

Bayesian inference of macromolecular ensembles

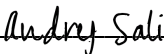
by
Matthew Hancock

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in
Biophysics


in the
GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

8F7A6AB94F2C4F4... Andrej Sali
Chair

DocuSigned by:

Matthew Jacobson

DocuSigned by:

430BBB9A04D24A3... James Fraser

Committee Members

Copyright 2024
by
Matthew Hancock

Acknowledgements

Andrej Sali, Ph.D.

Matt Jacobson, Ph.D., James Fraser, Ph.D.

Paul Hancock, Jeannie Hancock

Contributions

Chapter 2

Hancock, M., Peulen, T.-O., Webb, B., Fraser, J. S., Adams, P. & Sali, A. Integration of software tools for integrative modeling of biomolecular systems. en. *Biophysical Journal* **214** (2022)

Chapter 3

Taylor, M. S., Chen, M., Hancock, M., Wranik, M., Miller, B. D., Palanski, B. A., Ficarro, S. B., Groendyke, B. J., Xiang, Y., Linde-Garelli, K. Y., Mondal, D., Freund, D., Congreve, S., Matas, K., Hennink, M., Xibinaku, K., Marto, J. A., Sali, A., Shi, Y., Gray, N. S., Sabatini, D. M., Chu, N., Rogala, K. B. & Cole, P. A. A Long-Range Recruitment Mechanism for mTORC2 Phosphorylation of Akt. en

Chapter 4

Hancock, M., Fraser, J. S., Adams, P. D. & Sali, A. Bayesian multi-state multi-condition modeling of a protein structure from X-ray crystallography. en

Bayesian inference of macromolecular ensembles

Matthew Hancock

Abstract

Models of protein structure at atomic resolution are fundamental to many problems in biology. Increasingly, we are interested in computing models of protein structure based on noisy, sparse, and heterogeneous data; demanding a rigorous approach to modeling. Bayesian inference is one such approach. In Chapter 1, we formalize modeling as a search for all models consistent with the input information. In Chapter 2, we develop a computational framework for computing models from heterogeneous biophysical software. In Chapter 3, we develop a modeling method to compute a model of mTORC2 in complex with a native substrate, Akt1. In Chapter 4, we develop a modeling method to compute multi-state models from X-ray crystallography.

Contents

1	Modeling as an optimization problem	1
1.1	Gathering of Input information	2
1.2	Model search	3
1.2.1	Defining the model representation	3
1.2.2	Specification of a scoring function	4
1.2.3	Sampling good scoring models	5
1.3	Model validation	6
	References	7
2	Integration of software tools for integrative modeling of biomolecular systems	9
2.1	Abstract	9
2.2	Introduction	10
2.2.1	Introduction to integrative modeling	10
2.2.2	Developing software modeling tools is challenging	12
2.2.3	Collaborative development maximizes the efficiency and quality of modeling	12
2.2.4	Integration of software tools is an efficient approach to collaborative development	13
2.2.5	Why is integration of software tools challenging?	13

2.2.6	Opportunity for standardization	13
2.2.7	Article overview	14
2.3	Approach	15
2.3.1	Modeling as a search	15
2.3.2	Multistate model of Nup133 computed by a model search	15
2.3.3	Informed vs uninformed search	16
2.3.4	Demonstrative software standards (Fig. 4.2)	16
2.3.5	Illustrative modeling problem	18
2.3.6	<i>Phenix</i>	21
2.3.7	IMP	21
2.3.8	Integration of <i>Phenix</i> and IMP (Fig. 4.3)	22
2.3.9	Conclusions	24
2.4	Discussion	28
2.5	Conclusions	30
	References	31

3	A Long-Range Recruitment Mechanism for mTORC2 Phosphorylation of Akt	37
3.1	Abstract	37
3.2	Introduction	38
3.3	Results	42
3.3.1	mTORC2 directly phosphorylates Akt at Ser473	42
3.3.2	Specificity Determinants of mTORC2 Catalyzed Phosphorylation of Akt	43
3.3.3	Akt-ATP as a bisubstrate inhibitor of mTORC2	44
3.3.4	Akt-Torin as a bivalent ligand for mTORC2	48
3.3.5	Cryo-EM analysis of mTORC2:Akt-Torin co-complex	49
3.3.6	Interface 1: Akt N-Lobe – mSin1 CRIM	50

3.3.7	Interface 2: Akt C-Lobe – mSin1 N-Mooring	54
3.3.8	Integrative Modeling of the mTORC2:Akt co-Complex	57
3.3.9	Membrane recruitment in mTORC2’s phosphorylation of Akt	60
3.3.10	Mechanistic similarity in mTORC2 recognition and direct activation of Akt, PKC, and PKN	61
3.4	Discussion	64
3.5	Material and Methods	68
3.5.1	Reagents	68
3.5.2	Antibodies	69
3.5.3	Tissue Culture	69
3.5.4	Akt and PKC Cell Signaling Assay in HCT116 cells	70
3.5.5	Akt, PKC, and PKN Cell Signaling Assays in HEK293T cells	71
3.5.6	Generation of cells with loss-of-function mutations of mTORC2 com- ponents Rictor or mSin1	71
3.5.7	Molecular Cloning and Plasmids	72
3.5.8	Bacterial Expression and Purification of mSin1 fragments and 4E-BP1	73
3.5.9	mTORC2 Expression in Suspension HEK-293T Cells	74
3.5.10	mTORC2 Expression in Insect Cells	74
3.5.11	mTORC2 Purification	74
3.5.12	mTORC1 Expression and Purification	75
3.5.13	Small Uni-lamellar Vesicle (SUV) Preparation	75
3.5.14	Ultracentrifugation Vesicle Binding Assay	76
3.5.15	<i>In Vitro</i> Akt Phosphorylation Assay with mTORC2 and mTORC1	76
3.5.16	<i>In Vitro</i> Akt Phosphorylation Assay with γ 32P-ATP	77
3.5.17	<i>In Vitro</i> 4E-BP1 Phosphorylation Assay with mTORC1 and mTORC2	78
3.5.18	mTORC2-Akt reactions for mass spectrometry analysis	78
3.5.19	nanoLC-MS analysis of AKT reactions	79

3.5.20	Peptide Synthesis	80
3.5.21	Semisynthesis of Akt Proteins	80
3.5.22	Preparation of Lambda Phosphatase Dephosphorylated Akt	81
3.5.23	Semisynthesis of Akt-ATP Protein Bisubstrate	81
3.5.24	Synthesis of Torin Acrylamide	82
3.5.25	Protein Cross-linking and Mass Spectrometry (XL-MS)	85
3.5.26	Integrating Modeling of mTORC2-Akt	86
3.5.27	mTORC2 and mTORC2:Akt co-Complex Preparation for cryo-EM	92
3.5.28	Cryo-EM Image Processing	93
3.5.29	Cryo-EM Model Building and Refinement	94
3.5.30	Data availability	95
	References	122

4 Bayesian multi-state multi-condition modeling of a protein structure from X-ray crystallography 136

4.1	Abstract	136
4.2	Introduction	137
4.2.1	A protein crystal is a heterogeneous mix of structural states	137
4.2.2	Approaches to modeling a heterogenous mix of structural states	138
4.2.3	Multi-condition crystallography	138
4.2.4	Computing a multi-state multi-condition model	139
4.3	Methods	140
4.3.1	Overview of modeling method	140
4.3.2	Input information	140
4.3.3	Representation	142
4.3.4	Scoring	142
4.3.5	Sampling	147
4.3.6	Software availability	149

4.4	Results	150
4.4.1	Synthetic Benchmark	150
4.4.2	Sampling convergence	153
4.4.3	Benchmark: the shape of the scoring function	154
4.4.4	Multi-state multi-condition modeling of SARS-CoV-2 M ^{pro}	155
4.5	Discussion	159
	References	163

List of Figures

2.1	Coverage of molecular weight by structural technique.	11
2.2	Class diagram for proposed standard for implementing a model search.	18
2.3	Workflow for integrating modeling software, exemplified by <i>Phenix</i> and IMP.	25
2.4	Sample of SARS-CoV main protease posterior model density evaluated through integration of IMP and <i>Phenix</i>	26
3.1	mTORC2 directly and specifically phosphorylates Ser473 of Akt with little contribution from local Akt sequence	39
3.2	mTORC2-Akt Complex Structure Determined Using Akt-Torin	46
3.3	Interplay between Thr450 Phosphorylation and mTORC2-Akt Interface 1: Akt N-Lobe – mSin1 CRIM	51
3.4	mTORC2-Akt Interface 2: Akt C-Lobe – mSin1 N-mooring	55
3.5	Integrative Modeling and Membrane Association of mTORC2-Akt	58
3.6	mTORC2 Similarly Phosphorylates AGC Kinases at Both Turn and Hydrophobic Motifs	62
3.7	Extended Data Figure 1.1	96
3.8	Extended Data Figure 1.2	97
3.9	Extended Data Figure 1.3	98
3.10	Extended Data Figure 2.1	99
3.11	Extended Data Figure 2.2	100
3.12	Extended Data Figure 2.3	101

3.13	Extended Data Figure 2.4	102
3.14	Extended Data Figure 2.5	103
3.15	Extended Data Figure 2.6	104
3.16	Extended Data Figure 2.7	105
3.17	Extended Data Figure 3.1	106
3.18	Extended Data Figure 3.2	107
3.19	Extended Data Figure 4.1	108
3.20	Extended Data Figure 5.1	109
3.21	Extended Data Figure 5.2	110
3.22	Extended Data Figure 5.3	111
3.23	Extended Data Figure 6.1	112
3.24	Supplemental Figure 1.1	113
3.25	Supplemental Figure 1.2	114
3.26	Supplemental Figure 2.1	115
3.27	Supplemental Figure 2.2	116
3.28	Supplemental Figure 2.3	117
3.29	Supplemental Figure 2.4	118
3.30	Supplemental Figure 2.5	119
3.31	Supplemental Figure 2.6	120
3.32	Supplemental Figure 2.7	121
4.1	Method overview.	141
4.2	Degeneracy of the multi-state forward model	146
4.3	Synthetic benchmark	151
4.4	Multi-state multi-condition modeling of SARS-CoV-2 M^{pro}	156
4.5	Analysis of multi-state multi-condition modeling	158

Chapter 1

Modeling as an optimization problem

Scientific knowledge is gained through the relationship between experiments and models. A model is a depiction of the state of knowledge about a system or process. A model can be used to make predictions about future observations. An experiment will yield an observation that is either consistent or inconsistent with the model's predictions. The new observations can then be used in modeling to update the model.

The reason for the distinction between an observation and a model is that a model is an information structure that is generally more useful for making predictions. For example, the intensities of diffracted X-rays from a protein crystal are not useful for predicting the impact of mutations on the protein's function. On the other hand, a model of the protein's structure at atomic resolution is very useful for such questions. Here, the information contained by the observed diffraction and the protein model is the same, yet the protein model has an information structure that is far more useful in predicting new observations.

The process of modeling is a mapping of a space containing less useful objects (data space) to a space containing more useful objects (model space). The prediction of observations by a model is the inverse mapping. It is often the case that the mapping between data and model space is not one to one. In other words, a point in data space may map to many points in model space and *vice versa*. Many experimental observations contain noise and therefore we

aim to transform a distribution of points in data space to a distribution of points in model space.

Modeling can be formalized as the search for all models consistent with the input information¹. The model search can be formalized by the following steps: (i) specifying all model variables (representation), (ii) ranking alternative models by their agreement with the input information (scoring), and (iii) generating a sample of good-scoring models (sampling). A model should be validated before being interpreted. Multiple iterations of gathering input information, modeling, and validation are often necessary to compute a sufficiently precise model.

1.1 Gathering of Input information

Input information comes in the form of data and prior models and may be used to inform any step of the modeling process (representation, scoring, and sampling). By including all available input information, the model is maximally accurate, precise, and complete¹. In molecular biology, such an integrative approach is often used. For example, the double-helical structure of DNA was resolved through a fiber X-ray diffraction pattern, data about the composition and stoichiometry of the component nucleotides, and theoretical information about nucleotide complementarity².

A prior model is anything known about the model before data collection and is often the result of prior modeling. Most data involves several steps of abstraction before it is used in modeling, and thus, the distinction between a prior model and data is somewhat arbitrary. Three important prior models used in structural modeling are molecular mechanics force fields, statistical potentials, and previously computed structural models. Molecular mechanics force fields are mathematical models that describe the physical properties of proteins fit to spectroscopy, quantum mechanical calculations, and experimental energies³. Statistical potentials impute the statistical distribution of certain properties of protein structures from

large databases of previously computed models⁴. Machine learning algorithms for protein structure prediction rely on learning such patterns. Previously computed protein structure models, whether a prediction or determined from an experiment, are useful in a number of structural modeling problems. For example, a previously determined model may be used to provide a guess of the phases for X-ray diffraction patterns⁵ or flexibly fitted into a cryo-electron microscopy (Cryo-EM) density map⁶.

Historically, a model of a protein structure at atomic resolution was computed from data collected from X-ray crystallography, cryo electron microscopy (cryo-EM), or nuclear magnetic resonance (NMR) spectroscopy⁷. It is increasingly of interest to compute models of large, heterogeneous, and dynamic protein systems unsuitable to the above biophysical methods⁸. Such systems can often only be characterized by noisy, sparse, and incomplete experiments. Examples include cross-linking mass spectrometry (XL-MS), negative stain EM microscopy, small angle X-ray scattering (SAXS), hydrogen-deuterium exchange mass spectrometry (HDX), and Förster resonance energy transfer (FRET)¹.

1.2 Model search

1.2.1 Defining the model representation

The model representation includes all model coordinates. The span of all degrees of freedom is a space containing all possible models (model space). Any point in the model space is a model and the model space may be continuous or discrete. Based on the input information, constraints may also be placed on model coordinates to limit the space.

