. In addition to validating the benefit of using long linkers to avoid steric clashes, we demonstrate with a simple addition of nonspecific attractions between BTK and CRBN that the binding cooperativity can be promoted by shortening the PROTAC linker. Our minimal model is capable of unveiling such changes in cooperativity that are rooted in the configurational freedom of the ternary complexes rather than chemical properties.

Quantitative modeling of the cooperativity, however, remains difficult due to the lack of specificity in the minimal model. Previous studies have recognized the challenges brought by non-native PROTAC-mediated PPIs that are often weak, transient, and pliable, and have called for a paradigm shift towards an ensemble-based characterization beyond a handful of docked or crystal poses. [119, 144, 149]. While thermodynamic properties such as the binding cooperativity are inherently ensemble-based, we note that both accurate sampling of PPI conformations according to chemical properties and efficient computation to sample a diverse set of conformations are important for calculations. Currently, tuning the strength of nonspecific attractions cannot approximate favorable PPIs while retaining sensitivity to entropic constraints from the PROTAC linker length. Simply adding electrostatic interactions based on amino acid charges proved insufficient to capture the cooperativity trend enabled by JQ1-based PROTACs in $BRD4^{BD2}$-VHL. Additional parameterizations are needed to capture chemically specific PPIs.

Two main avenues are worth exploring for future improvement of our method — PROTAC linker conformations and protein sequence-dependence. Among a myriad of PROTAC properties [116] that we leave out, structural features such as the exit vector [138] and the linker rigidity [140, 167] in addition to the linker length can both entropically constrain the sampling of PPIs. Meanwhile, energy components of PPIs other than electrostatic interactions, notably the hydrophobic effects, are currently overlooked. Different energy components may have non-additive effects in optimizing the absolute cooperativity and relative cooperativities between target homologs such as $BRD4^{BD2}$ and $BRD4^{BD1}$. Although coarse-graining enables efficient computation, parameterization for both directions of forcefield development will be a major hurdle to overcome. This can be bottom-up using shorter-timescale higher-resolution simulations, similar to that of the CG ENM (Fig. 4.6) in this work. A top-down fitting might also become possible with rapidly growing experimental studies that develop platforms [168] for empirical SAR of PROTAC linkerology

[169, 170] or leverage promiscuous PROTACs and target homologs and mutants to investigate the molecular basis of specificity [171].

Acknowledgments

H.M. thanks William M. Clemons, Jr., Daniel Jacobson, Tomislav Begušić, Xuecheng Tao, Marta Gonzalvo, and Lixue Cheng for comments on the manuscript, and Zhen-Gang Wang and Christopher J. Balzer for technical discussions. We gratefully acknowledge support from the National Institutes of Health (NIH) R01GM138845 (8877_CIT, subaward), Amgen Chem-Bio-Engineering Award (CBEA), and De-Logi Trust Science and Technology Grant. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Bridges computer at the Pittsburgh Supercomputing Center through allocation MCB160013 [172]. XSEDE is supported by National Science Foundation grant number ACI-1548562. This work also used computational resources from the Resnick High Performance Computing Center, a facility supported by Resnick Sustainability Institute at the California Institute of Technology.

4.5 Appendix

CGMD Forcefield

The complete potential energy function for a ternary complex is

$$\begin{aligned} U(\boldsymbol{x};\boldsymbol{b},\boldsymbol{q}) = {} & U_{\text{ENM}}(\boldsymbol{x}_E) + U_{\text{ENM}}(\boldsymbol{x}_T) + U_{\text{spring}}(\boldsymbol{x}_P) + U_{\text{WCA}}(\boldsymbol{x}_P) \\ & + U_{\text{bind}}(\boldsymbol{x}_P,\boldsymbol{x}_T;\boldsymbol{b}) + U_{\text{bind}}(\boldsymbol{x}_P,\boldsymbol{x}_E;\boldsymbol{b}) \\ & + U_{\text{WCA}}(\boldsymbol{x}_P,\boldsymbol{x}_T) + U_{\text{WCA}}(\boldsymbol{x}_P,\boldsymbol{x}_E) + U_{\text{WCA}}(\boldsymbol{x}_E,\boldsymbol{x}_T) \\ & + U_{\text{elec}}(\boldsymbol{x}_E,\boldsymbol{x}_T;\boldsymbol{q}) + U_{\text{LJ}}(\boldsymbol{x}_E,\boldsymbol{x}_T;\epsilon_{\text{LJ}}) \end{aligned} \tag{4.1}$$

where $\boldsymbol{x}_E$, $\boldsymbol{x}_T$, and $\boldsymbol{x}_P$ indicate the coordinates of the E3 ligase, the target protein, and the PROTAC, respectively, $\boldsymbol{q}$ represent the charges of protein beads, and $\boldsymbol{b}$ are indicators of whether protein beads are at the binding pocket or not. All PROTAC beads are modeled with 0 charge and no attraction to the proteins. All parameters and variables are defined using a length scale of the large bead ($\sigma = 0.8$ nm) and an energy scale of $\epsilon = kT$ where k is the Boltzmann constant and $T = 310$ K.

Internal energy terms

Interactions within a protein are modeled by an elastic network model (ENM) such that every pair of beads within distance R_c is connected by a harmonic spring:

$$U_{\text{ENM}}(\boldsymbol{x}) = \sum_{(i,j)\in D} k_{\text{spring}}\left(\Delta x_{ij} - d_{ij}\right)^2 \tag{4.2}$$

where k_{spring} is the spring constant, d_{ij} is the optimal distance between x_i and x_j, and $D = \{(i,j)\,|d_{ij} < R_c\}$. The optimal distance between a pair of beads is its initial distance in the experimental structure. Experimental structures used in this work include VHL (PDB: 5T35[118] chain D), BRD4^{BD2} (PDB: 5T35[118] chain A), CRBN (PDB: 6BOY[119] chain B), and BTK (PDB: 6W7O[140] chain A), and Schrödinger Maestro [173] is used to fill in missing atoms and perform energy minimization before building the CG ENM. Additional details on the parameterization are described in a separate section below.

PROTAC is modeled as a linear molecule, where adjacent beads are connected by springs ($U_{\text{spring}}(\boldsymbol{x}_P)$) and non-adjacent beads are subjected to steric repulsions ($U_{\text{WCA}}(\boldsymbol{x}_P)$).

Interaction energy terms

PROTAC-protein interactions consist of binding interactions modeled by springs between a binding moiety bead in the PROTAC and all beads in the corresponding

binding pocket ($U_{\text{bind}}(\boldsymbol{x}_P, \boldsymbol{x}_T; \boldsymbol{b})$ and $U_{\text{bind}}(\boldsymbol{x}_P, \boldsymbol{x}_E; \boldsymbol{b})$ in eq.(4.1)) and steric repulsions ($U_{\text{WCA}}(\boldsymbol{x}_P, \boldsymbol{x}_T)$ and $U_{\text{WCA}}(\boldsymbol{x}_P, \boldsymbol{x}_E)$) between the remaining parts of PROTAC and protein. Steric repulsions in intra-PROTAC, PROTAC-protein, and inter-protein interactions are all modeled by the Weeks-Chandler-Andersen (WCA) potential, a shifted and truncated version of Lennard-Jones (LJ) potential.

Protein-protein interactions are captured by the steric repulsions ($U_{\text{WCA}}(\boldsymbol{x}_E, \boldsymbol{x}_T)$), and depending on the modeling purpose, electrostatics ($U_{\text{elec}}(\boldsymbol{x}_E, \boldsymbol{x}_T; \boldsymbol{q})$) or nonspecific attractions ($U_{\text{LJ}}(\boldsymbol{x}_E, \boldsymbol{x}_T; \epsilon_{\text{LJ}})$). The electrostatic interaction is modeled by a Debye-Hückel (DH) potential. The functional forms and parameterization of both potentials can be found in [153]. When reducing the screening of electrostatics between BRD4$^{\text{BD2}}$ and VHL, the Debye length κ is multiplied by 10. The solvent in our system is treated implicitly. Nonspecific attractions aimed at broadly including Van der Waals forces and hydrophobic interactions are modeled by LJ potentials. The strength of the attraction is kept under that of electrostatic interactions (Fig. 4.5). The well depth of LJ, ϵ_{LJ}, is currently set to be the same for all pairs of beads for nonspecific attraction. For future efforts, minor modifications to the formula [174] and parameterization of ϵ_{LJ} to depend on the Wimley-White hydrophobicity scale, for example, can capture more sequence-specific interactions such as the hydrophobic effects.

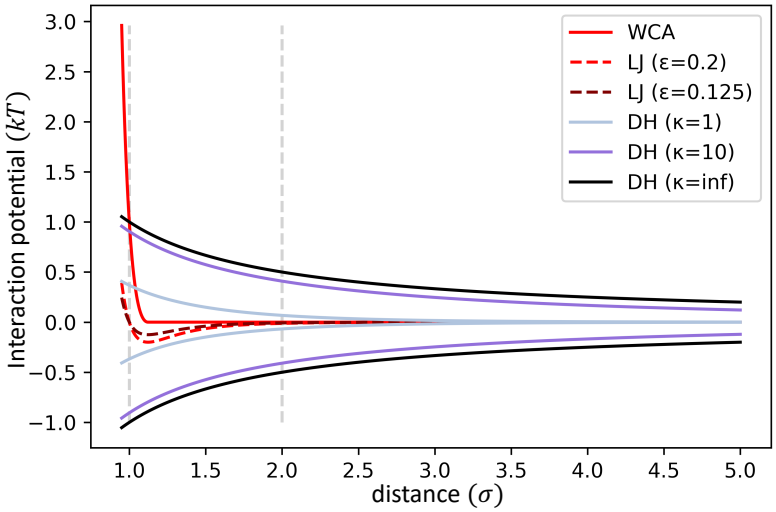

Figure 4.5: The strengths of various interaction potentials are plotted over the distance between protein beads. The two vertical dashed grey lines bound the distance between 1 and 2 σ. The electrostatic potentials (DH) are plotted for beads with +1 and +1 charges or +1 and -1 charges.

Parameterization of ENM

ENM is a model that represents the tertiary structure of a protein by connecting every pair of protein beads within a certain distance cutoff R_c by a Hookean spring of spring constant k_{spring}. Despite the simplicity of its parameterization, slow modes in ENM can capture biologically significant conformational changes [158, 159]. This structure-based model can also be used in combination with other physics-driven forcefields to model macromolecular complexes. Protein-protein associations and viral capsid assembly have both been successfully modeled by using Elnedyn, an ENM at the resolution of 1 residue per bead [175], on top of the MARTINI CG forcefield. By fitting to atomistic simulations, Elnedyn preserves both structural properties and dynamics within each protein subunit for the CG simulations.

We follow a similar protocol and fit our CG ENM parameters in eq.(4.2) to Elnedyn simulations results. Three proteins — IKZF1$^{\text{ZF2}}$ (PDB: 6H0F [176] chain C), BRD4$^{\text{BD1}}$ (PDB: 6BOY [119] chain C), and CRBN (PDB: 6BOY [119] chain B) — are chosen for the fitting to represent the range of protein sizes based on the publicly available crystal structures of PROTAC-mediated ternary complexes. Elnedyn is supported as an option in the MARTINI 2 CG forcefield [175], and we use the default parameters to generate Elnedyn simulations of these proteins with GROMACS version 5.0.7. Two equilibration stages were run, first at 1 fs timestep for 50 ps, and then at 10 fs timestep for 1 ns. Then, only the dynamics stage was used for fitting, which was run at 10 fs timestep for 40 ns. Four metrics are used to examine how well a particular combination of k_{spring} and R_c captures information in Elnedyn simulations: the difference of time-averaged root-mean-squared-deviation (ΔRMSD), bead-averaged root-mean-squared-fluctuation (ΔRMSF), Kullback–Leibler (KL) divergence of the RMSD distributions, and the root-mean-squared inner product of the principal components (RMSIP) of the trajectories.

Within a single metric, we usually observe a degeneracy within a certain region of k_{spring} and R_c values (Fig. 4.6), and this was also observed in Elnedyn fitting to atomistic simulations [175]. This is because increasing either k_{spring} or R_c can increase the stiffness of a protein and, therefore, can compensate for each other to some extent. Nevertheless, despite the degeneracy, given the wide range of protein sizes, there is no single combination of k_{spring} and R_c values that works best for all three proteins. We chose $k_{\text{spring}} = 100\epsilon/\sigma^2$ and $R_c = 2.0\sigma$ as they are near the optimal degeneracy region under most metrics and consistent with the values of Elnedyn parameters ($k_{\text{spring}} = 124.25\epsilon/\sigma^2$ and $R_c = 1.125\sigma$). This combination

of k_{spring} and R_c was selected without a global optimization function that combines all four metrics, and should be subjected to finer tuning if a specific system is of interest.

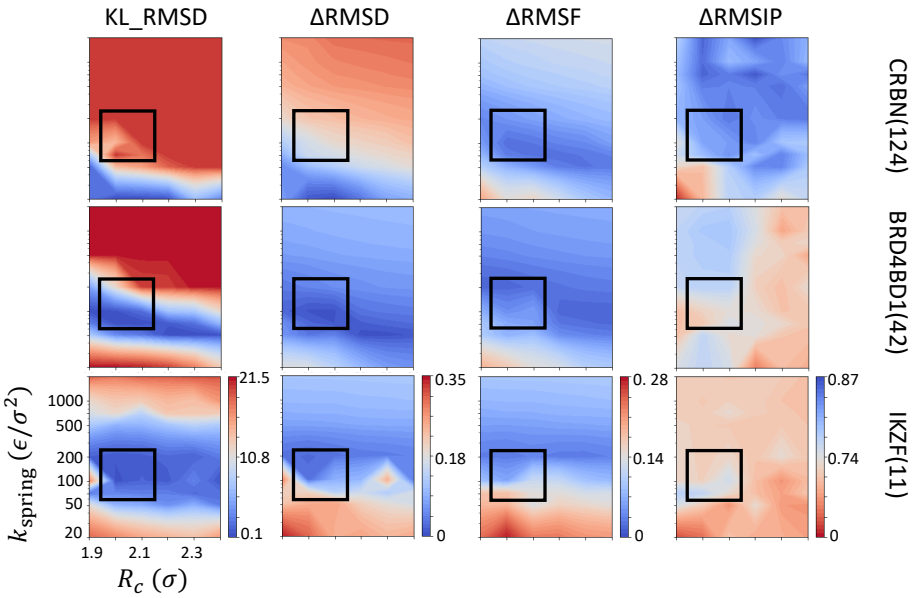

Figure 4.6: Fitting results of ENM parameters arranged by proteins (rows) and evaluation metrics (columns). Numbers in parenthesis next to protein names are the number of CG beads. For each plot, blue regions indicate k_{spring} and R_c values that result in good fitting, and red regions indicate significant differences between our simulations and Elnedyn simulations. Each column shares the same colorbar range. In general, the boxed regions around $k_{\text{spring}} = 100\epsilon/\sigma^2$ and $R_c = 2.0\sigma$ has good fitting.

Analysis of alchemical free energy calculations

We perform various checks to address two common concerns in alchemical simulations: 1) are there sufficient intermediate states along the alchemical reaction pathway, and 2) are there sufficient samples from each state for accurate free energy calculations. The BTK-PROTAC (10)-CRBN complex is used as an example for the analysis below.

We first validate that there are sufficient intermediate states for a converged estimation of $\Delta G^{\text{ternary(WCA)}}$. The convergence of free energy calculations depends on the overlap of the phase space, i.e., the distribution of sampled conformations, between neighboring states. Substantial overlap is achieved when the neighboring states are similar, which requires a fine spacing of the coupling parameter values. In practice, distributions of quantities such as ΔU and $\partial U/\partial\lambda$ that are directly involved in free energy estimations are often treated as proxies for the high-dimensional phase space

[16]. The similarity between distributions is quantified by KL divergence, where 0 indicates identical distributions and $\gg 1$ suggests concerning differences. Based on this metric, all neighboring states have substantial overlap, as the Kullback–Leibler (KL) divergence values of ΔU and of $\partial U/\partial\lambda$ distributions both stay below 1 (Fig. 4.7a).

Bennett's overlapping histogram [20] provides another qualitative test for the overlap of ΔU distributions. The difference between $g_{\lambda_{i+1}}(\Delta U_{\lambda_i,\lambda_{i+1}}) = P_{\lambda_i}(\Delta U_{\lambda_i,\lambda_{i+1}}) + (1-C)\,\Delta U_{\lambda_i,\lambda_{i+1}}$ and $g_{\lambda_i}(\Delta U_{\lambda_i,\lambda_{i+1}}) = P_{\lambda_{i+1}}(\Delta U_{\lambda_i,\lambda_{i+1}}) - C\Delta U_{\lambda_i,\lambda_{i+1}}$ is plotted over $\Delta U_{\lambda_i,\lambda_{i+1}}$ values, where C is an arbitrary constant between 0 and 1 and $P_{\lambda_i}(\Delta U_{\lambda_i,\lambda_{i+1}})$ and $P_{\lambda_i}(\Delta U_{\lambda_i,\lambda_{i+1}})$ are the distributions of $\Delta U_{\lambda_i,\lambda_{i+1}}$ obtained by sampling from neighboring alchemical states λ_i and λ_{i+1}, respectively. Continuous oscillations of $g_{\lambda_{i+1}}(\Delta U_{\lambda_i,\lambda_{i+1}}) - g_{\lambda_i}(\Delta U_{\lambda_i,\lambda_{i+1}})$ around the estimated $\Delta G_{\lambda_i,\lambda_{i+1}}$ over a range of $\Delta U_{\lambda_i,\lambda_{i+1}}$ values suggests good overlap (Fig. 4.7b) [17]. For states of higher λ_{LJ} values, higher energetic penalty of steric repulsions prevents sampling over a wide range of ΔU values, but the KL divergence and visualization of the distributions (Fig. 4.7a,c) both indicate the quality of the overlap.

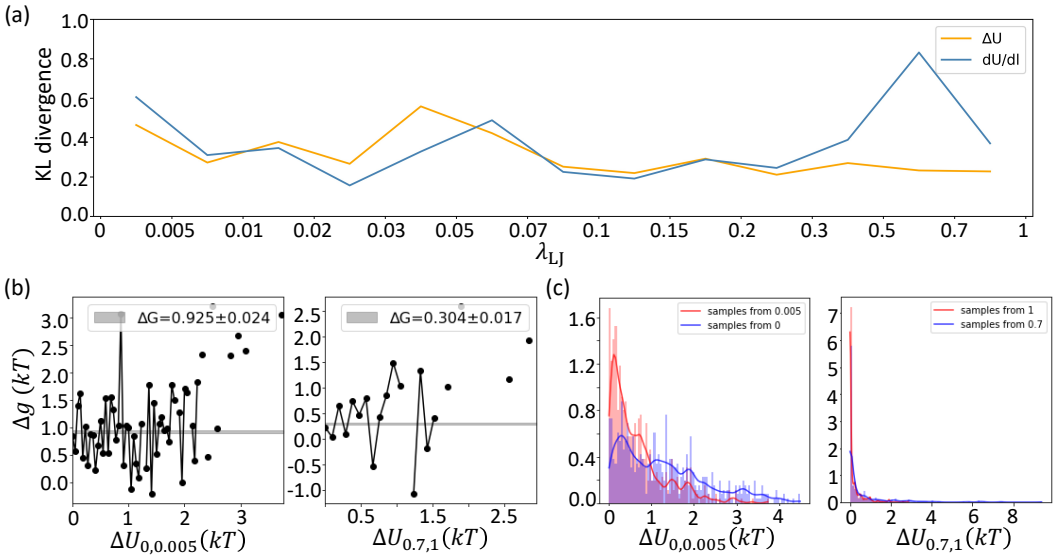

Figure 4.7: Phase space overlap in calculating $\Delta G^{\mathrm{ternary(WCA)}}$ for BTK-CRBN in Fig. 2. **(a)** Overlap of ΔU and $\partial U/\partial\lambda$ distributions between adjacent states are quantified by the KL divergence. **(b)** Example Bennett's overlapping plots for $\lambda_{\mathrm{LJ}} = 0, 0.005$ states (left) and $\lambda_{\mathrm{LJ}} = 0.7, 1$ states (right). The grey bands represent $\Delta G_{\lambda_i,\lambda_{i+1}} \pm 1$ std estimated using BAR. **(c)** Example distributions of $\Delta U_{i,i+1}$ are shown with Gaussian smoothing (red and blue solid curves) for better visualization.