The choice of the model coordinates to include is guided by the question being answered. The following considerations should guide the representation choice: (i) the model representation should be maximally simple, therefore maximizing the explanatory power and generality of the model. Kepler sought an analytical form to describe the elliptical orbits of planets rather than fitting a neural network. (ii) the model resolution should also be

commensurate with the resolution of the input information. It would be a poor choice of model representation to fit a fully parameterized atomic model to a 30Å map from negative stain electron microscopy. (iii) the model representation should be selected such that the available computational power may adequately sample it. It is often not possible to compute a sufficiently precise sample of a high-dimensional landscape on a single CPU core. The selection of the optimal model representation is an active area of research. Ideally, the type and number of model variables may be treated as a model variable.

Historically, a protein structure model included atomic coordinates, temperature factors, and atomic occupancies. Because of the increased complexity of the studied protein systems, new experimental techniques, and increased computational power, it is becoming more common to expand the description of the target system through the addition of model variables. 2 areas of particular importance to protein models are including variables to describe the system in a multi-scale and multi-state fashion⁸. A multi-scale model will describe the target system at multiple scales; for example, the model representation may include model variables describing the individual atoms and entire domains as rigid bodies. Different parts of the protein may be represented at different scales commensurate to the resolution of the input information. A multi-state model describes the protein’s various states and their transitions. A state may describe a distinct conformational or compositional state the protein exists. The transitions might represent a thermodynamic cycle or time-ordered assembly⁸.

1.2.2 Specification of a scoring function

The scoring function ranks all model configurations in the model space for compatibility with the input information. A Bayesian posterior model density is the most objective scoring function¹. The posterior model density is the probability of a model M conditional on the data D and prior models I :

$$p(M|D, I) \propto p(D|M, I) \times p(M|I) \tag{1.1}$$

The posterior model density is factored into the likelihood and prior. The prior, the probability of M given I , represents the state of knowledge about the model before data collection, and the likelihood, the probability of D given M and I represents what is learned from data observation.

It is often convenient to factor $p(D|M, I)$ into a forward model $f(M)$ that simulates the data observation for a given model in the absence of noise and a noise model $N(D; f(M), \sigma)$ that quantifies the difference between the noiseless prediction and the data measurement given some nuisance parameters σ .

As the posterior model density is a probability distribution over all possible models, it allows for the explicit quantification of uncertainty. Ideally, the addition of more and more input information would narrow the distribution, increasing certainty.

The iterative refinement of Bayesian inference reflects the way the human mind learns. As more observations are made, the state of knowledge increases; reflected by the increased certainty or accuracy of the posterior model density. Similarly, the mind stores models on all aspects of how the world works and humans act based on the prediction of said models. The models are constantly tested and updated based on new experiences and observations. For example, we have a model where if we throw a ball in the air; it will fall to the ground. Every instance where we observe the ball falling to the ground increases our confidence or certainty in the model. If, however, we were to throw the ball in the air and it did not fall to ground, our certainty in the model would dramatically decrease.

1.2.3 Sampling good scoring models

The goal of sampling is to find all models consistent with input information (ensemble of good-scoring models). Since it is often impossible to enumerate the model space, a sample of models is drawn from it.

Three important algorithms for computing a sample of good scoring protein structure models are conjugate gradients, Molecular dynamics (MD) simulation, and Markov chain

Monte Carlo (MCMC)⁹. Conjugate gradients minimizes an objective function by updating model variables in the opposite direction of the partial gradient with respect to the model variables and is useful for structure refinement¹⁰. MD simulation can generate samples of low energy structure models by predicting the movement of atoms from a model of the interatomic interactions¹¹. Finally, MCMC can draw accurate samples from an unknown distribution of some protein structure model coordinate(s) and is particularly useful when there are large energy barriers in the scoring function¹². These methods have several notable variations, including replica exchange and simulated annealing^{13,14}.

Sampling should continue until convergence is achieved. One test for sample convergence is ensuring sufficient sample precision¹⁵. A model should not be interpreted with a resolution greater than the sample precision.

1.3 Model validation

A model should be validated before being interpreted. Validation should include checking that a model satisfies data used in modeling and data withheld from modeling. In X-ray crystallography, it is common to withhold some fraction of the observed reflections as a test set¹⁶. Computing the satisfaction of reflections withheld from the modeling process ensures that the model is not overfit to the reflections used in modeling. It is also common to use other input information, such as the Ramachandran plot, to assess the quality of a structure model.

References

1. Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. en. *Cell* **177**, 1384–1403 (May 2019) (cit. on pp. 2–4).
2. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. en. *Nature* **171**, 737–738 (Apr. 1953) (cit. on p. 2).
3. Dauber-Osguthorpe, P. & Hagler, A. T. Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? en. *J. Comput. Aided Mol. Des.* **33**, 133–203 (Feb. 2019) (cit. on p. 2).
4. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. en. *Protein Sci.* **15**, 2507–2524 (Nov. 2006) (cit. on p. 3).
5. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. Phaser crystallographic software. en. *J. Appl. Crystallogr.* **40**, 658–674 (Aug. 2007) (cit. on p. 3).
6. Kim, D. N., Moriarty, N. W., Kirmizialtin, S., Afonine, P. V., Poon, B., Sobolev, O. V., Adams, P. D. & Sanbonmatsu, K. Cryo_fit: Democratization of flexible fitting for cryo-EM. en. *J. Struct. Biol.* **208**, 1–6 (Oct. 2019) (cit. on p. 3).
7. Hancock, M., Peulen, T.-O., Webb, B., Poon, B., Fraser, J. S., Adams, P. & Sali, A. Integration of software tools for integrative modeling of biomolecular systems. en. *J. Struct. Biol.* **214**, 107841 (Mar. 2022) (cit. on p. 3).
8. Sali, A. From integrative structural biology to cell biology. en. *J. Biol. Chem.* **296**, 100743 (Jan. 2021) (cit. on pp. 3, 4).
9. Liwo, A., Czaplowski, C., Ołdziej, S. & Scheraga, H. A. Computational techniques for efficient conformational sampling of proteins. en. *Curr. Opin. Struct. Biol.* **18**, 134–139 (Apr. 2008) (cit. on p. 6).

10. Ruder, S. An overview of gradient descent optimization algorithms (Sept. 2016) (cit. on p. 6).
11. Hollingsworth, S. A. & Dror, R. O. Molecular Dynamics Simulation for All. en. *Neuron* **99**, 1129–1143 (Sept. 2018) (cit. on p. 6).
12. Robert, C. P. & Changye, W. Markov Chain Monte Carlo Methods, a survey with some frequent misunderstandings (Jan. 2020) (cit. on p. 6).
13. Khachatryan, A., Semenovskaya, S. & Vainstein, B. Statistical-thermodynamic approach to determination of structure amplitude phases. *Sov. Phys. Crystallogr.* (cit. on p. 6).
14. Swendsen, R. H. & Wang, J. S. Replica Monte Carlo simulation of spin glasses. en. *Phys. Rev. Lett.* **57**, 2607–2609 (Nov. 1986) (cit. on p. 6).
15. Viswanath, S., Chemmama, I. E., Cimermancic, P. & Sali, A. Assessing Exhaustiveness of Stochastic Sampling for Integrative Modeling of Macromolecular Structures. en. *Biophys. J.* **113**, 2344–2353 (Dec. 2017) (cit. on p. 6).
16. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. en. *Nature* **355**, 472–475 (Jan. 1992) (cit. on p. 6).

Chapter 2

Integration of software tools for integrative modeling of biomolecular systems

2.1 Abstract

Integrative modeling computes a model based on varied types of input information, be it from experiments or prior models. Often, a type of input information will be best handled by a specific modeling software package. In such a case, we desire to integrate our integrative modeling software package, *Integrative Modeling Platform* (IMP), with software specialized to the computational demands of the modeling problem at hand. After several attempts, however, we have concluded that even in collaboration with the software's developers, integration is either impractical or impossible. The reasons for the intractability of integration include software incompatibilities, differing modeling logic, the costs of collaboration, and academic incentives. In the integrative modeling software ecosystem, several large modeling packages exist with often redundant tools. We reason, therefore, that the other development groups have similarly concluded that the benefit of integration does not justify the cost. As

a result, modelers are often restricted to the set of tools within a single software package. The inability to integrate tools from distinct software negatively impacts the quality of the models and the efficiency of the modeling. As the complexity of modeling problems grows, we seek to galvanize developers and modelers to consider the long-term benefit that software interoperability yields. In this article, we formulate a demonstrative set of software standards for implementing a model search using tools from independent software packages, and discuss our efforts to integrate the IMP and the crystallography suite *Phenix* within the Bayesian modeling framework.

2.2 Introduction

2.2.1 Introduction to integrative modeling

Integrative modeling combines information of different types into a model^{1,2}. When all available information is used, the accuracy, precision, and completeness of the model are maximized. An example of an integrative model is the double-helical structure of DNA, which could only be resolved through a joint consideration of a fiber X-ray diffraction pattern of the DNA, data about composition and stoichiometry of the component nucleotides, as well as theoretical information about physiochemical nucleotide complementarity³. Modern integrative modeling of biomolecular structures similarly considers experimental data (e.g., an X-ray diffraction pattern, a cryo-electron microscopy (cryo-EM) density map, and nuclear magnetic resonance (NMR) spectra) and prior information (e.g., a molecular mechanics force field, a statistical potential, and previously obtained structural models). As the complexity (e.g., size, resolution, heterogeneity, and dynamics) of biomolecular structural models grows, integration of diverse and often sparse experimental data will be critical for maximally exploiting experimental techniques and their complementarities (**Fig. 4.1**)⁴.

Figure 1

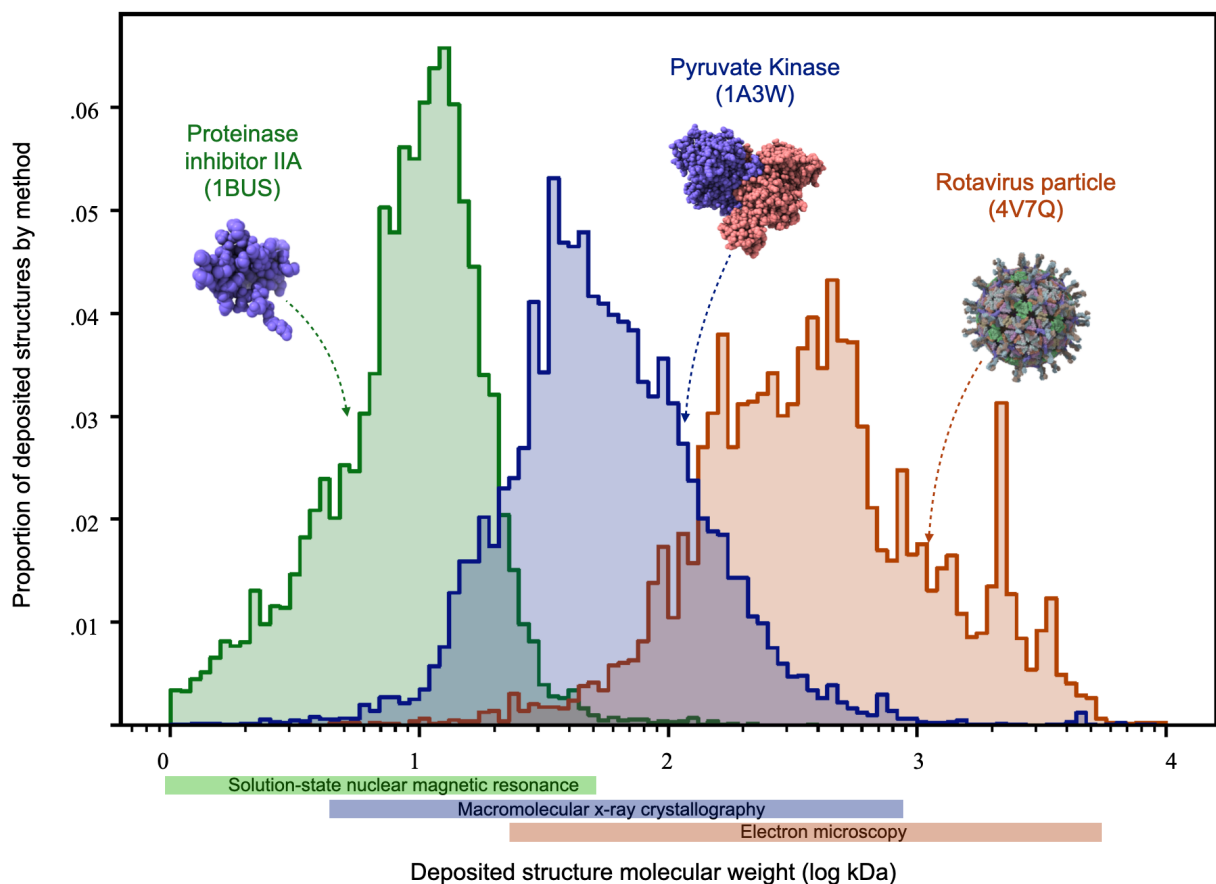


Figure 2.1: Coverage of molecular weight by structural technique. Histogram of the molecular weight of structures resolved by solution-state NMR spectroscopy, macromolecular X-ray crystallography, and electron microscopy (EM) deposited in the Protein Data Bank (PDB)⁵. Each technique has a different coverage with respect to the molecular weight of the studied system. This variation illustrates one reason why it is desirable to integrate varied types of information as well as the software tools used to compute models from them.

2.2.2 Developing software modeling tools is challenging

Modeling is often only possible through computation using software tools. The development of software tools for modeling is challenging because it demands both domain knowledge and technical expertise. Domain knowledge encompasses both an understanding of the general concepts of the field and a thorough understanding of both the theory and practices of the leveraged experimental techniques. The technical prerequisites include the ability to formulate the theory into stable numerical algorithms and to develop software of sufficient quality. Moreover, software needs to incorporate new technical and scientific advancements. Fulfilling the above prerequisites is difficult when developing software for an individual experimental technique. The challenge is compounded in the development of integrative modeling software, where several sources of experimental information generally are combined. Nevertheless, there is a large number of independent and complex integrative modeling software packages⁶⁻²⁰, as reviewed⁴.