Next, we examine sampling within each state. For each state, a simulation needs to be post-processed to discard the initial unequilibrated part and then subsampled to

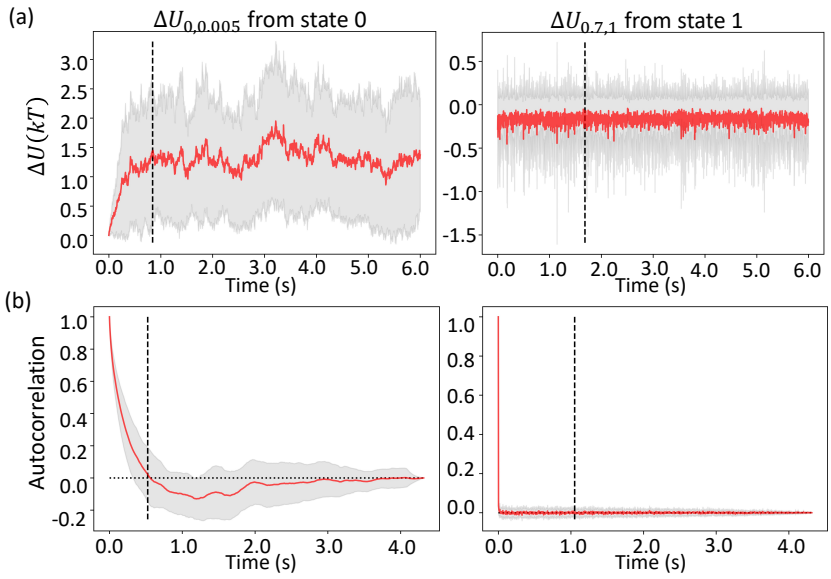

Figure 4.8: Detecting equilibration and autocorrelation time in calculating $\Delta G^{\text{ternary(WCA)}}$ for BTK-CRBN in Fig. 2. **(a)** $\Delta U_{\lambda_i,\lambda_{i+1}}$ over simulation time and **(b)** the autocorrelation of $\Delta U_{\lambda_i,\lambda_{i+1}}$ from $\lambda_{\text{LJ}} = 0$ (left) and $\lambda_{\text{LJ}} = 1$ (right). The red curves and the shaded regions represent the average value ±1 standard deviation based on 64 independent trajectories. The vertical dashed lines in this example mark 0.9 s in (a) and 0.63 s in (b). The horizontal dotted lines in (b) mark the 0 autocorrelation value.

obtain de-correlated data for accurate uncertainty quantification of the free energy estimation. Thus, the length of the simulations is dictated by the equilibration time, autocorrelation time, and the number of de-correlated samples needed for converged estimations. We examine the values of ΔU, $\partial U/\partial\lambda$, and other collective variables over the simulation time, which typically equilibrate after 0.9 s (Fig. 4.8a). To find out the decorrelation time, we discard the initial 0.9 s of simulations and plot the autocorrelation functions of these variables over different time lags up to half of the simulation time to ensure that the autocorrelation is calculated from a sufficient number of samples. The autocorrelation times all plummet to 0 before 0.63 s (Fig. 4.8b). Both equilibration time and decorrelation time are longer for simulations in lower value of λ_{LJ} states that retain more memory of previously sampled configurations due to lower energetic costs. Currently, the equilibration and autocorrelation cutoffs depend on each system. For convenience, we used the same cutoffs for all λ states. In the future, this can be customized for each state to maximize the number of samples, especially from states of high λ values that requires less equilibration and decorrelation time (Fig. 4.8b).

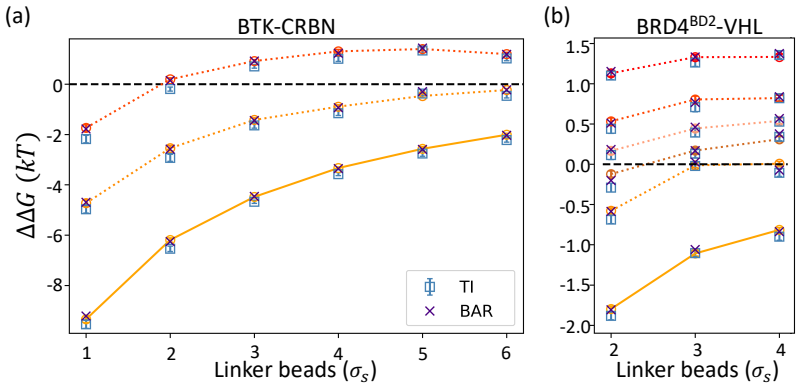

Figure 4.9: $\Delta\Delta G$s calculated by TI and BAR are superimposed onto the MBAR results shown in Figure 3 to show that all three alchemical free energy calculation methods agree within noise.

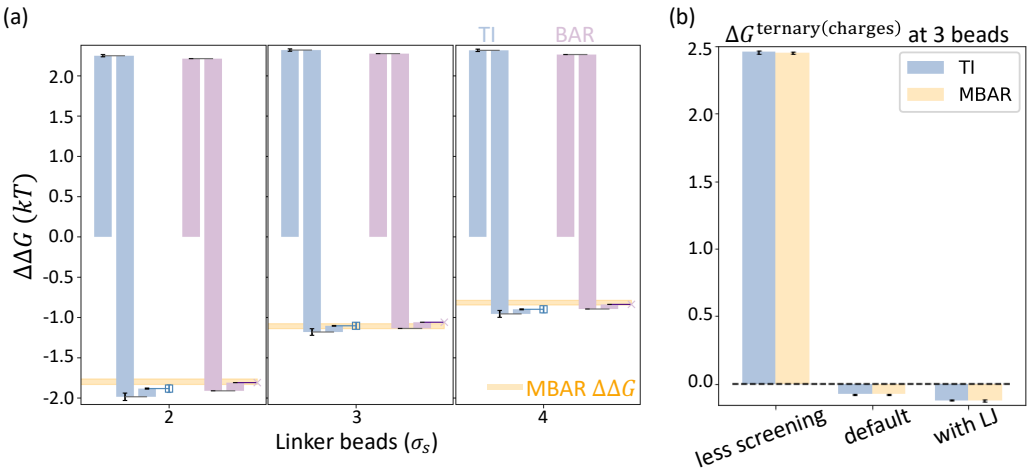

Figure 4.10: $\Delta\Delta G$s calculated by TI and BAR agree with MBAR results shown in Figure 4 for the BRD4$^{\text{BD2}}$-VHL system modeled with protein charges included. **(a)** $\Delta\Delta G$s at each PROTAC linker length calculated by TI and BAR are broken down using waterfall plots similar to Figure 4b. In each triplet, columns from left to right correspond to ΔG^{binary}, $-\Delta G^{\text{ternary(other)}}$, and $-\Delta G^{\text{ternary(charges)}}$. Columns are arranged cumulatively such that the end point of a triplet of columns represent the final $\Delta\Delta G$ value calculated by the corresponding method. MBAR $\Delta\Delta G$ values with ± 1 standard deviation are shown as horizontal yellow bands for reference. **(b)** TI and MBAR calculations of the electrostatic contribution to $\Delta\Delta G$ under different forcefield setups at the linker length of 3 beads agree with each other. Note that $\Delta G^{\text{ternary(charges)}}$ is shown here rather than $-\Delta G^{\text{ternary(charges)}}$ in panel **(a)**.

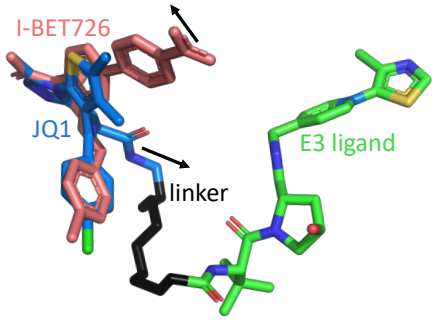

Figure 4.11: The structure of MZ1, which is a PROTAC with linker length of 3 beads using a JQ1 warhead, extracted from the ternary crystal structure (PDB: 5T35[118]) and the structure of I-BET726 warhead extracted from the crystal structure of a binary complex (PDB: 4BJX[177]) are superimposed to highlight the difference in exit vectors (black arrows).

BIBLIOGRAPHY

(1) Bracewell, R. N. Strip Integration in Radio Astronomy. *Australian Journal of Physics* **1956**, *9*, 198–217. DOI: 10.1071/ph560198.

(2) Kershner, R. The Number of Circles Covering a Set. *American Journal of Mathematics* **1939**, *61*, 665–671. DOI: 10.2307/2371320.

(3) Mastronarde, D. SerialEM: A Program for Automated Tilt Series Acquisition on Tecnai Microscopes Using Prediction of Specimen Position. *Microscopy and Microanalysis* **2003**, *9*, 1182–1183. DOI: 10.1017/S1431927603445911.

(4) Wade, R. H. A Brief Look at Imaging and Contrast Transfer. *Ultramicroscopy* **1992**, *46*, 145–156. DOI: 10.1016/0304-3991(92)90011-8.

(5) Rohou, A.; Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *Journal of Structural Biology* **2015**, *192*, 216–221. DOI: 10.1016/j.jsb.2015.08.008.

(6) Thon, F. Notizen: Zur Defokussierungsabhängigkeit des Phasenkontrastes bei der elektronenmikroskopischen Abbildung. *Zeitschrift für Naturforschung A* **1966**, *21*, 476–478. DOI: 10.1515/zna-1966-0417.

(7) Hagen, W. J. H.; Wan, W.; Briggs, J. A. G. Implementation of a Cryo-Electron Tomography Tilt-Scheme Optimized for High Resolution Subtomogram Averaging. *Journal of Structural Biology* **2017**, *197*, 191–198. DOI: 10.1016/j.jsb.2016.06.007.

(8) Dalton, K. M.; Greisman, J. B.; Hekstra, D. R. A Unifying Bayesian Framework for Merging X-Ray Diffraction Data. *Nature Communications* **2022**, *13*, 1–13. DOI: 10.1038/s41467-022-35280-8.

(9) Aldama, L. A.; Dalton, K. M.; Hekstra, D. R. Correcting Systematic Errors in Diffraction Data with Modern Scaling Algorithms. *Acta Crystallographica Section D: Structural Biology* **2023**, *79*, 796–805. DOI: 10.1107/S2059798323005776.

(10) Wilson, A. J. C. The Probability Distribution of X-Ray Intensities. *Acta Crystallographica* **1949**, *2*, 318–321. DOI: 10.1107/S0365110X49000813.

(11) Srinivasan, R.; Parthasarathy, S. *Some Statistical Applications in X-Ray Crystallography*; Elsevier Science & Technology Books, 1976.

(12) Howells, E. R.; Phillips, D. C.; Rogers, D. The Probability Distribution of X-Ray Intensities. Ii. Experimental Investigation and the X-Ray Detection of Centres of Symmetry. *Acta Crystallographica* **1950**, *3*, 210–214. DOI: 10.1107/S0365110X50000513.

(13) Rees, D. C. The Influence of Twinning by Merohedry on Intensity Statistics. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **1980**, *36*, 578–581. DOI: 10.1107/S0567739480001234.

(14) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **1954**, *22*, 1420–1426. DOI: 10.1063/1.1740409.

(15) Mey, A. S. J. S.; Allen, B.; Macdonald, H. E. B.; Chodera, J. D.; Kuhn, M.; Michel, J.; Mobley, D. L.; Naden, L. N.; Prasad, S.; Rizzi, A.; Scheen, J.; Shirts, M. R.; Tresadern, G.; Xu, H. Best Practices for Alchemical Free Energy Calculations. *Living Journal of Computational Molecular Science* **2020**, *2*, 18378. DOI: 10.33011/livecoms.2.1.18378.

(16) Pohorille, A.; Jarzynski, C.; Chipot, C. Good Practices in Free-Energy Calculations. *The Journal of Physical Chemistry B* **2010**, *114*, 10235–10253. DOI: 10.1021/jp102971x.

(17) Klimovich, P. V.; Shirts, M. R.; Mobley, D. L. Guidelines for the Analysis of Free Energy Calculations. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 397–411. DOI: 10.1007/s10822-015-9840-9.

(18) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *The Journal of Chemical Physics* **2008**, *129*, 124105. DOI: 10.1063/1.2978177.

(19) Shirts, M. R. Reweighting from the Mixture Distribution as a Better Way to Describe the Multistate Bennett Acceptance Ratio. *arXiv* **2017**, 1704.00891, ver. 4. DOI: 10.48550/arXiv.1704.00891.

(20) Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *Journal of Computational Physics* **1976**, *22*, 245–268. DOI: 10.1016/0021-9991(76)90078-4.

(21) Tocheva, E. I.; Li, Z.; Jensen, G. J. Electron Cryotomography. *Cold Spring Harb Perspectives in Biology* **2010**, *2*, a003442. DOI: 10.1101/cshperspect.a003442.

(22) Böhning, J.; Bharat, T. A. M. Towards High-Throughput In Situ Structural Biology Using Electron Cryotomography. *Progress in Biophysics and Molecular Biology* **2021**, *160*, 97–103. DOI: 10.1016/j.pbiomolbio.2020.05.010.

(23) Himes, B. A.; Zhang, P. emClarity: Software for High-Resolution Cryo-Electron Tomography and Subtomogram Averaging. *Nature Methods* **2018**, *15*, 955–961. DOI: 10.1038/s41592-018-0167-z.

(24) Tegunov, D.; Xue, L.; Dienemann, C.; Cramer, P.; Mahamid, J. Multi-Particle Cryo-EM Refinement with M Visualizes Ribosome-Antibiotic Complex at 3.5 Å in Cells. *Nature Methods* **2021**, *18*, 186–193. DOI: 10.1038/s41592-020-01054-7.

(25) Mastronarde, D. N. Automated Electron Microscope Tomography Using Robust Prediction of Specimen Movements. *Journal of Structural Biology* **2005**, *152*, 36–51. DOI: 10.1016/j.jsb.2005.07.007.

(26) Noske, A. B.; Costin, A. J.; Morgan, G. P.; Marsh, B. J. Expedited Approaches to Whole Cell Electron Tomography and Organelle Mark-Up In Situ in High-Pressure Frozen Pancreatic Islets. *Journal of Structural Biology* **2008**, *161*, 298–313. DOI: 10.1016/j.jsb.2007.09.015.

(27) Rog-Zielinska, E. A.; Johnston, C. M.; O'Toole, E. T.; Morphew, M.; Hoenger, A.; Kohl, P. Electron Tomography of Rabbit Cardiomyocyte Three-Dimensional Ultrastructure. *Progress in Biophysics and Molecular Biology* **2016**, *121*, 77–84. DOI: 10.1016/j.pbiomolbio.2016.05.005.

(28) Koning, R. I.; Zovko, S.; Bárcena, M.; Oostergetel, G. T.; Koerten, H. K.; Galjart, N.; Koster, A. J.; Mieke Mommaas, A. Cryo Electron Tomography of Vitrified Fibroblasts: Microtubule Plus Ends In Situ. *Journal of Structural Biology* **2008**, *161*, 459–468. DOI: 10.1016/j.jsb.2007.08.011.

(29) Baker, L. A.; Rubinstein, J. L. Radiation Damage in Electron Cryomicroscopy. *Methods in Enzymology* **2010**, *481*, 371–388. DOI: 10.1016/S0076-6879(10)81015-8.

(30) Chreifi, G.; Chen, S.; Metskas, L. A.; Kaplan, M.; Jensen, G. J. Rapid Tilt-Series Acquisition for Electron Cryotomography. *Journal of Structural Biology* **2019**, *205*, 163–169. DOI: 10.1016/j.jsb.2018.12.008.

(31) Konings, S.; Kuijper, M.; Keizer, J.; Grollios, F.; Spanjer, T.; Tiemeijer, P. Advances in Single Particle Analysis Data Acquisition. *Microscopy and Microanalysis* **2019**, *25*, 1012–1013. DOI: 10.1017/S1431927619005798.

(32) Weis, F.; Hagen, W. J. H. Combining High Throughput and High Quality for Cryo-Electron Microscopy Data Collection. *Acta Crystallographica Section D: Biological Crystallography* **2020**, *76*, 724–728. DOI: 10.1107/s2059798320008347.

(33) Bammes, B. E.; Rochat, R. H.; Jakana, J.; Chen, D. H.; Chiu, W. Direct Electron Detection Yields Cryo-EM Reconstructions at Resolutions Beyond 3/4 Nyquist Frequency. *Journal of Structural Biology* **2012**, *177*, 589–601. DOI: 10.1016/j.jsb.2012.01.008.

(34) Milazzo, A. C.; Cheng, A.; Moeller, A.; Lyumkis, D.; Jacovetty, E.; Polukas, J.; Ellisman, M. H.; Xuong, N. H.; Carragher, B.; Potter, C. S. Initial Evaluation of a Direct Detection Device Detector for Single Particle Cryo-Electron

Microscopy. *Journal of Structural Biology* **2011**, *176*, 404–408. DOI: 10.1016/j.jsb.2011.09.002.

(35) Kuijper, M.; van Hoften, G.; Janssen, B.; Geurink, R.; De Carlo, S.; Vos, M.; van Duinen, G.; van Haeringen, B.; Storms, M. FEI's Direct Electron Detector Developments: Embarking on a Revolution in Cryo-TEM. *Journal of Structural Biology* **2015**, *192*, 179–187. DOI: 10.1016/j.jsb.2015.09.014.

(36) McMullan, G.; Faruqi, A. R.; Henderson, R. Direct Electron Detectors. *Methods in Enzymology* **2016**, *579*, 1–17. DOI: 10.1016/bs.mie.2016.05.056.

(37) Marko, M.; Hsieh, C.; Schalek, R.; Frank, J.; Mannella, C. Focused-Ion-Beam Thinning of Frozen-Hhydrated Biological Specimens for Cryo-Electron Microscopy. *Nature Methods* **2007**, *4*, 215–217. DOI: 10.1038/nmeth1014.

(38) Rigort, A.; Bauerlein, F. J.; Villa, E.; Eibauer, M.; Laugks, T.; Baumeister, W.; Plitzko, J. M. Focused Ion Beam Micromachining of Eukaryotic Cells for Cryoelectron Tomography. *Proceedings of the National Academy of Sciences* **2012**, *109*, 4449–4454. DOI: 10.1073/pnas.1201333109.

(39) Mahamid, J.; Pfeffer, S.; Schaffer, M.; Villa, E.; Danev, R.; Cuellar, L. K.; Förster, F.; Hyman, A. A.; Plitzko, J. M.; Baumeister, W. Visualizing the Molecular Sociology at the HeLa Cell Nuclear Periphery. *Science* **2016**, *351*, 969–972. DOI: 10.1126/science.aad8857.

(40) Glaeser, R. M. Specimen Behavior in the Electron Beam. *Methods in Enzymology* **2016**, *579*, 19–50. DOI: 10.1016/bs.mie.2016.04.010.

(41) Bartesaghi, A.; Aguerrebere, C.; Falconieri, V.; Banerjee, S.; Earl, L. A.; Zhu, X.; Grigorieff, N.; Milne, J. L. S.; Sapiro, G.; Wu, X.; Subramaniam, S. Atomic Resolution Cryo-EM Structure of β-Galactosidase. *Structure* **2018**, *26*, 848–856. DOI: 10.1016/j.str.2018.04.004.

(42) Hattne, J.; Shi, D.; Glynn, C.; Zee, C. T.; Gallagher-Jones, M.; Martynowycz, M. W.; Rodriguez, J. A.; Gonen, T. Analysis of Global and Site-Specific Radiation Damage in Cryo-EM. *Structure* **2018**, *26*, 759–766. DOI: 10.1016/j.str.2018.03.021.

(43) Lippens, S.; Kremer, A.; Borghgraef, P.; Guérin, C. J. Serial Block Face-Scanning Electron Microscopy for Volume Electron Microscopy. *Methods in Cell Biology* **2019**, *152*, 69–85. DOI: 10.1016/bs.mcb.2019.04.002.

(44) Le Gros, M. A.; McDermott, G.; Cinquin, B. P.; Smith, E. A.; Do, M.; Chao, W. L.; Naulleau, P. P.; Larabell, C. A. Biological Soft X-ray Tomography on Beamline 2.1 at the Advanced Light Source. *Journal of Synchrotron Radiation* **2014**, *21*, 1370–1377. DOI: 10.1107/S1600577514015033.

(45) Loconte, V.; Singla, J.; Li, A.; Chen, J. H.; Ekman, A.; McDermott, G.; Sali, A.; Le Gros, M.; White, K. L.; Larabell, C. A. Soft X-ray Tomography to Map and Quantify Organelle Interactions at the Mesoscale. *Structure* **2022**, DOI: 10.1016/j.str.2022.01.006.

(46) Xu, C. S.; Hayworth, K. J.; Lu, Z.; Grob, P.; Hassan, A. M.; García-Cerdán, J. G.; Niyogi, K. K.; Nogales, E.; Weinberg, R. J.; Hess, H. F. Enhanced FIB-SEM Systems for Large-Volume 3D Imaging. *eLife* **2017**, *6*, DOI: 10.7554/eLife.25916.

(47) Hagen, W. J. H.; Wan, W.; Briggs, J. A. G. Implementation of a Cryo-Electron Tomography Tilt-Scheme Optimized for High Resolution Subtomogram Averaging. *Journal of Structural Biology* **2017**, *197*, 191–198. DOI: 10.1016/j.jsb.2016.06.007.