2.2.3 Collaborative development maximizes the efficiency and quality of modeling

Given the technical and scientific difficulty of developing software modeling tools, it is desirable to tackle integrative modeling as a collaborative development effort of multiple research groups and development teams. Such collaborative development has the benefits of better integration of domain expertise into modeling software. There is also a benefit in the distribution of development costs over multiple research groups. For these reasons, collaborative software development maximizes both the efficiency of the software development and the quality of the resulting models.

2.2.4 Integration of software tools is an efficient approach to collaborative development

One way to achieve collaborative development is by developing software that is designed to solve a modeling problem through integration with one or more additional modeling tools. Such integration can occur between tools within and across integrative modeling software packages. The ability to combine the tools from existing software packages can extend the set of methods that can be implemented with these tools. It is often better to approach an integrative modeling problem by mixing and matching existing tools rather than by building new ones because the developer benefits from prior work. For example, Rosetta¹⁴ integrates high-level tools within Rosetta through RosettaScripts²¹ and lower-level Rosetta functionality with tools outside of Rosetta through the C++/Python layers, such as `phenix.rosetta_refine`²².

2.2.5 Why is integration of software tools challenging?

There are numerous structural biology software packages, including tens of integrative modeling programs alone². While there is consensus that software interoperability is beneficial, little has been done to address the issue. It is often attractive to implement new tools within one's own software ecosystem, even if similar tools already exist elsewhere. It is not unusual to see, for example, unique implementations of vector classes within modeling packages. Possible reasons include the difficulty of coordinating a large number of contributors, lack of support and motivation for rigorous development and maintenance standards for academic software, and the incentive for individual programmers and research groups to publish new software.

2.2.6 Opportunity for standardization

As the complexity of integrative modeling problems grows, they will be increasingly diffi-

cult to solve by relying on tools from a single software package. Integration can be done on an ad hoc basis, where specific tools are combined when needed. Such an integration is not efficient, however, as the pool of tools grows because any integration would incur additional development costs. An alternative is the adoption of software standards for tools developed by the integrative modeling community. Software standards offer guarantees on some aspect of the software’s implementation or function. Well-defined standards would benefit the integrative modeling field by ensuring that software tools are of sufficient quality and generality to be interoperable with other software tools that adhere to the standards. If a standard is adopted by the community, an integrative modeler may be able to readily mix-and-match software that has been produced from multiple development groups. Similar to the PDB/mmCIF standards for archival, the nature and extent of the standards must be agreed upon by the integrative modeling development community.

2.2.7 Article overview

In this article, we develop a demonstrative software standard for integrating one or more independent tools in a model search. We begin by describing modeling as a 5-step search for a model that satisfies input information and distinguishing between informed and uninformed model searches. We then describe how the model search is achieved through basic function definitions. We illustrate the standards by integrating *Phenix* and *Integrative Modeling Platform* (IMP) for computing an atomic model from X-ray crystallography datasets and a molecular mechanics force field. Key to the integration is the factorization of the model posterior density into likelihoods and priors. We discuss our attempts to integrate IMP and *Phenix* tools as independent processes via file input/output and within a single process using both libraries’ application programming interface (API). We conclude by discussing how software integration may increase the quality and potential complexity of the model as well as the efficiency of the modeling.

2.3 Approach

2.3.1 Modeling as a search

A model is a depiction of a system or process that we would like to inform from input information, consisting of experimental data and prior information². A model can then be used to rationalize input information and make testable predictions. Modeling is the search for a set of models consistent with the input information. Ideally, we aim to find all models that satisfy the input information, reflecting the uncertainty of the input information. It is convenient to divide the search into the following three steps: (i) defining the model representation that specifies all degrees of freedom whose values are determined by modeling, (ii) defining a scoring function for ranking alternative models for their agreement with the input information, and (iii) generating a sample of good-scoring models. As an aside, these models can be optionally filtered based on the input information and should also be validated before interpretation².

2.3.2 Multistate model of Nup133 computed by a model search

For example, a model search is used to compute a multi-state model of the Nup133 nucleoporin from small-angle x-ray scattering (SAXS), electron microscopy class averages, and cross-linking mass spectrometry (XL-MS) data²³. To reflect the structural heterogeneity of Nup133 in solution, the authors defined the model representation as an ensemble of fully atomic structures. The degrees of freedom, to be fit to the input information, include the number of models in the ensemble as well as the positions of atoms in each structure. Therefore, the objective of the model search is to find all Nup133 ensembles that satisfy the input information. A sample of Nup133 ensembles was generated via molecular dynamics simulations (MD) such that sufficient coverage of the energy landscape was achieved. The scoring function then evaluated the consistency of any model ensemble with the SAXS profile, the EM class averages, and chemical cross-links by simulating data via physical principles and

comparing it to the observed data. The model search framework is a general description of modeling that can describe most modeling protocols.

2.3.3 Informed vs uninformed search

Generally, integrative modeling software either explicitly or implicitly implements tools for each step of the model search. The model search can be categorized as either uninformed or informed, relative to some input information. In the uninformed search, candidate models are systematically generated to explore the search space without consideration for a specified subset of input information²⁴. For example, the minimal ensemble approach to computing protein structure ensembles based on SAXS data generates an ensemble of structures without consideration of the experimental data²⁵. In the informed search, a specified subset of input information is used to bias the generation of solutions²⁴. For example, partial derivatives based on the SAXS data could be used to guide sampling.

It is easier to isolate software tools in an uninformed search because outputs of relevant modeling steps can be combined as an additional post-processing step. However, due to a large model space generally required to be searched when solving integrative modeling problems, we are interested in the integration of software tools that enable informed search, in addition to the uninformed search. Informed search demands the passage of information between the tools that implement representation, scoring, and sampling during the modeling procedure.

2.3.4 Demonstrative software standards (Fig. 4.2)

While tools within a given software package generally pass information to each other, they have not been designed to do so across different software packages. Interoperability may sometimes be achieved by engineering connections between a particular set of software tools in an ad hoc fashion. However, to maximize generalizability, it is better if information is communicated in well-defined channels defined by a software standard. Here, we develop basic

standards to illustrate how information passing could be accomplished between independent modeling tools.

To facilitate the integration of two modeling tools, they should share the minimal amount of information necessary. It is also desirable that the tools be limited in scope to maximize modularity. For example, if a tool implemented a specific modeling step (i.e., representation, scoring, sampling, and optionally filtering plus validation). Tools with a well-defined purpose that hide their implementation details help manage the technical complexity of an integrative modeling problem. In the example of the Nup133 model search, a distinct software tool could be used to implement the multi-state model representation, the molecular dynamics sampler, the SAXS scoring function, the XL-MS scoring function, and the EM scoring function. The tools must be able to communicate with one another, but at the same time, they should be encapsulated from each other’s technical complexity.

To design message passing, we first define the function of a model representation, scoring function, and sampling algorithm tool based on an informed search. The model representation manages the model state (the current value of the model parameters), which we partition between structural, X , and nuisance parameters, σ . The model representation manages the model state at step i of the model search, $\{X, \sigma\}_i$, as well as previously visited states, $\{X, \sigma\}_1 \dots \{X, \sigma\}_{i-1}$. The scoring function computes the scores s , an assessment of the compatibility of $\{X, \sigma\}_i$ with input information, D . The scoring function also returns heuristics, h , as a function of D , which help inform the search process (eg, gradients for finding local minima in the search space), $f(\{X, \sigma\}_i, D) = s, h$. A sampling tool updates the model state based on $\{X, \sigma\}_i$, s , and h , $g(\{X, \sigma\}_i, s, h) = \{X, \sigma\}_{i+1}$.

Based on the above definitions, the minimal informed search is implemented by the following 4 functions that facilitate communication between the model representation, the scoring function, and the sampling algorithm tools (**Fig. 4.2**). First, the scoring function must be able to access the model state from the model representation (`get_state`). Second, the sampling algorithm must be able to access the current model parameters from the model

Figure 2

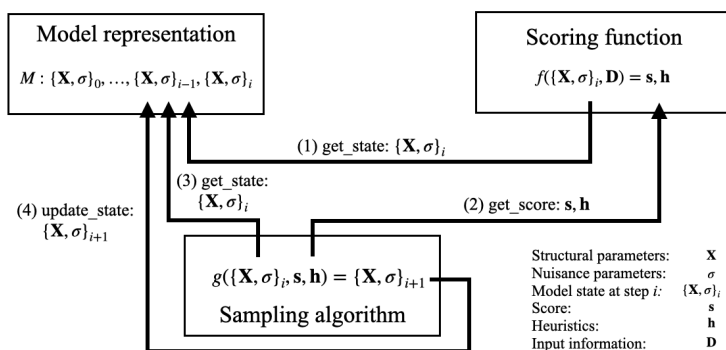


Figure 2.2: Class diagram for proposed standard for implementing a model search. A model search may be implemented through the communication of software tools implementing a modeling step (representation, scoring, sampling). A box represents an independent tool and the arrows represent function calls.

representation (`get_state`). Third, the sampling algorithm must be able to also access the score and heuristic information from the scoring function (`get_score`). Finally, the sampling algorithm must be able to update the current model parameters (`update_state`).

Our demonstrative standard is sufficiently general to enable the mixing-and-matching of modeling tools. For example, when computing a model of Nup133, an integrative modeler wishes to score their model against multiple sources of input information (SAXS, XL-MS, and EM) that are not easily handled by a single modeling tool. 3 distinct scoring tools may be drawn from different modeling packages specialized to the computational demands unique to each experimental datatype. Optimally, these scoring function tools operate independently on each information source to manage the complexity of the tool’s implementation. If all tools provide a uniform interface for returning the score and computed heuristics (`get_score`), they may be used interchangeably while remaining isolated from one another.

2.3.5 Illustrative modeling problem

We demonstrate the software standard by integrating IMP and *Phenix* to solve a specific problem in X-ray crystallography. Namely, we are interested in computing a model of a

set of atomic protein structures (multi-state model), based on multiple diffraction datasets collected at different temperatures and physical principles. Such a multi-state model can be useful for mapping the dynamics and allostery of proteins^{26,27}. Next, we describe the three necessary steps in modeling.

Input Information

We are interested in informing our model by both structure factors from the X-ray crystallography experiments in addition to an empirical molecular mechanics force field. Satisfaction of the X-ray crystallography data restrains the overall model geometry while an empirical potential energy function restrains the stereochemistry and nonbonded interactions of local sets of atoms. Utilization of experimental X-ray diffraction datasets in conjunction with a force-field has been applied previously in computing X-ray models^{28,29}.

Representation

The multi-state model M is defined by Cartesian atomic coordinates for each of a small number of discrete structural states of a protein; the model also includes the relative weight of each state.

Bayesian scoring function

In general, Bayes' theorem states that the posterior model density, $p(M|D, I)$, (the conditional probability density of model, M , given experimental measurements, D , and prior information, I) is proportional to the product of the data likelihood, $p(D|M, I)$, (the probability of D given M and I) and prior distribution, $p(M|I)$, (the probability of M given I):

$$p(M|D, I) \propto p(D|M, I) \times p(M|I) \tag{2.1}$$

For our multi-temperature model where we have multiple diffraction datasets, D_i , we

assume that the likelihood is the product of independent likelihoods for each diffraction dataset. The posterior model density is:

$$p(M|D, I) \propto \prod p(D_i|M, I) \times P(M|I) \tag{2.2}$$

where $p(D_i|M, I)$ is the likelihood for the diffraction data from the i -th experiment. In Bayesian modeling, the model is not a single model instance, rather the model is the posterior density over the entire model space spanned by the degrees of freedom defined by the model representation. The uncertainty of the model is the posterior model density spread. Bayesian modeling is conducive for integrative modeling because likelihoods and priors can be combined from diverse experimental datasets and prior information. As is often the case with probabilistic modeling, the score, s , is the negative logarithm of the model posterior density:

$$s = -\log p(M|D, I) \tag{2.3}$$

$$= -\log p(M|I) - \log \prod p(D_i|M, I) \tag{2.4}$$

$$= -\log p(M|I) - \sum \log p(D_i|M, I) \tag{2.5}$$

Though the scores can be weighed as pure probabilities, it is often useful to weigh the likelihood and priors so that the gradients have comparable magnitudes²⁸. The weights, w_1, w_2 , may be optimized to a target function (eg, Phenix.refine) or empirically chosen:

$$s = -w_1 \log p(M|I) - w_2 \sum \log p(D_i|M, I) \tag{2.6}$$

Sampling

A sample can be generated from the model posterior via Molecular Dynamic simulation where atomic positions are updated based on a force computed from the potential energy

surface. The stochastic Monte Carlo method, where parameter moves are accepted and rejected based on relative energy levels, is also useful for generating molecular ensembles. Furthermore, enhanced sampling variations of the Monte Carlo method that leverage derivatives (eg, Hamiltonian Monte Carlo) are useful for improving sample convergence from complex posterior model densities.

To implement the representation, scoring, and sampling outlined above, we integrated software tools from the *Phenix* software suite and IMP modeling package using our demonstrative standard.

2.3.6 *Phenix*

Phenix (Python-based Hierarchical Environment for Integrated Xtallography) provides a suite of programs for manipulating experimental data, computing models, and validating structures from cryo-EM and X-ray/neutron/electron crystallography data. *Phenix* includes tools for the entire data processing and model generation workflow, including computing data quality indicators (phenix.xtrriage), maximum likelihood estimation of phases from molecular replacement via a homologous structure (phenix.phaser³⁰), phase optimization (phenix.density_modification³¹), model building (phenix.autobuild³²), refinement of the model to better fit both experimental and empirical restraints (phenix.refine³³), and finally model validation (access to MolProbity webservice³⁴). Despite a large number of algorithms in *Phenix*, there would be great benefit from integration with IMP, for example by providing access to the flexible model representations, incorporation of non-crystallographic information, and enhanced sampling techniques in IMP.