(48) Swinbank, R.; James Purser, R. Fibonacci Grids: A Novel Approach to Global Modelling. *Quarterly Journal of the Royal Meteorological Society* **2006**, *132*, 1769–1793. DOI: 10.1256/qj.05.227.

(49) Chreifi, G.; Chen, S.; Jensen, G. J. Rapid Tilt-Series Method for Cryo-Electron Tomography: Characterizing Stage Behavior during FISE Acquisition. *Journal of Structural Biology* **2021**, *213*, 107716. DOI: 10.1016/j.jsb.2021.107716.

(50) Xiong, Q.; Morphew, M. K.; Schwartz, C. L.; Hoenger, A. H.; Mastronarde, D. N. CTF Determination and Correction for Low Dose Tomographic Tilt Series. *Journal of Structural Biology* **2009**, *168*, 378–387. DOI: 10.1016/j.jsb.2009.08.016.

(51) Saalfeld, S.; Fetter, R.; Cardona, A.; Tomancak, P. Elastic Volume Reconstruction From Series of Ultra-Thin Microscopy Sections. *Nature Methods* **2012**, *9*, 717–720. DOI: 10.1038/nmeth.2072.

(52) Wolf, S. G.; Mutsafi, Y.; Dadosh, T.; Ilani, T.; Lansky, Z.; Horowitz, B.; Rubin, S.; Elbaum, M.; Fass, D. 3D Visualization of Mitochondrial Solid-Phase Calcium Stores in Whole Cells. *elife* **2017**, *6*, DOI: 10.7554/eLife.29929.

(53) Ni, T.; Frosio, T.; Mendonça, L.; Sheng, Y.; Clare, D.; Himes, B. A.; Zhang, P. High-Resolution In Situ Structure Determination by Cryo-Electron Tomography and Subtomogram Averaging Using emClarity. *Nature Protocols* **2022**, *17*, 421–444. DOI: 10.1038/s41596-021-00648-5.

(54) Chen, M.; Bell, J. M.; Shi, X.; Sun, S. Y.; Wang, Z.; Ludtke, S. J. A Complete Data Processing Workflow for Cryo-ET and Subtomogram Averaging. *Nature Methods* **2019**, *16*, 1161–1168. DOI: 10.1038/s41592-019-0591-8.

(55) Nickell, S.; Kofler, C.; Leis, A. P.; Baumeister, W. A Visual Approach to Proteomics. *Nature Reviews Molecular Cell Biology* **2006**, *7*, 225–230. DOI: 10.1038/nrm1861.

(56) Moebel, E.; Martinez-Sanchez, A.; Lamm, L.; Righetto, R. D.; Wietrzynski, W.; Albert, S.; Larivière, D.; Fourmentin, E.; Pfeffer, S.; Ortiz, J.; Baumeister, W.; Peng, T.; Engel, B. D.; Kervrann, C. Deep Learning Improves Macromolecule Identification in 3D Cellular Cryo-Electron Tomograms. *Nature Methods* **2021**, DOI: 10.1038/s41592-021-01275-4.

(57) Smith, J. L.; Fischetti, R. F.; Yamamoto, M. Micro-Crystallography Comes of Age. *Current Opinion in Structural Biology* **2012**, *22*, 602–612. DOI: 10.1016/j.sbi.2012.09.001.

(58) Gemmi, M.; Mugnaioli, E.; Gorelik, T. E.; Kolb, U.; Palatinus, L.; Boullay, P.; Hovmöller, S.; Abrahams, J. P. 3D Electron Diffraction: The Nanocrystallography Revolution. *ACS Central Science* **2019**, *5*, 1315–1329. DOI: 10.1021/acscentsci.9b00394.

(59) Grimes, J. M.; Hall, D. R.; Ashton, A. W.; Evans, G.; Owen, R. L.; Wagner, A.; McAuley, K. E.; von Delft, F.; Orville, A. M.; Sorensen, T.; Walsh, M. A.; Ginn, H. M.; Stuart, D. I. Where Is Crystallography Going? *Acta Crystallographica Section D: Structural Biology* **2018**, *74*, 152–166. DOI: 10.1107/S2059798317016709.

(60) Chapman, H. N. *et al.* Femtosecond X-Ray Protein Nanocrystallography. *Nature* **2011**, *470*, 73–77. DOI: 10.1038/nature09750.

(61) Shi, D.; Nannenga, B. L.; Iadanza, M. G.; Gonen, T. Three-Dimensional Electron Crystallography of Protein Microcrystals. *eLife* **2013**, *2*, ed. by Harrison, S. C., e01345. DOI: 10.7554/eLife.01345.

(62) Wang, Y.; Takki, S.; Cheung, O.; Xu, H.; Wan, W.; Öhrström, L.; Ken Inge, A. Elucidation of the Elusive Structure and Formula of the Active Pharmaceutical Ingredient Bismuth Subgallate by Continuous Rotation Electron Diffraction. *Chemical Communications* **2017**, *53*, 7018–7021. DOI: 10.1039/C7CC03180G.

(63) Jones, C. G.; Martynowycz, M. W.; Hattne, J.; Fulton, T. J.; Stoltz, B. M.; Rodriguez, J. A.; Nelson, H. M.; Gonen, T. The CryoEM Method MicroED as a Powerful Tool for Small Molecule Structure Determination. *ACS Central Science* **2018**, *4*, 1587–1592. DOI: 10.1021/acscentsci.8b00760.

(64) Gruene, T. *et al.* Rapid Structure Determination of Microcrystalline Molecular Compounds Using Electron Diffraction. *Angewandte Chemie International Edition* **2018**, *57*, 16313–16317. DOI: 10.1002/anie.201811318.

(65) Kim, L. J.; Ohashi, M.; Zhang, Z.; Tan, D.; Asay, M.; Cascio, D.; Rodriguez, J. A.; Tang, Y.; Nelson, H. M. Prospecting for Natural Products by Genome Mining and Microcrystal Electron Diffraction. *Nature Chemical Biology* **2021**, *17*, 872–877. DOI: 10.1038/s41589-021-00834-2.

(66) Bruhn, J. F. *et al.* Small Molecule Microcrystal Electron Diffraction for the Pharmaceutical Industry–Lessons Learned From Examining Over Fifty Samples. *Frontiers in Molecular Biosciences* **2021**, *8*, 648603. DOI: 10.3389/fmolb.2021.648603.

(67) Park, J.-D.; Li, Y.; Moon, K.; Han, E. J.; Lee, S. R.; Seyedsayamdost, M. R. Structural Elucidation of Cryptic Algaecides in Marine Algal-Bacterial Symbioses by NMR Spectroscopy and MicroED. *Angewandte Chemie International Edition* **2022**, *61*, e202114022. DOI: 10.1002/anie.202114022.

(68) E. Gorelik, T.; E. Tehrani, K. H. M.; Gruene, T.; Monecke, T.; Niessing, D.; Kaiser, U.; Blankenfeldt, W.; Müller, R. Crystal Structure of Natural Product Argyrin-D Determined by 3D Electron Diffraction. *CrystEngComm* **2022**, *24*, 5885–5889. DOI: 10.1039/D2CE00707J.

(69) Delgadillo, D. A. *et al.* High-Throughput Identification of Crystalline Natural Products from Crude Extracts Enabled by Microarray Technology and microED. *ACS Central Science* **2024**, *10*, 176–183. DOI: 10.1021/acscentsci.3c01365.

(70) Cichocka, M. O.; Ångström, J.; Wang, B.; Zou, X.; Smeets, S. High-Throughput Continuous Rotation Electron Diffraction Data Acquisition Via Software Automation. *Journal of Applied Crystallography* **2018**, *51*, 1652–1661. DOI: 10.1107/S1600576718015145.

(71) Wang, B.; Zou, X.; Smeets, S. Automated Serial Rotation Electron Diffraction Combined with Cluster Analysis: An Efficient Multi-Crystal Workflow for Structure Determination. *IUCrJ* **2019**, *6*, 854–867. DOI: 10.1107/S2052252519007681.

(72) Lightowler, M.; Li, S.; Ou, X.; Cho, J.; Li, A.; Hofer, G.; Xu, J.; Yang, T.; Zou, X.; Lu, M.; Xu, H. Phase Identification and Discovery of Hidden Crystal Forms in a Polycrystalline Pharmaceutical Sample Using High-Throughput 3D Electron Diffraction. *ChemRxiv* **2023**, ver. 1. DOI: 10.26434/chemrxiv-2023-2rh9j.

(73) Unge, J.; Lin, J.; Weaver, S. J.; Sae Her, A.; Gonen, T. Compositional Analysis of Complex Mixtures using Automatic MicroED Data Collection. *Advanced Science* **2024**, *11*, 2400081. DOI: 10.1002/advs.202400081.

(74) Powell, S. M.; Novikova, I. V.; Kim, D. N.; Evans, J. E. AutoMicroED: A Semi-Automated MicroED Processing Pipeline. *bioRxiv* **2021**, ver. 1. DOI: 10.1101/2021.12.13.472146.

(75) Kabsch, W. XDS. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 125–132. DOI: 10.1107/S0907444909047337.

(76) Sheldrick, G. M. Phase Annealing in SHELX-90: Direct Methods for Larger Structures. *Acta Crystallographica Section A: Foundations of Crystallography* **1990**, *46*, 467–473. DOI: 10.1107/S0108767390000277.

(77) Nannenga, B. L.; Gonen, T. Protein Structure Determination by MicroED. *Current Opinion in Structural Biology* **2014**, *27*, 24–31. DOI: 10.1016/j.sbi.2014.03.004.

(78) Palatinus, L.; Petříček, V.; Corrêa, C. A. Structure Refinement Using Precession Electron Diffraction Tomography and Dynamical Diffraction: Theory and Implementation. *Acta Crystallographica Section A: Foundations and Advances* **2015**, *71*, 235–244. DOI: 10.1107/S2053273315001266.

(79) Khouchen, M.; Klar, P. B.; Chintakindi, H.; Suresh, A.; Palatinus, L. Optimal Estimated Standard Uncertainties of Reflection Intensities for Kinematical Refinement from 3D Electron Diffraction Data. *Acta Crystallographica Section A: Foundations and Advances* **2023**, *79*, 427–439. DOI: 10.1107/S2053273323005053.

(80) Saha, A.; Mecklenburg, M.; Pattison, A. J.; Brewster, A. S.; Rodriguez, J. A.; Ercius, P. Mapping Electron Beam-Induced Radiolytic Damage in Molecular Crystals. *arXiv* **2024**, 2404.18011, ver. 1. DOI: 10.48550/arXiv.2404.18011.

(81) Karplus, P. A.; Diederichs, K. Assessing and Maximizing Data Quality in Macromolecular Crystallography. *Current Opinion in Structural Biology* **2015**, *34*, 60–68. DOI: 10.1016/j.sbi.2015.07.003.

(82) Uervirojnangkoorn, M.; Zeldin, O. B.; Lyubimov, A. Y.; Hattne, J.; Brewster, A. S.; Sauter, N. K.; Brunger, A. T.; Weis, W. I. Enabling X-Ray Free Electron Laser Crystallography for Challenging Biological Systems from a Limited Number of Crystals. *eLife* **2015**, *4*, ed. by Harrison, S. C., e05421. DOI: 10.7554/eLife.05421.

(83) White, T. A.; Mariani, V.; Brehm, W.; Yefanov, O.; Barty, A.; Beyerlein, K. R.; Chervinskii, F.; Galli, L.; Gati, C.; Nakane, T.; Tolstikova, A.; Yamashita, K.; Yoon, C. H.; Diederichs, K.; Chapman, H. N. Recent Developments in CrystFEL. *Journal of Applied Crystallography* **2016**, *49*, 680–689. DOI: 10.1107/S1600576716004751.

(84) Beilsten-Edmands, J.; Winter, G.; Gildea, R.; Parkhurst, J.; Waterman, D.; Evans, G. Scaling Diffraction Data in the DIALS Software Package: Algorithms and New Approaches for Multi-Crystal Scaling. *Acta Crystallographica Section D: Structural Biology* **2020**, *76*, 385–399. DOI: 10.1107/S2059798320003198.

(85) Gildea, R. J.; Beilsten-Edmands, J.; Axford, D.; Horrell, S.; Aller, P.; Sandy, J.; Sanchez-Weatherby, J.; Owen, C. D.; Lukacik, P.; Strain-Damerell, C.; Owen, R. L.; Walsh, M. A.; Winter, G. Xia2.multiplex: A Multi-Crystal Data-Analysis Pipeline. *Acta Crystallographica Section D: Structural Biology* **2022**, *78*, 752–769. DOI: 10.1107/S2059798322004399.

(86) Assmann, G.; Brehm, W.; Diederichs, K. Identification of Rogue Datasets in Serial Crystallography. *Journal of Applied Crystallography* **2016**, *49*, 1021–1028. DOI: 10.1107/S1600576716005471.

(87) Giordano, R.; Leal, R. M. F.; Bourenkov, G. P.; McSweeney, S.; Popov, A. N. The Application of Hierarchical Cluster Analysis to the Selection of Isomorphous Crystals. *Acta Crystallographica Section D: Biological Crystallography* **2012**, *68*, 649–658. DOI: 10.1107/S0907444912006841.

(88) Foadi, J.; Aller, P.; Alguel, Y.; Cameron, A.; Axford, D.; Owen, R. L.; Armour, W.; Waterman, D. G.; Iwata, S.; Evans, G. Clustering Procedures for the Optimal Selection of Data Sets from Multiple Crystals in Macromolecular Crystallography. *Acta Crystallographica Section D: Biological Crystallography* **2013**, *69*, 1617–1632. DOI: 10.1107/S0907444913012274.

(89) Yamashita, K.; Hirata, K.; Yamamoto, M. KAMO: Towards Automated Data Processing for Microcrystals. *Acta Crystallographica Section D: Structural Biology* **2018**, *74*, 441–449. DOI: 10.1107/S2059798318004576.

(90) Burch, J. E.; Smith, A. G.; Caille, S.; Walker, S. D.; Wurz, R.; Cee, V. J.; Rodriguez, J.; Gostovic, D.; Quasdorf, K.; Nelson, H. M. Putting MicroED to the Test: An Account of the Evaluation of 30 Diverse Pharmaceutical Compounds, 2021, DOI: 10.26434/chemrxiv-2021-h3tqz.

(91) Chhetri, B. K.; Mojib, N.; Moore, S. G.; Delgadillo, D. A.; Burch, J. E.; Barrett, N. H.; Gaul, D. A.; Marquez, L.; Soapi, K.; Nelson, H. M.; Quave, C. L.; Kubanek, J. Cryptic Chemical Variation in a Marine Red Alga as Revealed by Nontargeted Metabolomics. *ACS Omega* **2023**, *8*, 13899–13910. DOI: 10.1021/acsomega.3c00301.

(92) Abad, A. N. D.; Seshadri, K.; Ohashi, M.; Delgadillo, D. A.; de Moraes, L. S.; Nagasawa, K. K.; Liu, M.; Johnson, S.; Nelson, H. M.; Tang, Y. Discovery and Characterization of Pyridoxal 5'-Phosphate-Dependent Cycloleucine Synthases. *Journal of the American Chemical Society* **2024**, *146*, 14672–14684. DOI: 10.1021/jacs.4c02142.

(93) Kabsch, W. Integration, Scaling, Space-Group Assignment and Post-Refinement. *Acta Crystallographica Section D: Biological Crystallography* **2010**, *66*, 133–144. DOI: 10.1107/S0907444909047374.

(94) Mott, N. F.; Bragg, W. L. The Scattering of Electrons by Atoms. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **1997**, *127*, 658–665. DOI: 10.1098/rspa.1930.0082.

(95) Dorset, D. L. Electron Crystallography. *Acta Crystallographica Section B: Structural Science* **1996**, *52*, 753–769. DOI: 10.1107/S0108768196005599.

(96) DeLaBarre, B.; Brunger, A. T. Considerations for the Refinement of Low-Resolution Crystal Structures. *Acta Crystallographica Section D: Biological Crystallography* **2006**, *62*, 923–932. DOI: 10.1107/S0907444906012650.

(97) Evans, P. R.; Murshudov, G. N. How Good Are My Data and What Is the Resolution? *Acta Crystallographica Section D: Biological Crystallography* **2013**, *69*, 1204–1214. DOI: 10.1107/S0907444913000061.

(98) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **2017**, *112*, 859–877. DOI: 10.1080/01621459.2017.1285773.

(99) Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; Saul, L. K. An Introduction to Variational Methods for Graphical Models. *Machine Learning* **1999**, *37*, 183–233. DOI: 10.1023/A:1007665907178.

(100) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2014**, 1312.6114, ver. 10. DOI: 10.48550/arXiv.1312.6114.

(101) Evans, P. R. An Introduction to Data Reduction: Space-Group Determination, Scaling and Intensity Statistics. *Acta Crystallographica Section D: Biological Crystallography* **2011**, *67*, 282–292. DOI: 10.1107/S090744491003982X.

(102) Hattne, J.; Reyes, F. E.; Nannenga, B. L.; Shi, D.; de la Cruz, M. J.; Leslie, A. G. W.; Gonen, T. MicroED Data Collection and Processing. *Acta Crystallographica Section A: Foundations and Advances* **2015**, *71*, 353–360. DOI: 10.1107/S2053273315010669.

(103) Winter, G.; Waterman, D. G.; Parkhurst, J. M.; Brewster, A. S.; Gildea, R. J.; Gerstel, M.; Fuentes-Montero, L.; Vollmar, M.; Michels-Clark, T.; Young, I. D.; Sauter, N. K.; Evans, G. DIALS: Implementation and Evaluation of a New Integration Package. *Acta Crystallographica Section D: Structural Biology* **2018**, *74*, 85–97. DOI: 10.1107/S2059798317017235.

(104) Clabbers, M. T. B.; Gruene, T.; Parkhurst, J. M.; Abrahams, J. P.; Waterman, D. G. Electron Diffraction Data Processing with DIALS. *Acta Crystallographica Section D: Structural Biology* **2018**, *74*, 506–518. DOI: 10.1107/S2059798318007726.

(105) Petříček, V.; Palatinus, L.; Plášil, J.; Dušek, M. Jana2020 – a New Version of the Crystallographic Computing System Jana. *Zeitschrift für Kristallographie - Crystalline Materials* **2023**, *238*, 271–282. DOI: 10.1515/zkri-2023-0005.

(106) Ito, S.; White, F. J.; Okunishi, E.; Aoyama, Y.; Yamano, A.; Sato, H.; Ferrara, J. D.; Jasnowski, M.; Meyer, M. Structure determination of small molecule compounds by an electron diffractometer for 3D ED/MicroED. *CrystEngComm* **2021**, *23*, 8622–8630. DOI: 10.1039/D1CE01172C.

(107) Greisman, J. B.; Dalton, K. M.; Hekstra, D. R. Reciprocalspaceship: A Python Library for Crystallographic Data Analysis. *Journal of Applied Crystallography* **2021**, *54*, 1521–1529. DOI: 10.1107/S160057672100755X.

(108) Sheldrick, G. M. SHELXT – Integrated Space-Group and Crystal-Structure Determination. *Acta Crystallographica Section A: Foundations and Advances* **2015**, *71*, 3–8. DOI: 10.1107/S2053273314026370.

(109) Usón, I.; Sheldrick, G. M. Advances in Direct Methods for Protein Crystallography. *Current Opinion in Structural Biology* **1999**, *9*, 643–648. DOI: 10.1016/S0959-440X(99)00020-2.

(110) Schrödinger, LLC The PyMOL Molecular Graphics System, Version 2.5, 2022.

(111) An, S.; Fu, L. Small-Molecule PROTACs: An Emerging and Promising Approach for the Development of Targeted Therapy Drugs. *EBioMedicine* **2018**, *36*, 553–562. DOI: 10.1016/j.ebiom.2018.09.005.

(112) Burslem, G. M.; Crews, C. M. Proteolysis-Targeting Chimeras as Therapeutics and Tools for Biological Discovery. *Cell* **2020**, *181*, 102–114. DOI: 10.1016/j.cell.2019.11.031.

(113) Sakamoto, K. M.; Kim, K. B.; Kumagai, A.; Mercurio, F.; Crews, C. M.; Deshaies, R. J. PROTACs: Chimeric Molecules That Target Proteins to the Skp1–Cullin–F Box Complex for Ubiquitination and Degradation. *Proceedings of the National Academy of Sciences* **2001**, *98*, 8554–8559. DOI: 10.1073/pnas.141230798.

(114) Mullard, A. Targeted Protein Degraders Crowd into the Clinic. *Nature Reviews Drug Discovery* **2021**, *20*, 247–250. DOI: 10.1038/d41573-021-00052-4.

(115) Maniaci, C.; Ciulli, A. Bifunctional Chemical Probes Inducing Protein–Protein Interactions. *Current Opinion in Chemical Biology* **2019**, *52*, 145–156. DOI: 10.1016/j.cbpa.2019.07.003.