2.3.7 IMP

Integrative Modeling Platform (IMP) is open-source software that contains a large number of libraries and programs for flexibly computing integrative models of biomolecular systems (<https://integrativemodeling.org>)¹⁵. IMP supports a diverse set of model representa-

tions that can be flexibly coarse-grained to suit the problem. Restraints can be formulated to score models against various experimental data (*eg*, chemical cross-links identified by mass spectrometry, electron microscopy density maps, and small-angle X-ray scattering profiles) as well as prior models (*eg*, excluded volume, comparative models, molecular mechanics force fields, and statistical potentials). Models can be sampled or enumerated through numerical integration techniques, variations of the Monte Carlo method, as well as Molecular and Brownian Dynamics simulations. IMP’s relative strengths include a large variety of model representations, scoring functions based on different data, and sampling schemes, all of which can be mixed and matched relatively easily with each other to facilitate integrative structure modeling. Another distinction is an increasingly Bayesian perspective on uncertainties in input information, model representations, and scoring functions. In contrast, IMP does not include tools for X-ray crystallography. However, *Phenix* is a premier program for this task.

2.3.8 Integration of *Phenix* and IMP (Fig. 4.3)

As IMP and *Phenix* use independent modeling frameworks, we were required to implement the standard functions by using and modifying software tools from both packages. Engineering the interfaces, therefore, required familiarity with the application and implementation of tools from both software ecosystems. For example, IMP and *Phenix* employ their own model hierarchy to represent atomic structures. Yet for the model search, there must be a shared definition of the model representation. Efficiently managing and interconverting between the unique model hierarchies proved challenging. Such challenges of handling specific modeling packages further motivates the need for general standards to facilitate software interoperability without additional development effort.

The multi-state model representation is defined in IMP (IMP.Model). The sampler is also defined in IMP (IMP.MolecularDynamics). We implemented 2 scoring function tools, to evaluate the prior and likelihood respectively. The distinction between data likelihood and prior provides an opportunity for distributing an evaluation of the scoring function across

both *Phenix* and IMP, taking advantage of the comparative strengths of each software package. The prior scoring function uses the IMP.atom library to evaluate stereochemical and non-bonded scores based on the CHARMM22 empirical force field. The likelihood scoring function uses *Phenix*'s maximum likelihood target function for a given set of experimentally observed structure factors. Evaluation of the crystallography likelihood includes several computationally demanding tasks such as the inference of distribution parameters and determination of the solvent region in the unit cell. Both the prior and likelihood scoring function tools were implemented with an identical interface to return the score and gradients, ensuring compatibility with IMP.Model and IMP.MolecularDynamics.

The model search could be implemented in two ways: in separate runtime environments via data integration or in the same runtime environment via library integration, as follows.

Data integration

Our first attempt to integrate *Phenix* and IMP was by exchanging data between independent executions of custom scoring evaluation programs written separately from the *Phenix* (Computational Crystallography Toolbox (CCTBX)) and IMP libraries, respectively. The stochastic sampling algorithm proposes a move and saves the coordinates to a disk file in the PDB format, which is then read separately by the two evaluation programs. The *Phenix* program computes the likelihood while the IMP program computes the prior independently. The posterior is the product of these two terms. The advantage of this strategy is that the runtime environments are completely separated. However, the design presents a runtime challenge because all model data must be saved from the IMP address space to disk and then be read by the IMP and *Phenix* programs. As the structure factor calculations using Fast Fourier Transform (FFT) are extremely fast, the addition of the computational overhead presented a significant challenge to generating and scoring a sufficiently large sample. *Phenix* can compute 6224 structure factors for a ubiquitin molecule in 0.015 seconds on a single computational core. *Phenix* can read and write a PDB file in 0.08 seconds, while IMP

is even slower. As millions of samples may be necessary to sufficiently sample the Bayesian posterior model density, this additional overhead makes the data integration expensive. For the informed search to be computationally feasible, the likelihood and prior evaluations must occur in the same runtime environment, which is achieved via library integration.

Library integration

We also engineered evaluation of both the likelihood and prior with *Phenix* and IMP, respectively, within the same runtime environment. A single evaluation function accepts a model and computes both the likelihood and prior using *Phenix* and IMP library calls, respectively. To integrate the scoring functionalities of *Phenix* and IMP, it was essential to develop a code for translating between the IMP and CCTBX hierarchical representations. As a result, the IMP sampling algorithm can make proposals based on IMP’s implementation of the model representation that is automatically reflected in CCTBX’s implementation representation and can be used natively with other CCTBX tools. Although the release of both IMP and *Phenix*’s underlying libraries (CCTBX) in conda-forge enables a consistent Python environment, we opted to build our integration in C++ for its performance advantages, followed by wrapping in Python for usability. We handled the technical integration of both IMP and *Phenix* shared dynamic (.so) libraries along with their dependencies through a custom compilation of IMP facilitated by CMake.

2.3.9 Conclusions

We were successful in incorporating IMP and *Phenix* functions, data structures, and numerical calculators in a model search. Rather than integrating IMP and *Phenix* tools in an ad hoc fashion, we organized them as proposed by our standard where the scoring function, sampling, and model representation tools communicate through defined channels. Importantly, the crystallographic likelihood and derivative calculator were completely independent of the molecular mechanics force field likelihood and derivative calculator. Using the model

Figure 3

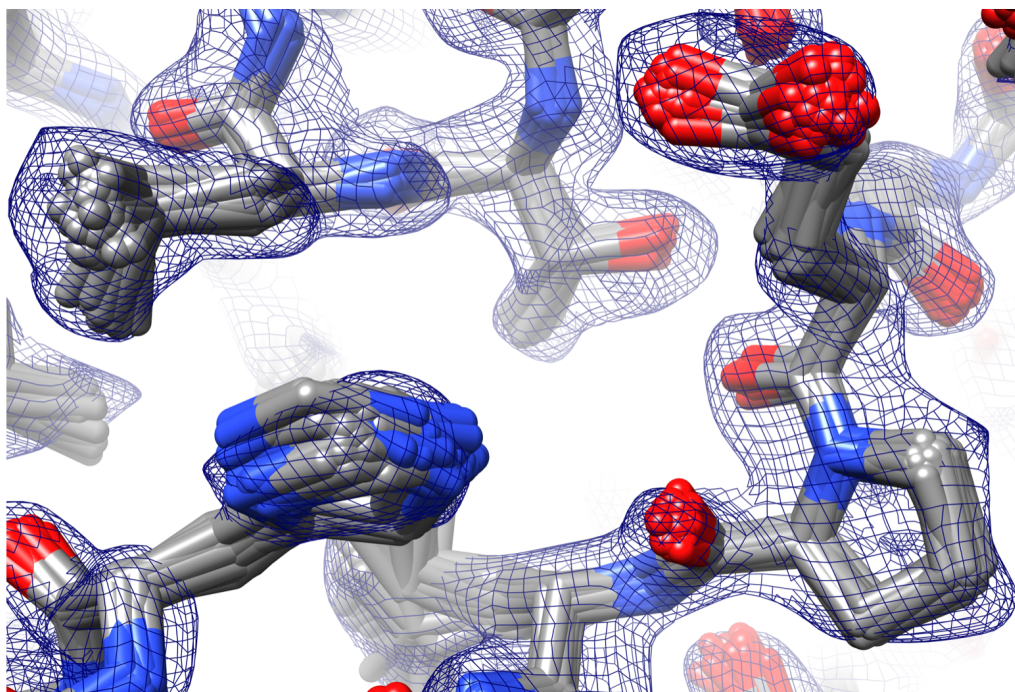


Figure 2.3: Workflow for integrating modeling software, exemplified by *Phenix* and IMP. The sampler is implemented by IMP.MolecularDynamics. The model representation is implemented by IMP.Model. Two scoring tools are used: ForceFieldRestraint for computing the prior based on an empirical force field and XtalRestraint for computing the likelihood based on the observed diffraction data. IMP.MolecularDynamics calls to the model representation evaluate both restraints. The restraint then calls the functions to compute the score and gradients. For ForceFieldRestraint, calls were made to the IMP.atom library to return individual stereochemical and nonbonded scores which are combined to compute in the total prior and gradients. For XtalRestraint, calls were made to *Phenix* functions within the mmtbx and cctbx libraries. IMP.MolecularDynamics updates the model parameters based on Newton's second law of motion where the force is derived from the sum of the returned gradients. The modularity of the design enables the substitution of alternative model representation and sampling tools (IMP.MonteCarlo). The isolation of *Phenix* and IMP scoring evaluations demonstrates how software integration is facilitated by the factorization of the Bayesian posterior model density into a data likelihood and a prior.

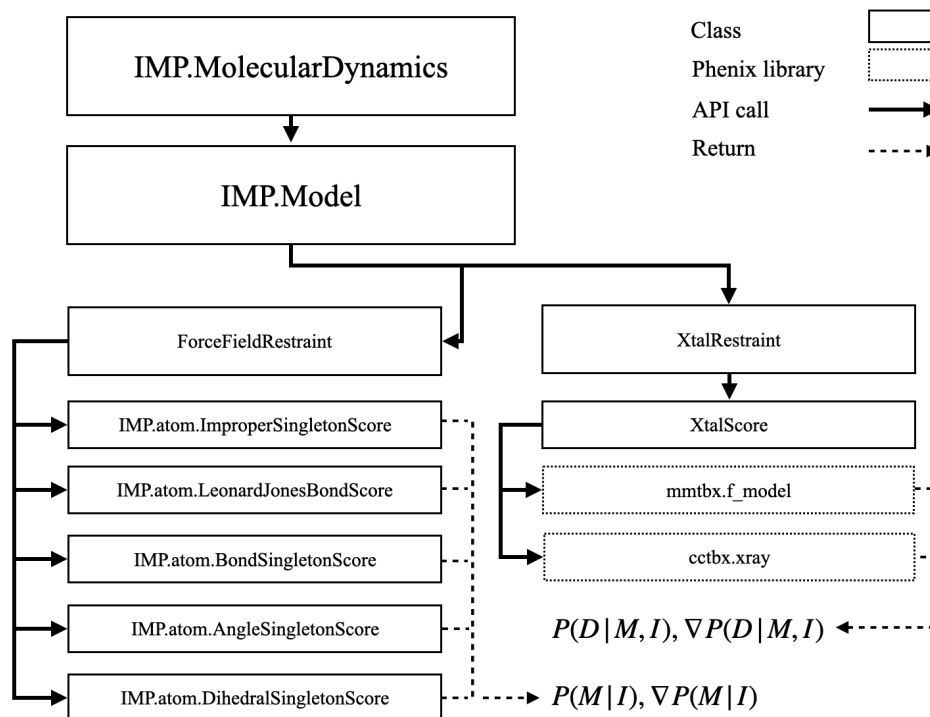
Figure 4

Figure 2.4: Sample of SARS-CoV main protease posterior model density evaluated through integration of IMP and *Phenix*. Sample of Histidine 246, Proline 241, Glutamic Acid 240, and Leucine 202 conformations from the SARS-Cov main protease model posterior density. The sample of 100 structures was generated via Molecular Dynamics sampling of the model posterior density. The posterior was evaluated as the product of the molecular mechanics prior and the X-ray likelihood, computed by IMP and *Phenix* libraries respectively. The model posterior density sample is overlaid by the all features (2Fo-Fc) map.

search implemented through the integration of IMP and *Phenix*, we generated a sample of the model posterior density from the structure of SARS-Cov main protease (PDB ID: 2H2Z) (**Fig. 4.4**). Structures contained within the sample are drawn from the potential energy landscape formulated from the satisfaction of empirical stereochemical and non-bonded relationships as well as the observed X-ray data.

As introduced above, the first major advantage of integration is that little crystallography-specific source code must be implemented in IMP. By leveraging the crystallographic functionality of *Phenix*, we do not have to implement a large number of crystallo-

graphic data structures and subroutines in IMP. CCTBX and IMP consist of 6210 (955,323 lines of code) and 3450 unique source files (327,966 lines of code), respectively, supported by multiple groups around the world. Direct integration of source has historically been the de facto approach to accommodating new data types in IMP. This integration saves significant development time and also prevents inflation of either codebase. We also benefit from the significant amount of previous *Phenix* development and will continue to benefit from future *Phenix* development. IMP also benefits from the *Phenix* authors' expertise in computational crystallography, which includes significant runtime optimizations (e.g., testing whether FFT or direct summation is faster for a given crystal system).

Secondly, the standard is sufficiently general to enable substitution of other model representation, sampling algorithm, or scoring function tools. For example, the modularity is well suited for introducing additional scoring function tools as they simply need to parse the model representation and return the computed score plus heuristics. Using the Bayesian framework, new tools could be easily introduced for computing likelihoods from new forms of experimental data, for example nuclear magnetic resonance (NMR) restraints, or prior information, for example statistical potentials from the PDB.

More complex sampling and scoring procedures may also be employed. The optimal strategy for sampling a model space is often unknown prior to modeling. Methods often rely on iterations of sampling procedures that may vary the sampling parameters (*eg*, temperature in simulated annealing) or the scoring functions (optimization of satisfaction of one source of input information over the other). Based on our framework, the model representation and scoring functions, are encapsulated from the details of the sampling procedure (in other words, *g*). New sampling procedures can easily be substituted so long as they can access the model representation to update the parameter state and accept scores and heuristics from the scoring function.

2.4 Discussion

Next, we discuss the implications of software integration for the field of integrative modeling. In integrative modeling, the goal is to build increasingly complex models based on increasingly varied sets of data⁴. Key to any modeling is input information, which determines the type of model that can be computed; the choice of model representation is of course also informed by the questions asked of a model. In addition to model representation, a general description of modeling requires a Bayesian posterior model density that specifies the probability density of a model, given the input information, and a scheme for sampling this posterior density. Using Bayes' theorem, the posterior can be factorized into data likelihoods, which depend on data, and priors, which depend on prior information. The evaluation of a posterior based on independent likelihoods and priors is simple in concept, but may be technically challenging; in other words, writing the code that implements the component likelihoods and priors may require a significant amount of effort and expertise that is difficult to duplicate by non-experts and wasteful to duplicate by experts. Thus, software integration can be seen as the major method for maximizing the quality and efficiency of modeling based on varied data and prior models. If the experts encoded their expertise in the code that can be easily mixed and matched, integrative modelers would in turn be able to rigorously combine likelihoods and priors for all the available input information to solve their integrative modeling problems efficiently. To illustrate this point in more detail, we discuss three specific examples of the posterior model density factorization next.

The first example is that used above (Approach), corresponding to computing an atomic multi-state model based on crystallography data and physical principles. Posterior model density is factorized into a data likelihood based on X-ray diffraction patterns and a prior based on a molecular mechanics force field. Consequently, input information is conveniently isolated in the software (IMP for prior models and *Phenix* for diffraction patterns) that is best suited to evaluate a model based on it. The separation of software manages the technical complexity of the prior and likelihood implementation.

The second example is computing a coarse-grained structural model of the heteroheptameric Nup84 complex based on a negative-stain electron microscopy map and residue-specific cross-links as well as prior models of the subunits³⁵. Posterior model density is factorized into a data likelihood based on the map, a data likelihood based on the cross-links, a prior based on prior subunit models, and a prior based on excluded volume. Although all the terms were evaluated in IMP in this case, it is conceivable that a more accurate encoding of a subset of input information is or will be available in another software package. In such a case, software integration would facilitate computing a higher quality model more efficiently.

The third example is computing a multi-scale model of glucose-stimulated secretion in human pancreatic beta-cells, based on 8 prior models of different aspects of different parts of the entire system³⁶. These prior models are a coarse-grained spatiotemporal simulation of insulin vesicle trafficking, docking, and exocytosis; a molecular network model of glucose-stimulated insulin secretion signaling; a network model of insulin metabolism; a structural model of glucagon-like peptide-1 receptor activation; a linear model of a pancreatic cell population; and ordinary differential equations for systemic postprandial insulin response. When dealing with a complex multi-scale model, it is often not reasonable to assume independence of the input information. In this case, the prior models must be coupled by additional terms in the scoring function. In addition to the prior models, simple models of statistical coupling between the prior models are also defined. The prior models and the couplers are the priors in a posterior model density for a model of the entire system. Bayesian metamodeling estimates the posterior density via backpropagation. Thus, Bayesian metamodeling decomposes the problem of modeling a large, complex system into smaller, more tractable modeling problems. It is likely that more sophisticated and physically realistic coupling of prior models would be facilitated by software integration, where each type of prior model is evaluated in a separate specialized code developed by experts in the domain of that model.

2.5 Conclusions

In summary, software interoperability would greatly benefit the field of integrative modeling as the integration of software specialized for handling specific types of information supports more efficient building of higher quality and more complex models. We proposed a demonstrative standard that facilitates simple software integration by representing modeling as a model search with defined information passing. We then implemented the above standards to integrate *Phenix* and IMP. Finally, we discussed the ability of the Bayesian formulation to facilitate collaborative integrative modeling by mixing-and-matching priors and likelihoods. Ultimately, we suggest that the software development community in the field of integrative modeling consider the definition and adoption of de facto protocol standards for improving the interoperability of their software.

Acknowledgments

This work was funded by NIH grants R01 GM083960 (Sali), R01 GM123159 (Fraser), P41 GM109824 (Sali), P01 AG002132 (Sali), P50 AI150476 (Sali), and U19 AI135990 (Sali) as well as NSF grants DBI-1756250 (Sali) and DBI-1832184 (Sali). The development of *Phenix* is supported by NIH grants P01GM063210 and R24GM141254 to PDA, the *Phenix* Industrial Consortium, and in part by the US Department of Energy under Contract DE-AC02-05CH11231.

References

1. Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprpto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P. & Sali, A. Determining the architectures of macromolecular assemblies. en. *Nature* **450**, 683–694 (Nov. 2007) (cit. on p. 10).
2. Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. en. *Cell* **177**, 1384–1403 (May 2019) (cit. on pp. 10, 13, 15).
3. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. en. *Nature* **171**, 737–738 (Apr. 1953) (cit. on p. 10).
4. Sali, A. From integrative structural biology to cell biology. en. *J. Biol. Chem.* **296**, 100743 (Jan. 2021) (cit. on pp. 10, 12, 28).
5. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. en. *Nat. Struct. Biol.* **10**, 980 (Dec. 2003) (cit. on p. 11).
6. Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. PHENIX: a comprehensive Python-based system for macromolecular structure solution. en. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (Feb. 2010) (cit. on p. 12).
7. Das, R. & Baker, D. Macromolecular modeling with rosetta. en. *Annu. Rev. Biochem.* **77**, 363–382 (2008) (cit. on p. 12).
8. Dimura, M., Peulen, T. O., Hanke, C. A., Prakash, A., Gohlke, H. & Seidel, C. A. Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. en. *Curr. Opin. Struct. Biol.* **40**, 163–185 (Oct. 2016) (cit. on p. 12).

9. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. en. *J. Am. Chem. Soc.* **125**, 1731–1737 (Feb. 2003) (cit. on p. 12).
10. Hsieh, A., Lu, L., Chance, M. R. & Yang, S. A Practical Guide to iSPOT Modeling: An Integrative Structural Biology Platform. en. *Adv. Exp. Med. Biol.* **1009**, 229–238 (2017) (cit. on p. 12).
11. Hua, N., Tjong, H., Shin, H., Gong, K., Zhou, X. J. & Alber, F. Producing genome structure populations with the dynamic and automated PGS software. en. *Nat. Protoc.* **13**, 915–926 (May 2018) (cit. on p. 12).
12. Hummer, G. & Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. en. *J. Chem. Phys.* **143**, 243150 (Dec. 2015) (cit. on p. 12).
13. Karakaş, M., Woetzel, N., Staritzbichler, R., Alexander, N., Weiner, B. E. & Meiler, J. BCL::Fold—de novo prediction of complex and large protein topologies by assembly of secondary structure elements. en. *PLoS One* **7**, e49240 (Nov. 2012) (cit. on p. 12).
14. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K. W., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D. & Bradley, P. en. in *Methods in Enzymology* (eds Johnson, M. L. & Brand, L.) 545–574 (Academic Press, Jan. 2011) (cit. on pp. 12, 13).
15. Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B. & Sali, A. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. en. *PLoS Biol.* **10**, e1001244 (Jan. 2012) (cit. on pp. 12, 21).

16. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. en. *Nucleic Acids Res.* **33**, W363–7 (July 2005) (cit. on p. 12).
17. Schwieters, C. D., Bermejo, G. A. & Clore, G. M. Xplor-NIH for molecular structure determination from NMR and other data sources. en. *Protein Sci.* **27**, 26–40 (Jan. 2018) (cit. on p. 12).
18. Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G. J. & Marti-Renom, M. A. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. en. *PLoS Comput. Biol.* **13**, e1005665 (July 2017) (cit. on p. 12).
19. Trussart, M., Serra, F., Baù, D., Junier, I., Serrano, L. & Marti-Renom, M. A. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. en. *Nucleic Acids Res.* **43**, 3465–3477 (Apr. 2015) (cit. on p. 12).
20. Van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S. J., van Dijk, M., de Vries, S. J. & Bonvin, A. M. J. J. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. en. *J. Mol. Biol.* **428**, 720–725 (Feb. 2016) (cit. on p. 12).
21. Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E.-M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., Meiler, J. & Baker, D. RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. en. *PLoS One* **6**, e20161 (June 2011) (cit. on p. 13).
22. DiMaio, F., Echols, N., Headd, J. J., Terwilliger, T. C., Adams, P. D. & Baker, D. Improved low-resolution crystallographic refinement with Phenix and Rosetta. en. *Nat. Methods* **10**, 1102–1104 (Nov. 2013) (cit. on p. 13).
23. Kim, S. J., Fernandez-Martinez, J., Nudelman, I., Shi, Y., Zhang, W., Raveh, B., Herricks, T., Slaughter, B. D., Hogan, J. A., Upla, P., Chemmama, I. E., Pellarin, R.,

- Echeverria, I., Shivaraju, M., Chaudhury, A. S., Wang, J., Williams, R., Unruh, J. R., Greenberg, C. H., Jacobs, E. Y., Yu, Z., de la Cruz, M. J., Mironska, R., Stokes, D. L., Aitchison, J. D., Jarrold, M. F., Gerton, J. L., Ludtke, S. J., Akey, C. W., Chait, B. T., Sali, A. & Rout, M. P. Integrative structure and functional anatomy of a nuclear pore complex. en. *Nature* **555**, 475–482 (Mar. 2018) (cit. on p. 15).
24. Grosan, C. & Abraham, A. *Intelligent Systems: A Modern Approach* en (Springer Berlin Heidelberg, July 2011) (cit. on p. 16).
25. Köfinger, J., Różycki, B. & Hummer, G. in *Biomolecular Simulations: Methods and Protocols* (eds Bonomi, M. & Camilloni, C.) 341–352 (Springer New York, New York, NY, 2019) (cit. on p. 16).
26. Fraser, J. S., van den Bedem, H., Samelson, A. J., Lang, P. T., Holton, J. M., Echols, N. & Alber, T. Accessing protein conformational ensembles using room-temperature X-ray crystallography. en. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16247–16252 (Sept. 2011) (cit. on p. 19).
27. Keedy, D. A., Kenner, L. R., Warkentin, M., Woldeyes, R. A., Hopkins, J. B., Thompson, M. C., Brewster, A. S., Van Benschoten, A. H., Baxter, E. L., Uervirojnangkoorn, M., McPhillips, S. E., Song, J., Alonso-Mori, R., Holton, J. M., Weis, W. I., Brunger, A. T., Soltis, S. M., Lemke, H., Gonzalez, A., Sauter, N. K., Cohen, A. E., van den Bedem, H., Thorne, R. E. & Fraser, J. S. Mapping the conformational landscape of a dynamic enzyme by multitemperature and XFEL crystallography. en. *Elife* **4** (Sept. 2015) (cit. on p. 19).
28. Brünger, A. T., Karplus, M. & Petsko, G. A. Crystallographic refinement by simulated annealing: application to crambin. en. *Acta Crystallogr. A* **45**, 50–61 (Jan. 1989) (cit. on pp. 19, 20).
29. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. en. *Elife* **1**, e00311 (Dec. 2012) (cit. on p. 19).

30. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. Phaser crystallographic software. en. *J. Appl. Crystallogr.* **40**, 658–674 (Aug. 2007) (cit. on p. 21).
31. Terwilliger, T. C., Ludtke, S. J., Read, R. J., Adams, P. D. & Afonine, P. V. *Improvement of cryo-EM maps by density modification* en. Mar. 2020 (cit. on p. 21).
32. Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L. W., Read, R. J. & Adams, P. D. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. en. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (Jan. 2008) (cit. on p. 21).
33. Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. Towards automated crystallographic structure refinement with phenix.refine. en. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (Apr. 2012) (cit. on p. 21).
34. Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis, S. M., Arendall 3rd, W. B., Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S. & Richardson, D. C. MolProbity: More and better reference data for improved all-atom structure validation. en. *Protein Sci.* **27**, 293–315 (Jan. 2018) (cit. on p. 21).
35. Shi, Y., Fernandez-Martinez, J., Tjioe, E., Pellarin, R., Kim, S. J., Williams, R., Schneidman-Duhovny, D., Sali, A., Rout, M. P. & Chait, B. T. Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. en. *Mol. Cell. Proteomics* **13**, 2927–2943 (Nov. 2014) (cit. on p. 29).
36. Raveh, B., Sun, L., White, K. L., Sanyal, T., Tempkin, J., Zheng, D., Bharat, K., Singla, J., Wang, C., Zhao, J., Li, A., Graham, N. A., Kesselman, C., Stevens, R. C. & Sali, A. *Bayesian metamodeling of complex biological systems across varying representations*

en. Mar. 2021 (cit. on p. 29).

Chapter 3

A Long-Range Recruitment

Mechanism for mTORC2

Phosphorylation of Akt

3.1 Abstract

In response to growth factors like insulin, the mTOR Complex 2 (mTORC2) phosphorylates and activates Akt, PKC, and SGK protein kinases, which in turn control diverse cellular processes¹⁻³. How mTORC2, a large protein complex containing mTOR, Rictor, mSin1, and mLST8, recognizes these kinases to phosphorylate them on specific sites is uncertain³⁻⁶. To address these questions, we studied the phosphorylation by mTORC2 of Ser473 of Akt, a central regulator of growth and metabolism⁷. Enzymatic and cellular data revealed that mTORC2 directly phosphorylates Akt Ser473, and that the local sequence makes limited contributions to recognition. To understand why, we prepared bivalent semisynthetic Akt proteins to stabilize the interactions with mTORC2, enabling cross-linking mass spectrometry and cryo-electron microscopy (cryo-EM) analysis of the mTORC2-Akt co-complex structure. We find that mTORC2 phosphorylation of Akt is driven by recognition of the three-

dimensional fold of Akt at two interfaces: (1) mSin1 CRIM domain with the Akt N-lobe, and (2) mSin1/Rictor N-terminal regions with the Akt C-lobe, and is augmented by PIP3-induced membrane association via mTORC2 and Akt PH domains. The surface on the Akt N-lobe that binds to mSin1 CRIM overlaps with that previously proposed to mediate the intramolecular interaction between Akt phospho-Thr450 and its N-lobe, suggesting an intricate interplay between mTORC2-mediated phosphorylation of Thr450 and Ser473. We show that this mechanism of molecular recognition by mTORC2 is conserved across a subset of AGC kinases. Together, these results reveal the long-range interactions that enable mTORC2 to specifically phosphorylate and activate Akt and other key cellular signaling substrates.