(116) Troup, R. I.; Fallan, C.; Baud, M. G. J. Current Strategies for the Design of Protac Linkers: A Critical Review. *Exploration of Targeted Anti-tumor Therapy* **2020**, *1*, 273–312. DOI: 10.37349/etat.2020.00018.

(117) Alabi, S.; Crews, C. Major Advances in Targeted Protein Degradation: PROTACs, LYTACs, and MADTACs. *Journal of Biological Chemistry* **2021**, 100647. DOI: 10.1016/j.jbc.2021.100647.

(118) Gadd, M. S.; Testa, A.; Lucas, X.; Chan, K.-H.; Chen, W.; Lamont, D. J.; Zengerle, M.; Ciulli, A. Structural Basis of PROTAC Cooperative Recognition for Selective Protein Degradation. *Nature Chemical Biology* **2017**, *13*, 514–521. DOI: 10.1038/nchembio.2329.

(119) Nowak, R. P.; DeAngelo, S. L.; Buckley, D.; He, Z.; Donovan, K. A.; An, J.; Safaee, N.; Jedrychowski, M. P.; Ponthier, C. M.; Ishoey, M.; Zhang, T.; Mancias, J. D.; Gray, N. S.; Bradner, J. E.; Fischer, E. S. Plasticity in Binding Confers Selectivity in Ligand-Induced Protein Degradation. *Nature Chemical Biology* **2018**, *14*, 706–714. DOI: 10.1038/s41589-018-0055-y.

(120) Smith, B. E.; Wang, S. L.; Jaime-Figueroa, S.; Harbin, A.; Wang, J.; Hamman, B. D.; Crews, C. M. Differential PROTAC Substrate Specificity Dictated by Orientation of Recruited E3 Ligase. *Nature Communications* **2019**, *10*, 131. DOI: 10.1038/s41467-018-08027-7.

(121) Riching, K. M.; Mahan, S.; Corona, C. R.; McDougall, M.; Vasta, J. D.; Robers, M. B.; Urh, M.; Daniels, D. L. Quantitative Live-Cell Kinetic Degradation and Mechanistic Profiling of PROTAC Mode of Action. *ACS Chemical Biology* **2018**, *13*, 2758–2770. DOI: 10.1021/acschembio.8b00692.

(122) Roy, M. J.; Winkler, S.; Hughes, S. J.; Whitworth, C.; Galant, M.; Farnaby, W.; Rumpel, K.; Ciulli, A. SPR-Measured Dissociation Kinetics of PROTAC Ternary Complexes Influence Target Degradation Rate. *ACS Chemical Biology* **2019**, *14*, 361–368. DOI: 10.1021/acschembio.9b00092.

(123) Farnaby, W. *et al.* BAF Complex Vulnerabilities in Cancer Demonstrated Via Structure-Based PROTAC Design. *Nature Chemical Biology* **2019**, *15*, 672–680. DOI: 10.1038/s41589-019-0294-6.

(124) Du, X.; Volkov, O. A.; Czerwinski, R. M.; Tan, H.; Huerta, C.; Morton, E. R.; Rizzi, J. P.; Wehn, P. M.; Xu, R.; Nijhawan, D.; Wallace, E. M. Structural Basis and Kinetic Pathway of RBM39 Recruitment to DCAF15 by a Sulfonamide Molecular Glue E7820. *Structure* **2019**, *27*, 1625–1633.e3. DOI: 10.1016/j.str.2019.10.005.

(125) Lai, A. C.; Toure, M.; Hellerschmied, D.; Salami, J.; Jaime-Figueroa, S.; Ko, E.; Hines, J.; Crews, C. M. Modular PROTAC Design for the Degradation of Oncogenic BCR-ABL. *Angewandte Chemie International Edition* **2016**, *55*, 807–810. DOI: 10.1002/anie.201507634.

(126) Bondeson, D. P.; Smith, B. E.; Burslem, G. M.; Buhimschi, A. D.; Hines, J.; Jaime-Figueroa, S.; Wang, J.; Hamman, B. D.; Ishchenko, A.; Crews, C. M. Lessons in PROTAC Design from Selective Degradation with a Promiscuous Warhead. *Cell Chemical Biology* **2018**, *25*, 78–87.e5. DOI: 10.1016/j.chembiol.2017.09.010.

(127) Donovan, K. A. *et al.* Mapping the Degradable Kinome Provides a Resource for Expedited Degrader Development. *Cell* **2020**, *183*, 1714–1731.e10. DOI: 10.1016/j.cell.2020.10.038.

(128) Spradlin, J. N. *et al.* Harnessing the Anti-Cancer Natural Product Nimbolide for Targeted Protein Degradation. *Nature Chemical Biology* **2019**, *15*, 747–755. DOI: 10.1038/s41589-019-0304-8.

(129) Ward, C. C.; Kleinman, J. I.; Brittain, S. M.; Lee, P. S.; Chung, C. Y. S.; Kim, K.; Petri, Y.; Thomas, J. R.; Tallarico, J. A.; McKenna, J. M.; Schirle, M.; Nomura, D. K. Covalent Ligand Screening Uncovers a RNF4 E3 Ligase Recruiter for Targeted Protein Degradation Applications. *ACS Chemical Biology* **2019**, *14*, 2430–2440. DOI: 10.1021/acschembio.8b01083.

(130) Zhang, X.; Crowley, V. M.; Wucherpfennig, T. G.; Dix, M. M.; Cravatt, B. F. Electrophilic PROTACs That Degrade Nuclear Proteins by Engaging DCAF16. *Nature Chemical Biology* **2019**, *15*, 737–746. DOI: 10.1038/s41589-019-0279-5.

(131) Kuljanin, M.; Mitchell, D. C.; Schweppe, D. K.; Gikandi, A. S.; Nusinow, D. P.; Bulloch, N. J.; Vinogradova, E. V.; Wilson, D. L.; Kool, E. T.; Mancias, J. D.; Cravatt, B. F.; Gygi, S. P. Reimagining High-Throughput Profiling of Reactive Cysteines for Cell-Based Screening of Large Electrophile Libraries. *Nature Biotechnology* **2021**, *39*, 630–641. DOI: 10.1038/s41587-020-00778-3.

(132) Li, W.; Bengtson, M. H.; Ulbrich, A.; Matsuda, A.; Reddy, V. A.; Orth, A.; Chanda, S. K.; Batalov, S.; Joazeiro, C. A. P. Genome-Wide and Functional Annotation of Human E3 Ubiquitin Ligases Identifies MULAN, a Mitochondrial E3 that Regulates the Organelle's Dynamics and Signaling. *PLOS ONE* **2008**, *3*, e1487. DOI: 10.1371/journal.pone.0001487.

(133) Jevtić, P.; Haakonsen, D. L.; Rapé, M. An E3 Ligase Guide to the Galaxy of Small-Molecule-Induced Protein Degradation. *Cell Chemical Biology* **2021**, DOI: 10.1016/j.chembiol.2021.04.002.

(134) Scholes, N. S.; Mayor-Ruiz, C.; Winter, G. E. Identification and Selectivity Profiling of Small-Molecule Degraders Via Multi-Omics Approaches. *Cell Chemical Biology* **2021**, DOI: 10.1016/j.chembiol.2021.03.007.

(135) Huang, H.-T. *et al.* A Chemoproteomic Approach to Query the Degradable Kinome Using a Multi-kinase Degrader. *Cell Chemical Biology* **2018**, *25*, 88–99.e6. DOI: 10.1016/j.chembiol.2017.10.005.

(136) Rodriguez-Rivera, F. P.; Levi, S. M. Unifying Catalysis Framework to Dissect Proteasomal Degradation Paradigms. *ACS Central Science* **2021**, DOI: 10.1021/acscentsci.1c00389.

(137) Maniaci, C.; Hughes, S. J.; Testa, A.; Chen, W.; Lamont, D. J.; Rocha, S.; Alessi, D. R.; Romeo, R.; Ciulli, A. Homo-PROTACs: Bivalent Small-Molecule Dimerizers of the VHL E3 Ubiquitin Ligase to Induce Self-Degradation. *Nature Communications* **2017**, *8*, 830. DOI: 10.1038/s41467-017-00954-1.

(138) Chan, K.-H.; Zengerle, M.; Testa, A.; Ciulli, A. Impact of Target Warhead and Linkage Vector on Inducing Protein Degradation: Comparison of Bromodomain and Extra-Terminal (BET) Degraders Derived from Triazolodiazepine (JQ1) and Tetrahydroquinoline (I-BET726) BET Inhibitor

Scaffolds. *Journal of Medicinal Chemistry* **2018**, *61*, 504–513. DOI: 10.1021/acs.jmedchem.6b01912.

(139) Zorba, A. *et al.* Delineating the Role of Cooperativity in the Design of Potent PROTACs for BTK. *Proceedings of the National Academy of Sciences* **2018**, *115*, E7285–E7292. DOI: 10.1073/pnas.1803662115.

(140) Schiemer, J. *et al.* Snapshots and Ensembles of BTK and cIAP1 Protein Degrader Ternary Complexes. *Nature Chemical Biology* **2021**, *17*, 152–160. DOI: 10.1038/s41589-020-00686-2.

(141) Drummond, M. L.; Henry, A.; Li, H.; Williams, C. I. Improved Accuracy for Modeling PROTAC-Mediated Ternary Complex Formation and Targeted Protein Degradation via New In Silico Methodologies. *Journal of Chemical Information and Modeling* **2020**, *60*, 5234–5254. DOI: 10.1021/acs.jcim.0c00897.

(142) Zaidman, D.; Prilusky, J.; London, N. PRosettaC: Rosetta Based Modeling of PROTAC Mediated Ternary Complexes. *Journal of Chemical Information and Modeling* **2020**, *60*, 4894–4903. DOI: 10.1021/acs.jcim.0c00589.

(143) Weng, G.; Li, D.; Kang, Y.; Hou, T. Integrative Modeling of PROTAC-Mediated Ternary Complexes. *Journal of Medicinal Chemistry* **2021**, *64*, 16271–16281. DOI: 10.1021/acs.jmedchem.1c01576.

(144) Bai, N.; Miller, S. A.; Andrianov, G. V.; Yates, M.; Kirubakaran, P.; Karanicolas, J. Rationalizing PROTAC-Mediated Ternary Complex Formation Using Rosetta. *Journal of Chemical Information and Modeling* **2021**, *61*, 1368–1382. DOI: 10.1021/acs.jcim.0c01451.