3.2 Introduction

The mTOR complexes mTORC1 and mTORC2 are critical for diverse cellular functions, including the control of growth, metabolism, and proliferation, and are conserved from yeast to human^{2,8}. Extracellular signals including insulin and other growth factors activate mTORC2 at the plasma membrane⁹. mTORC2 phosphorylates and activates AGC family kinases including Akt (protein kinase B)^{1,10-12}, protein kinase C (PKC)¹³, and SGK¹⁴ at their hydrophobic motifs (Ser473 in Akt) in the C-terminal tail (C-tail, residues 422-480 in Akt) (**Fig. 4.1a**). Dysregulated mTOR signaling is central to the pathology of aging and diseases including diabetes/metabolic syndrome and many human cancers⁸. Because the two mTOR complexes have distinct and often opposing signaling roles and cellular feedback loops, selective modulation of both complexes is of interest in these diseases⁸. While both mTORC1 and mTORC2 are sensitive to ATP-site small molecule inhibitors, inhibition of both complexes by existing agents limits their clinical utility¹⁵, and efforts to develop complex-specific pharmacology have been stymied because both complexes share the kinase mTOR and component mLST8 in nearly identical conformations^{4,16}.

Figure 1

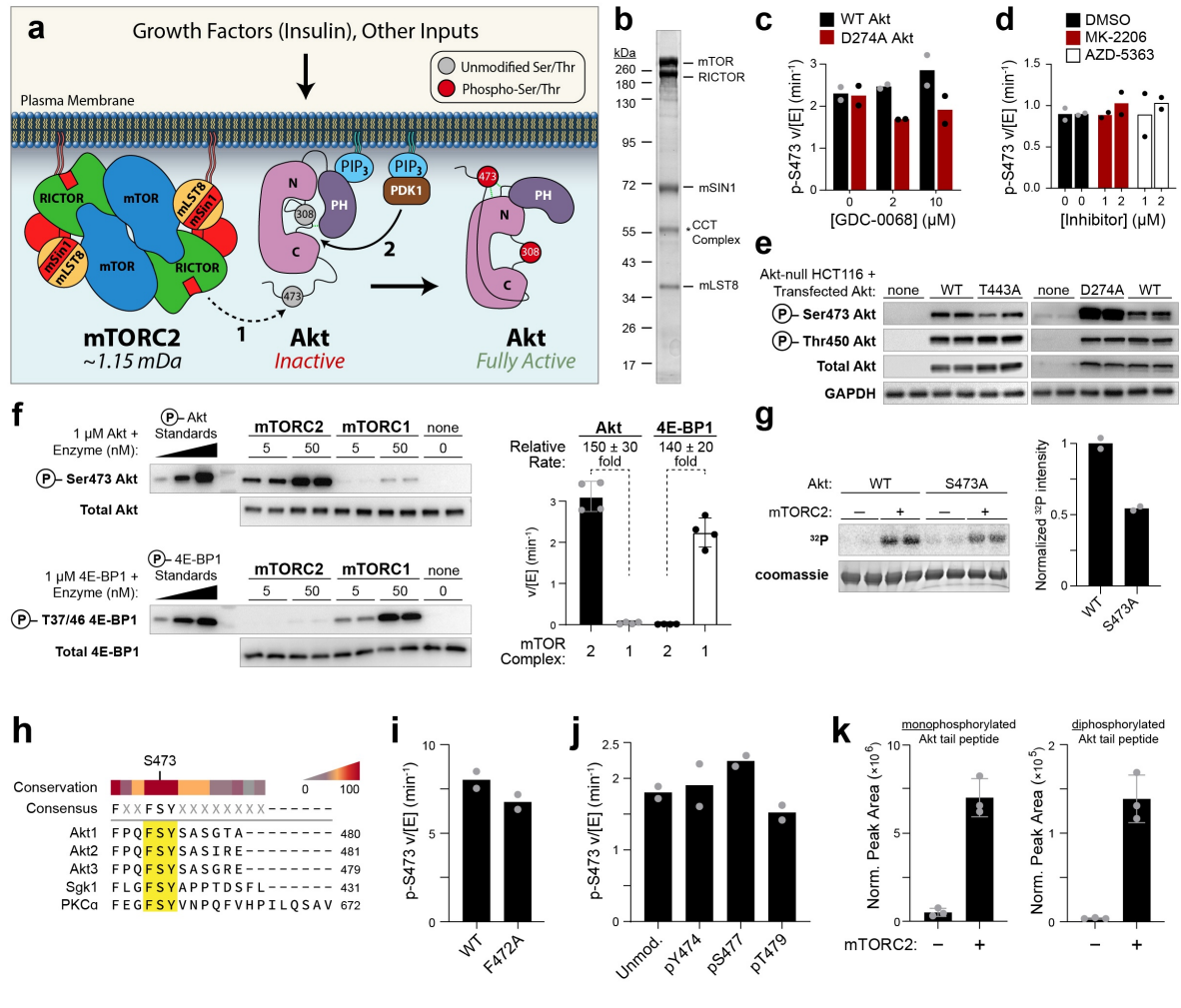


Figure 3.1: mTORC2 directly and specifically phosphorylates Ser473 of Akt with little contribution from local Akt sequence. (Figure caption continued on next page.)

(Figure caption continued from previous page.) **a**, Model for two-step activation of Akt by mTORC2 at the plasma membrane. Following PIP3-mediated recruitment, Akt is sequentially modified at Ser473 by mTORC2, facilitating a conformational change of the PH domain, and then at Thr308 by PDK1, together resulting in full activity. **b**, Purified monodisperse mTORC2 expressed in SF9 insect cells. **c,d** No significant effect of Akt inhibitors and Akt catalytic activity of mTORC2 phosphorylation of Akt Ser473 *in vitro*; inhibitors used have reported Akt IC50 <10 nM; $\frac{v}{[E]} = \frac{\text{velocity}}{[\text{enzyme (mTORC2)}]}$; 5 nM mTORC2, 1 mM ATP, 1 μ M Akt, 10 mM MgCl2. **c**, Neither ATP-competitive Akt inhibitor GDC-0068 nor Akt catalytic status affects mTORC2 phosphorylation of Akt (D274A Akt: catalytic null). **d**, Allosteric Akt inhibitor MK-2206 and ATP-competitive Akt inhibitor AZD-5360 do not affect Akt phosphorylation by mTORC2 at 1-2 μ M. **e**, Akt Ser473 phosphorylation assay in Akt1/2-null HCT116 cells; transfected cells were serum starved for 20 hours, stimulated for 40 min with insulin and hIGF4, and analyzed by immunoblotting. **f**, mTORC2 and mTORC1 are specific in phosphorylating canonical substrates Akt and 4E-BP1, respectively under identical conditions *in vitro* as read by quantitative western blotting; 1 μ M substrate, 1 mM ATP, and 10 mM MgCl2; n=4, representative of 2 experiments (p<0.0001, both rate comparisons, t-test). **g**, Around half of mTORC2-Akt phosphorylation is lost with S473A mutation as measured by γ^{32} P-ATP assay; 1 μ Ci γ^{32} P-ATP was used at 1 mM final, 5 μ M Akt, 5 nM mTORC2, 10 mM MgCl2; notably WT Akt and AktS473A preparations have minimal pT30812. **h-j**, Local sequence around Ser473 has little impact on Akt Ser473 phosphorylation despite conservation. **h**, C-tail sequences of human AGC kinases shows conservation of Ser473-flanking Phe472 and Tyr474. **i,j**, Neither mutation of Phe472 to Ala nor phosphorylation of Tyr474, Ser477, or Thr479 has a significant effect on Ser473 phosphorylation *in vitro* (quantitative western blotting; i, 5 μ M Akt, j, 1 μ M Akt). **k**, Mass spectrometry measurement of *in vitro* phosphorylated Akt tails shows monophosphorylated tail peptide is largely pSer473, and mTORC2 produces doubly phosphorylated pSer473-pSer475 tails.

Recent structural analysis has revealed how two distinct motifs, TOS¹⁷⁻¹⁹ and RAIP²⁰⁻²² control phosphorylation of mTORC1 substrates including 4E-BP1. In contrast to mTORC1, the enzymological properties and structural interactions of mTORC2 with its substrates have not been characterized in depth^{4,5,23,24}. Biochemical and cellular studies have suggested potential roles for the CRIM²⁵ and pleckstrin homology (PH) domains²⁶ of mSin1 in Akt C-tail modification, and the role of the atypical Ras binding domain (aRBD) is controversial^{27,28}. However, the relative importance of proximal amino acid residues around Akt Ser473 versus distal docking sites for mTORC2 recognition remain unclear. Moreover, a recent report proposed a mechanism in which mTORC2 may not even directly phosphorylate Akt, SGK, and PKC at the C-tail⁶. Instead, Akt was proposed to autophosphorylate Ser473 follow-

ing mTORC2 phosphorylation of an upstream site in the ‘TIM’ motif (Thr443 in Akt), with analogous mechanisms proposed in SGK and PKC⁶. mTORC2 has also been shown to have a role in turn motif phosphorylation of Akt (Thr450), PKC^{13,29}, and PKN³⁰, and these phospho-sites are proposed to engage in intramolecular N-lobe interactions in these kinases that promote their stability^{13,29}. Currently, there are no experimental structures of mTORC2 bound to Akt, SGK, or other substrates^{4,5}. Although there are increasing numbers of high-resolution structures of protein kinases bound to synthetic peptide substrate analogs, there have been very few of protein kinases bound to their folded protein substrates. Beyond kinases that autophosphorylate, we are aware of only a few such examples including: Csk bound to Src³¹, BRAF complexed with MEK³², and p38 engaged with MK2³³. A principal challenge in capturing protein kinase/protein substrate structures is the modest affinity of these interactions, which likely facilitates efficient catalytic turnover.

Akt1 (referred to here as Akt) and its paralogs Akt2 and Akt3 are Ser/Thr protein kinases that govern cell proliferation and metabolism and play key roles in health and diseases including cancer and diabetes^{7,34}. Growth factor signaling through PI3-kinase first recruits Akt to the plasma membrane via elevation of the phospholipid phosphatidyl-3,4,5-triphosphate (PIP3)^{7,34}. Two protein kinases, mTORC2 and PDK1, then sequentially activate Akt (**Fig. 4.1a**): mTORC2 targets Akt at Ser473 on the C-tail hydrophobic motif, relieving autoinhibition and facilitating subsequent PDK1 phosphorylation of Thr308 in Akt’s activation loop^{1,10–12,35–37}. Turn motif phosphorylation at Akt Thr450 requires mTORC2 and is thought to be constitutive, but it may be regulated in other AGC members^{13,29,38}, such as PKN¹⁹. The combination of activation loop and C-tail phosphorylation results in full stimulation of Akt, which then phosphorylates key substrates such as GSK3, PRAS40, and the FOXO transcription factors^{7,34}. Whereas the molecular basis for PDK1 recognition and activation loop phosphorylation of the AGC family of kinases has been partially illuminated³⁹, much less is known about how mTORC2 targets the C-tails of Akt and those of related AGC kinases including SGK and PKC. To shed light on this mechanism, we began

by establishing a system to study mTORC2 phosphorylation of Akt Ser473 *in vitro*.

3.3 Results

3.3.1 mTORC2 directly phosphorylates Akt at Ser473

Understanding of mTORC2-Akt interactions have been limited by the challenges of producing mTORC2 as well as Akt with controlled post-translational modifications. To generate homogeneous Akt substrates and standards, we employed expressed protein ligation⁴⁰; the resultant semisynthetic Akt proteins lack Thr308 phosphorylation and contain either a uniformly unmodified or a uniformly Ser473 mono-phosphorylated synthetic C-tail¹² (**Fig. 3.7a**). We overexpressed and purified human mTORC2 from both HEK293T and insect cells (**Fig. 4.1b**, **Fig. 3.7b-d**) and measured mTORC2 catalytic parameters by quantitative western blots using semisynthetic Akt substrates and standards. These studies revealed apparent K_M values with insect cell-produced mTORC2 for Akt and ATP of 6 μM and 1 mM, respectively (**Fig. 3.7e**). Comparable results were seen with HEK293T cell-produced mTORC2, although a lower apparent ATP K_M was measured (apparent K_M of 0.27 mM, **Fig. 3.8a-c**). Due to low yield of monodisperse mTORC2 from HEK293T cells, further analyses were performed with insect cell expressed mTORC2.

To distinguish between mTORC2 catalyzed phosphorylation of Ser473 and Akt autophosphorylation, we tested the catalytically inactive D274A mutant Akt as an mTORC2 substrate and several Akt inhibitors including the ATP-competitive inhibitors GDC-0068 (ipatasertib)⁴¹ and FDA-approved AZD5363 (capiasertib)⁴² and the Akt allosteric inhibitor MK2206⁴³. The rate of mTORC2 phosphorylation of AktD274A on Ser473 was similar to that of wild-type (WT) Akt and was minimally affected by Akt inhibitors (**Fig. 1c-d**, **Fig. 3.8d-f**). These results show that the catalytic activity of Akt is not essential for its C-tail phosphorylation by mTORC2. In contrast, the mTOR inhibitors Torin1 and Torin2 potently inhibit Akt Ser473 phosphorylation (**Fig. 3.7e**, **Fig. 3.8c**), evidence that mTORC2

directly catalyzes Ser473 phosphorylation. We also investigated the role of Thr443 and Akt catalytic activity on C-tail phosphorylation in mammalian cells. We expressed WT and mutant Akt forms in HCT116 Akt1/Akt2 knockout cells⁴⁴ and found that T443A mutation resulted in little impact on Ser473 phosphorylation whereas catalytically inactive AktD274A showed enhanced Ser473 phosphorylation, perhaps due to a cellular feedback mechanism^{7,8} (**Fig. 4.1e**, **Fig. 3.8f**). Taken together, the biochemical and cellular experiments indicate that mTORC2 directly phosphorylates Akt Ser473.

3.3.2 Specificity Determinants of mTORC2 Catalyzed Phosphorylation of Akt

To understand the specificity of mTORC2 in C-tail phosphorylation of Akt, we compared the activity of purified mTORC1 versus mTORC2. Both mTOR complexes showed high specificity for canonical substrates *in vitro*, with mTORC2 ~150-fold faster in phosphorylating Akt and mTORC1 ~140-fold faster towards 4E-BP1 (**Fig. 4.1f**, **Fig. 3.9a,b**). We also looked at the potential role of the PH domain of Akt in C-tail phosphorylation and found that mTORC2 effectively phosphorylates Akt in solution, with or without the Akt PH domain: the apparent k_{cat}/K_M was within ~2-fold for WT vs Δ PH-Akt (amino acids 123-480), although the apparent K_M of PH-Akt was elevated (>20 M) relative to full length Akt (**Fig. 3.9c**). The Akt PH domain is therefore not essential for C-terminal phosphorylation of Akt by mTORC2 in solution.

We next tested how the residues flanking Ser473 in Akt, namely Phe472 and Tyr474, affect mTORC2 catalysis. Phe472 and Tyr474 are conserved across Akt1-3 as well as SGK1 and PKC α (**Fig. 4.1h**). We chose to test F472A and pTyr474 Akt forms⁴⁵. To enable accurate western blot calibration, it was necessary to prepare compound-modified AktF472A-pSer473 and AktpSer473-pTyr474 by expressed protein ligation. Unexpectedly, mTORC2 phosphorylation of AktF472A and AktpTyr474 occurs with rates nearly identical to WT Akt (**Fig. 4.1i**, **Fig. 3.9d-e**). These results suggest that, despite their conservation, the specific

residues neighboring Ser473 are relatively unimportant in mTORC2-mediated phosphorylation of the Akt C-tail. However, these residues are likely important for pSer473 effects on Akt kinase activity⁴⁵ and may influence susceptibility of pSer473 to dephosphorylation. Similar results were obtained with pSer477 or pSer479 Akt, two previously identified neighboring phospho-sites on Akt⁴⁶, which showed little change in mTORC2 phosphorylation of Ser473 despite the addition of bulky phosphate groups in proximity (**Fig. 4.1j**, **Fig. 3.9e**).

We also probed mTORC2 site-selectivity using a radioactivity-based assay with $\gamma^{32}\text{P}$ -ATP with WT and S473A mutant Akt as substrates. Analysis of $\gamma^{32}\text{P}$ -incorporation into WT Akt vs AktS473A indicated that Ser473 accounts for about 50% of the overall phosphorylation of Akt, indicating a fair degree of promiscuity by mTORC2 (**Fig. 4.1g**). Interestingly, AktS473A showed little inhibition of mTORC2-catalyzed phosphorylation of WT Akt even at 10 μM ΔPH -Akt competitor, suggesting that the Kd of Akt is relatively weak and greater than its apparent KM (**Fig. 3.9g**). We next assayed the specificity of Akt C-tail phosphorylation using nanoLC-MS/MS and found that monophosphorylation of Ser473 of purified Akt is the major phosphorylation site by purified mTORC2, with lesser amounts of diphosphorylation of Ser473+Ser475 (**Fig. 4.1k**, **Fig. 3.24**, **Fig. 3.25**). Together, these data show that local substrate sequence has little effect on mTORC2 activity and suggest that other mechanisms likely mediate mTORC2 recognition of Akt.

3.3.3 Akt-ATP as a bisubstrate inhibitor of mTORC2

Our kinetic experiments reveal a low baseline affinity of Akt for mTORC2, which has likely impeded the acquisition of a stable co-complex suitable for structural analysis^{4,5}. Appropriately designed ATP-peptide bisubstrate analogs typically show enhanced affinity for their cognate kinases relative to the unmodified peptides and have been used as structural tools to analyze substrate molecular recognition for several protein kinases^{12,47-49}. We employed a related approach to prepare an Akt-ATP protein bisubstrate analog using semisynthesis with the linkage through an aminoAla-acetyl-ATP S at Ser473 of Akt (**Fig. 3.10a**). While

producing the Akt-ATP protein bisubstrate analog, we generated the Akt-ATP peptide conjugate (aa460-480) as a useful comparator. We analyzed the affinities of Akt-ATP peptide and protein by testing these agents as inhibitors of mTORC2 phosphorylation of unmodified Akt. These experiments showed that Akt-ATP peptide possessed an IC₅₀ of 180 μ M whereas Akt-ATP protein displayed an IC₅₀ of 1.5 μ M (**Fig. 3.10b**). The ~100-fold difference in IC₅₀ values of the peptide versus protein reveals the importance of the full-length Akt versus the truncated C-tail in molecular recognition of substrates by mTORC2.

Akt-ATP showed sufficient affinity for mTORC2 to partly co-elute with mTORC2 using size-exclusion chromatography (**Fig. 3.10c**) but cryo-EM analysis of this complex did not reveal a strong density for Akt (data not shown). Unable to extract additional structural details from microscopy, we employed crosslinking mass spectrometry (XL-MS) on the reconstituted mTORC2:Akt-ATP co-complex. Using complementary cross-linkers including DSS (disuccinimidyl suberate) or EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride) we aimed to reveal potential binding or contact interfaces among the components. This revealed several Akt-mSin1 cross-links including (1) Akt Lys183-mSin1-Lys295 and (2) Akt Lys183-mSin1-Lys313, and (3) Akt Lys154-mSin1 Lys175, as well as numerous other cross-links (**Fig. 3.10d**, discussed below). Lysines 154 and 183 of Akt are in the kinase N-lobe while mSin1 Lys175 is in the CRIM domain, and Lys295 and Lys313 of mSin1 are C-terminal to CRIM in the adjacent aRBD²⁷.

Because Lys154 and Lys183 are conserved among mTORC2 substrates Akt1-3, SGK1, and PKC α and solvent exposed (**Fig. 3.10e,f**), we examined the importance of the Akt surfaces containing these residues in mTORC2 recognition. Transient expression experiments in HCT116 Akt1/2 knockout cells revealed that trialanine mutation of residues 153-155 had a modest effect on pSer473 levels, and mutation of 182-184 nearly abrogated it (**Fig. 3.10g**). Among alanine scanning of 182-184 residues, K182A and E184A show moderate reduction in Akt pSer473 levels, whereas K183A almost completely abolished pSer473 (**Fig. 4.2a, Fig. 3.10h**). pThr450 levels of mutant Akt forms are comparable or slightly reduced relative

to WT. pThr450 levels are often used as a marker of proper folding and stability of Akt45. As discussed later in this study, we show that pThr450 is directly formed by mTORC2, making it more complex to interpret as such a biomarker.

Figure 2

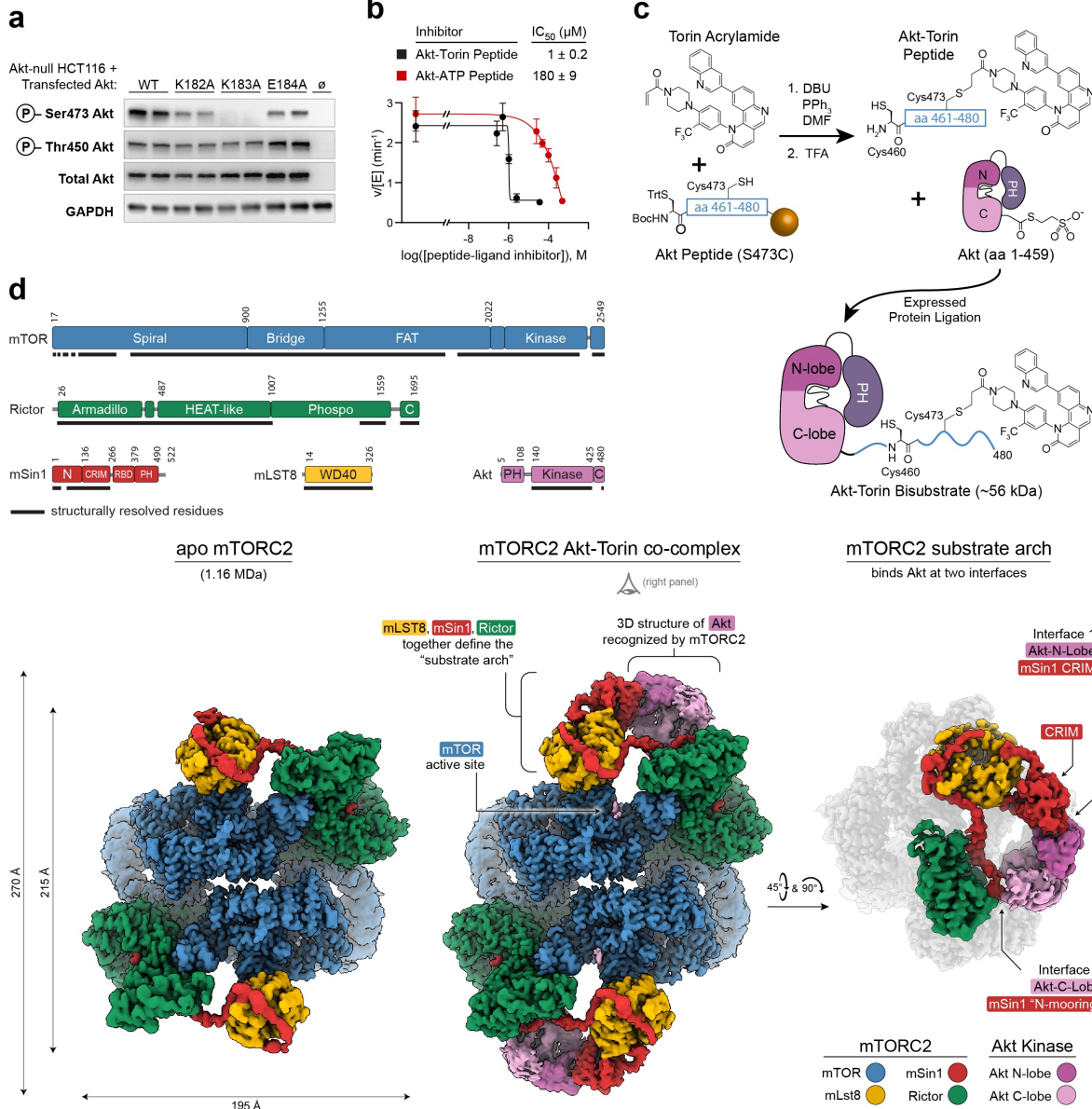


Figure 3.2: mTORC2-Akt Complex Structure Determined Using Akt-Torin. (Figure caption continued on next page.)

(Figure caption continued from previous page.) **a**, First-generation bivalent mTORC2 ligand Akt-ATP was insufficient for structure determination, but cross-linking mass spectrometry (Extended Data Fig. 2.1) revealed a contact between Akt K183 and mSin1 CRIM. Mutation of K183 and neighboring residues was tested in Akt phosphorylation in transfected HCT116 cells; \emptyset , untransfected cells. **b**, Peptide inhibition of mTORC2 phosphorylation of Akt *in vitro* by Akt-Torin peptide is 180-fold more potent than Akt-ATP peptide; 5 μ M Akt, 1 mM ATP, 5 nM mTORC2, 10 mM MgCl₂ (n=4). **c**, Design of second-generation bivalent mTORC2 ligand Akt-Torin. Torin acrylamide is conjugated to Cys473 of an Akt tail peptide via a Michael addition, and purified Akt-Torin peptide (460-480) is ligated to Akt(1-459) thioester; the expressed protein ligation reaction produces a native peptide backbone, facilitated by the presence of Cys460 in Akt. **d**, Structural determination of apo-mTORC2 and mTORC2-Akt-Torin by cryo-EM, resolved to 3.12 Å and 2.55 Å, respectively. Diagram shows the components of the complex, their compositions, and structurally resolved regions. Akt binds to mTORC2 in the “substrate arch”, where mLST8, mSin1, and Rictor together form two distinct docking interfaces \sim 75 Å from the mTOR active site. First, the Akt kinase domain N-lobe binds mSin1 CRIM, which is tethered and stabilized at two points against mLST8. Second, the Akt kinase domain C-lobe binds to the N-terminal region of mSin1, a flexible interaction we name the “mSin1 N-mooring”. The critical component for mTORC2 substrate binding is mSin1, which takes a long and circuitous path, snaking from a pocket in Rictor against which it builds the N-mooring, across the substrate arch, around the back of mLST8, and then back to build the CRIM interface. Subsequent mSin1 domains aRBD (atypical Ras binding domain) and PH are involved in regulation and tether mTORC2 to the plasma membrane, respectively.

To further explore the role of Lys183, which was crosslinked to mSin1, we pursued an *in vitro* approach and generated semisynthetic single mutant K183A and triple mutant KKE182-184AAA Akt purified proteins and examined them as mTORC2 substrates in kinase assays. These assays showed that mTORC2-catalyzed phosphorylation of Ser473 was \sim 50% slower for AktK183A relative to WT Akt, whereas AktKKE182-184AAA displayed \sim 10-fold reduction (**Fig. 3.11a**). Taken together, these results demonstrate that aa182-184 of Akt are influential determinants of mTORC2-catalyzed phosphorylation of Akt Ser473. This is especially notable, given both (1) the large spatial separation between Akt N-lobe and C-tail and (2) the apparent lack of importance of residues neighboring Ser473 in this reaction.