(145) Bai, N.; Riching, K. M.; Makaju, A.; Wu, H.; Acker, T. M.; Ou, S.-C.; Zhang, Y.; Shen, X.; Bulloch, D.; Rui, H.; Gibson, B.; Daniels, D. L.; Urh, M.; Rock, B.; Humphreys, S. C. Modeling the CRL4A Ligase Complex to Predict Target Protein Ubiquitination Induced by Cereblon-Recruiting PROTACs. *Journal of Biological Chemistry* **2022**, 101653. DOI: 10.1016/j.jbc.2022.101653.

(146) Moreira, I. S.; Fernandes, P. A.; Ramos, M. J. Protein–Protein Docking Dealing with the Unknown. *Journal of Computational Chemistry* **2010**, *31*, 317–342. DOI: 10.1002/jcc.21276.

(147) Gromiha, M. M.; Yugandhar, K.; Jemimah, S. Protein–Protein Interactions: Scoring Schemes and Binding Affinity. *Current Opinion in Structural Biology* **2017**, *44*, 31–38. DOI: 10.1016/j.sbi.2016.10.016.

(148) Bemis, T. A.; La Clair, J. J.; Burkart, M. D. Unraveling the Role of Linker Design in Proteolysis Targeting Chimeras. *Journal of Medicinal Chemistry* **2021**, DOI: 10.1021/acs.jmedchem.1c00482.

(149) Eron, S. J.; Huang, H.; Agafonov, R. V.; Fitzgerald, M. E.; Patel, J.; Michael, R. E.; Lee, T. D.; Hart, A. A.; Shaulsky, J.; Nasveschuk, C. G.; Phillips, A. J.; Fisher, S. L.; Good, A. Structural Characterization of Degrader-Induced Ternary Complexes Using Hydrogen–Deuterium Exchange Mass Spectrometry and Computational Modeling: Implications for Structure-Based Design. *ACS Chemical Biology* **2021**, DOI: 10.1021/acschembio.1c00376.

(150) Dixon, T. *et al.* Atomic-Resolution Prediction of Degrader-mediated Ternary Complex Structures by Combining Molecular Simulations with Hydrogen Deuterium Exchange. *bioRxiv* **2021**, ver. 1. DOI: 10.1101/2021.09.26.461830.

(151) Li, W.; Zhang, J.; Guo, L.; Wang, Q. Importance of Three-Body Problems and Protein–Protein Interactions in Proteolysis-Targeting Chimera Modeling: Insights from Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* **2022**, DOI: 10.1021/acs.jcim.1c01150.

(152) Liao, J.; Nie, X.; Unarta, I. C.; Ericksen, S. S.; Tang, W. In Silico Modeling and Scoring of PROTAC-Mediated Ternary Complex Poses. *Journal of Medicinal Chemistry* **2022**, *65*, 6116–6132. DOI: 10.1021/acs.jmedchem.1c02155.

(153) Niesen, M. J. M.; Wang, C. Y.; Lehn, R. C. V.; Miller, T. F. Structurally Detailed Coarse-Grained Model for Sec-Facilitated Co-Translational Protein Translocation and Membrane Integration. *PLOS Computational Biology* **2017**, *13*, e1005427. DOI: 10.1371/journal.pcbi.1005427.

(154) Zhang, B.; Miller, T. F. Long-Timescale Dynamics and Regulation of Sec-Facilitated Protein Translocation. *Cell Reports* **2012**, *2*, 927–937. DOI: 10.1016/j.celrep.2012.08.039.

(155) Hanke, F.; Serr, A.; Kreuzer, H. J.; Netz, R. R. Stretching Single Polypeptides: The Effect of Rotational Constraints in the Backbone. *EPL (Europhysics Letters)* **2010**, *92*, 53001. DOI: 10.1209/0295-5075/92/53001.

(156) Staple, D. B.; Payne, S. H.; Reddin, A. L. C.; Kreuzer, H. J. Model for Stretching and Unfolding the Giant Multidomain Muscle Protein Using Single-Molecule Force Spectroscopy. *Physical Review Letters* **2008**, *101*, 248301. DOI: 10.1103/PhysRevLett.101.248301.

(157) Cruje, C.; Chithrani, D. B. Polyethylene Glycol Density and Length Affects Nanoparticle Uptake by Cancer Cells. *Journal of Nanomedicine Research* **2014**, *1*, 00006. DOI: 10.15406/jnmr.2014.01.00006.

(158) Lezon, T. R.; Shrivastava, I. H.; Yang, Z.; Bahar, I. *Handbook on Biological Networks*; World Scientific Lecture Notes in Complex Systems, 2009; Vol. 10, DOI: 10.1142/9789812838803_0007.

(159) Ricardo Batista, P.; Herbert Robert, C.; Maréchal, J.-D.; Ben Hamida-Rebaï, M.; Geraldo Pascutti, P.; Mascarello Bisch, P.; Perahia, D. Consensus Modes, a Robust Description of Protein Collective Motions from Multiple-Minima Normal Mode Analysis—Application to the HIV-1 Protease. *Physical Chemistry Chemical Physics* **2010**, *12*, 2850–2859. DOI: 10.1039/B919148H.

(160) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *The Journal of Chemical Physics* **1935**, *3*, 300–313. DOI: 10.1063/1.1749657.

(161) Clark, A. J.; Gindin, T.; Zhang, B.; Wang, L.; Abel, R.; Murret, C. S.; Xu, F.; Bao, A.; Lu, N. J.; Zhou, T.; Kwong, P. D.; Shapiro, L.; Honig, B.; Friesner, R. A. Free Energy Perturbation Calculation of Relative Binding Free Energy between Broadly Neutralizing Antibodies and the gp120 Glycoprotein of HIV-1. *Journal of Molecular Biology* **2017**, *429*, 930–947. DOI: 10.1016/j.jmb.2016.11.021.

(162) Clark, A. J.; Negron, C.; Hauser, K.; Sun, M.; Wang, L.; Abel, R.; Friesner, R. A. Relative Binding Affinity Prediction of Charge-Changing Sequence Mutations with FEP in Protein–Protein Interfaces. *Journal of Molecular Biology* **2019**, *431*, 1481–1493. DOI: 10.1016/j.jmb.2019.02.003.

(163) Patel, D.; Patel, J. S.; Ytreberg, F. M. Implementing and Assessing an Alchemical Method for Calculating Protein–Protein Binding Free Energy. *Journal of Chemical Theory and Computation* **2021**, *17*, 2457–2464. DOI: 10.1021/acs.jctc.0c01045.

(164) La Serra, M. A.; Vidossich, P.; Acquistapace, I.; Ganesan, A. K.; De Vivo, M. Alchemical Free Energy Calculations to Investigate Protein–Protein Interactions: the Case of the CDC42/PAK1 Complex. *Journal of Chemical Information and Modeling* **2022**, *62*, 3023–3033. DOI: 10.1021/acs.jcim.2c00348.

(165) Nandigrami, P.; Szczepaniak, F.; Boughter, C. T.; Dehez, F.; Chipot, C.; Roux, B. Computational Assessment of Protein–Protein Binding Specificity within a Family of Synaptic Surface Receptors. *The Journal of Physical Chemistry B* **2022**, DOI: 10.1021/acs.jpcb.2c02173.

(166) Sun, Q.; Ramaswamy, V. S. K.; Levy, R.; Deng, N. Computational Design of Small Molecular Modulators of Protein–Protein Interactions with a Novel Thermodynamic Cycle: Allosteric Inhibitors of HIV-1 Integrase. *Protein Science* **2021**, *30*, 438–447. DOI: 10.1002/pro.4004.

(167) Testa, A.; Hughes, S. J.; Lucas, X.; Wright, J. E.; Ciulli, A. Structure-Based Design of a Macrocyclic PROTAC. *Angewandte Chemie International Edition* **2020**, *59*, 1727–1734. DOI: 10.1002/anie.201914396.

(168) Hendrick, C. E.; Jorgensen, J. R.; Chaudhry, C.; Strambeanu, I. I.; Brazeau, J.-F.; Schiffer, J.; Shi, Z.; Venable, J. D.; Wolkenberg, S. E. Direct-to-Biology Accelerates PROTAC Synthesis and the Evaluation of Linker Effects on

Permeability and Degradation. *ACS Medicinal Chemistry Letters* **2022**, *13*, 1182–1190. DOI: 10.1021/acsmedchemlett.2c00124.

(169) J. Maple, H.; Clayden, N.; Baron, A.; Stacey, C.; Felix, R. Developing Degraders: Principles and Perspectives on Design and Chemical Space. *MedChemComm* **2019**, *10*, 1755–1764. DOI: 10.1039/C9MD00272C.

(170) Ermondi, G.; Vallaro, M.; Caron, G. Degraders Early Developability Assessment: Face-to-Face with Molecular Properties. *Drug Discovery Today* **2020**, *25*, 1585–1591. DOI: 10.1016/j.drudis.2020.06.015.

(171) Gopalsamy, A. Selectivity through Targeted Protein Degradation (TPD). *Journal of Medicinal Chemistry* **2022**, *65*, 8113–8126. DOI: 10.1021/acs.jmedchem.2c00397.

(172) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; Wilkins-Diehr, N. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **2014**, *16*, 62–74. DOI: 10.1109/MCSE.2014.80.

(173) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 221–234. DOI: 10.1007/s10822-013-9644-8.

(174) Liwo, A.; Ołdziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. A United-Residue Force Field for Off-Lattice Protein-Structure Simulations. I. Functional Forms and Parameters of Long-Range Side-Chain Interaction Potentials from Protein Crystal Data. *Journal of Computational Chemistry* **1997**, *18*, 849–873. DOI: 10.1002/(SICI)1096-987X(19970 5)18:7<849::AID-JCC1>3.0.CO;2-R.

(175) Periole, X.; Cavalli, M.; Marrink, S.-J.; Ceruso, M. A. Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *Journal of Chemical Theory and Computation* **2009**, *5*, 2531–2543. DOI: 10.1021/ct9002114.

(176) Sievers, Q. L.; Petzold, G.; Bunker, R. D.; Renneville, A.; Słabicki, M.; Liddicoat, B. J.; Abdulrahman, W.; Mikkelsen, T.; Ebert, B. L.; Thomä, N. H. Defining the Human C2H2 Zinc Finger Degrome Targeted by Thalidomide Analogs Through CRBN. *Science* **2018**, *362*, eaat0572. DOI: 10.1126/science.aat0572.

(177) Wyce, A. *et al.* BET Inhibition Silences Expression of MYCN and BCL2 and Induces Cytotoxicity in Neuroblastoma Tumor Models. *PLOS ONE* **2013**, *8*, e72967. DOI: 10.1371/journal.pone.0072967.