To further explore the potential mode of interaction between mTORC2 and Akt, we modeled the interaction based on cross-linking data⁵⁰ using our open-source Integrative Modeling

Platform (IMP) software⁵¹. The results together satisfy 192/201 cross-links (95.5%) and show two predominant binding configurations of Akt and mTORC2 components; (1) docked in the substrate arch as identified in our preliminary cryo-EM map, 75 Å away from the mTOR active site, and (2) at remote locations surrounding the recently-identified ATP binding site on Rictor (the “A-site”)⁴ (**Fig. 3.11b-c**). We rationalize this second binding configuration as an artefact of the ATP in our bisubstrate, given that added ATP γ S was seen in this site in prior studies⁴.

3.3.4 Akt-Torin as a bivalent ligand for mTORC2

To generate a more stable mTORC2:Akt co-complex, we reasoned that replacing ATP in the bisubstrate with an ATP-competitive inhibitor of mTOR, such as an analog of Torin1⁵², could lead to a better bivalent ligand with both higher affinity and specificity for the mTORC2 active site, avoiding binding to the Rictor “A-site”. Inspection of an X-ray structure of mTOR Δ N bound to Torin2 suggested that the propionyl-piperazine moiety of Torin1 should be oriented away from the catalytic pocket and toward an approaching substrate Ser hydroxy (PDB 4JSX, **Fig. 3.12a,b**)⁵³. Moreover, structure-activity-relationship analysis of Torin1 suggested that substitutions of its propionyl group could preserve high affinity for mTORC2⁵⁴. We thus prepared Torin-acrylamide and conjugated it to the Akt C-tail peptide (aa460-480) via Michael addition, enabled by cysteine substitution at position 473 (S473C). The resulting Akt-Torin peptide displayed an IC₅₀ of 1 nM in blocking mTORC2 phosphorylation of Akt, greater than 100-fold more potent than Akt-ATP peptide (**Fig. 4.2b-c, Fig. 3.12b-d**). We next employed Akt-Torin peptide in expressed protein ligation to generate Akt-Torin protein. Although the ligation proceeded with >50% conversion, the resulting monodisperse fraction contained <20% Akt-Torin (1-480) and >80% unligated (C-tail deleted) Akt (aa1-459), likely due to limited solubility conferred by attachment of the Torin analog. Despite the presence of mixed protein species, mTORC2 selectively bound to the fully conjugated Akt-Torin and co-migrated in size-exclusion chromatography, which yielded

stoichiometric mTORC2:Akt-Torin co-complex suitable for structural analysis (**Fig. 3.12e**).

3.3.5 Cryo-EM analysis of mTORC2:Akt-Torin co-complex

We performed single-particle cryo-EM of apo-mTORC2 and the mTORC2:Akt-Torin co-complex and determined their structures to overall resolutions of 3.1 Å and 2.5 Å, respectively (**Fig. 4.2d, Fig. 3.26, Fig. 3.27, Fig. 3.28, Fig. 3.29, Fig. 3.30**). This facilitated building full atomic models of both complexes, including numerous regions of the complex not previously resolved such as the mSin1 CRIM domain and large sections of the mTOR spiral. The resulting cryo-EM map with bound Akt-Torin substrate revealed that the three-dimensional kinase fold of Akt engages mTORC2, creating what we refer to as the “substrate arch”. The local environment was highly dynamic, allowing us to resolve it only to 5.5 Å. Yet this resolution was sufficient to reveal two distinct structural interfaces of mTORC2 that enable substrate docking. First, the Akt N-lobe makes contacts to the mSin1 CRIM domain over a broad surface (“Interface 1”), including directly engaging the CRIM “acidic loop”, which was previously reported to be involved in Akt binding²⁵. Second, at the opposite end of the mTORC2-docked Akt, the kinase domain C-lobe engages a flexible interface we named the “mSin1 N-mooring” comprising N-terminal regions of mSin1 (“Interface 2”). The Akt kinase itself was captured in the inactive conformation⁵⁵, with no apparent density corresponding to helices α B, α C, or the PH domain. This suggests high conformational flexibility, consistent with our observation that the Akt PH domain is not essential for mTORC2-catalyzed phosphorylation of the Akt C-tail. mSin1 CRIM is stabilized against mLST8 by binding of the α 6-2 loop (mSin1 residues 163-174) to mLST8 1.4 (residues 35-39).

In the active site of mTORC2, 11 residues of the Akt C-tail are resolved (Arg466-Ala476), filling the mTOR active site (**Fig. 3.13a**). Akt-linked Torin1 assumes a nearly identical conformation to the analogous Torin2-bound crystal structure from which it was designed⁵³, with key contacts to mTOR Trp2339, Val2240, Ile2356, and Asp2357 (**Fig. 3.13b**). The peptide backbone takes a similar path like that seen in the two other PIKKs (phosphatidylinosi-

tol 3-kinase-related kinases) previously resolved with substrates: SMG1 (PDB 6Z3R) and ATM (PDB 8OXO), although the orientation is reversed (**Fig. 3.13c**). The backbone and sidechains of the Akt C-tail make a series of electrostatic and hydrophobic interactions with mTOR. Notably, Akt Phe472 engages a deep hydrophobic pocket (**Fig. 3.13d**). Together these contacts may partially explain mTOR's slight preference for Ser473 and accommodation of diverse sequences, but further study is needed using unmodified peptides.

3.3.6 Interface 1: Akt N-Lobe – mSin1 CRIM

The Akt N-lobe binding to mSin1 CRIM is facilitated by the positively charged surface spanning the five-stranded β -sheet of the Akt N-lobe that is well conserved (**Fig. 4.3a**, **Fig. 3.10e**, **Fig. 3.17a,b**). The basic sidechains of Akt N-lobe residues Lys158, Lys163, Lys182, and Arg222 directly engage the negatively-charged CRIM acidic loop (aa236-245), as does adjacent Akt Thr219, (**Fig. 4.3a**, right panel). As mentioned above, cellular phosphorylation of Ser473 was moderately impaired with transiently expressed K182A Akt (**Fig. 4.2a**, **Fig. 3.10g,h**). To further investigate the importance of this Akt interface, we explored a series of single and multiple Ala replacements of Akt Lys158, Lys163, Lys182, and Arg222 as well as flanking Thr219, His220, and Asp221. Interestingly, single K158A, K163A, and R222A mutant transfected Akts showed similar pSer473 levels to WT and reduced pThr450, especially in non-growth factor stimulated cells. Double mutant KK158/163AA and the quadruple mutant KKKR158/163/182/222AAAA transfected Akt forms displayed sharply decreased pSer473 and pThr450 levels (**Fig. 4.3b**, **Fig. 3.17b**). Although transiently expressed H220A and D221A Akt phosphorylation levels appeared similar to that of WT Akt, T219A Akt displayed a steep fall in pSer473 without change in pThr450 (**Fig. 4.3b**, **Fig. 3.17c-d**). *In vitro* kinase assays with mTORC2 revealed approximately 50% lower Ser473 phosphorylation of purified T219A Akt protein compared with WT Akt. (**Fig. 3.17e**).

Figure 3

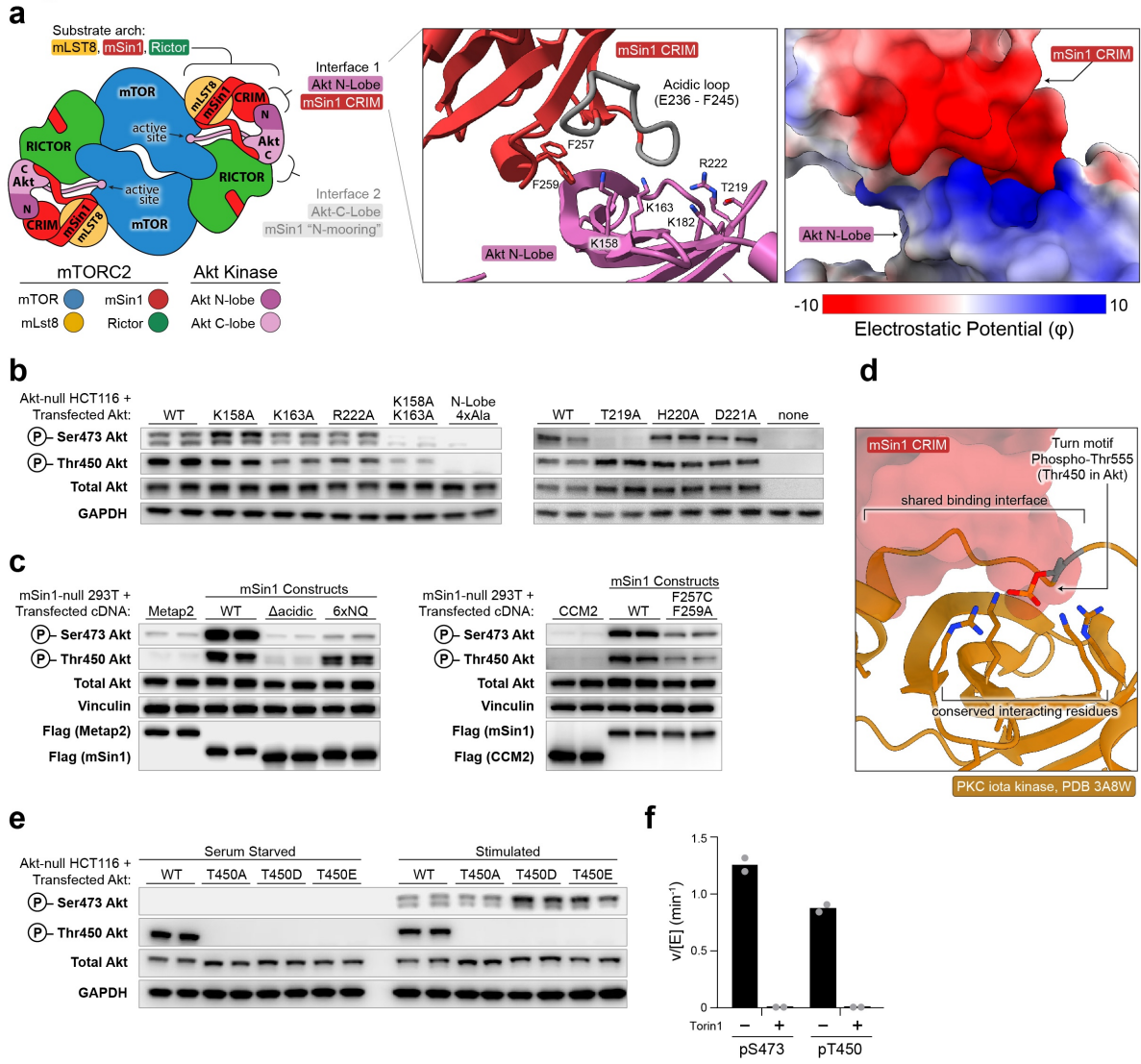


Figure 3.3: Interplay between Thr450 Phosphorylation and mTORC2-Akt Interface 1: Akt N-Lobe – mSin1 CRIM. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) **a**, Cryo-EM structure and model of this interface shows a broad docking surface that includes two loops on each mSin1 CRIM and Akt kinase domain N-lobe that have strongly complementary charges, as seen by the Coulombic electrostatic potential (calculated with ChimeraX). **b**, Key Akt residues assayed by transfection in Akt-null HCT116 for impacts on both hydrophobic motif pSer473 and turn motif pThr450 phosphorylation show (1) Thr219 and (2) a basic patch of four residues K158, K163, K182, and R222A together drive the interaction (all four basic residues mutated in N-lobe 4xAla construct). Double K158A-K163A and quadruple 4xAla Akt mutations largely abrogate both pSer473 and pThr450 levels despite equal total Akt expression. **c**, Both CRIM loops are required for Akt phosphorylation, as assayed by complementation transfection experiments in mSin1-null HEK293T cells with endogenous Akt substrate. Deletion of the acidic loop (Δ acidic, aa236-245) or mutation of all 6 acidic residues in the acidic loop (6xNQ), or mutation of Phe259 + Phe259 on the adjacent loop all reduce pSer473 and pThr450. **d**, mSin1-CRIM binds to AGC kinase N-lobe at a shared binding interface with the stabilizing intramolecular interaction with phosphorylated turn motif (pThr450 in Akt, pThr555 in the PKC-iota crystal structure PDB 3A8W, shown). Remarkably, overlaying the mTORC2-Akt structure PKC-iota crystal highlights that CRIM and phosphorylated turn motif threonine bind the same four key basic residues, and that both could not simultaneously bind. **e**, Test of the hypothesis that pThr450 may compete with Ser473 phosphorylation by transfection of ‘phospho-mimic’ and control alanine mutants in HCT116 cells. Increased pSer473 with T450D and T450E mutants is consistent with the expected weaker binding of sidechain carbonyl groups than a phosphate to the basic patch. **f**, *In vitro* rate of phosphorylation of Akt Thr450 and Ser473 by mTORC2 is similar, supporting a similar direct phosphorylation mechanism of both turn and hydrophobic motifs (full gels for b-f in **Fig. 3.17**).

To test the importance of the mSin1 CRIM surfaces in recruiting the Akt N-lobe, we generated mSin1-null HEK293T cells and complemented them in transient expression experiments with WT mSin1 or variants. mSin1 lacking the acidic loop (acidic) or with replacement of the 6 acidic loop aspartates and glutamates with asparagine and glutamine (6xNQ, previously ‘poly NQ’)²⁵ demonstrated greatly reduced pSer473 Akt levels compared with WT mSin1. Interestingly, acidic loop deleted mSin1 also displayed a large decrease in pThr450, whereas the effect of 6xNQ mutation on pThr450 was less pronounced (**Fig. 4.3c**, **Fig. 3.18a,b**). Neither of these mutants compromise the integrity of mTORC2 complex formation as measured by co-immunoprecipitation with mSin1 of mTOR, mLST8, or Rictor (**Fig. 3.18c**, **Fig. 3.23b**). We also tested the role in binding of the adjacent loop on CRIM, where residues Phe247 and Phe259 appear to contact Akt N-lobe. Mutation of these residues

markedly reduced both Akt pSer473 and pSer450 levels (**Fig. 4.3c**).

In addition, we prepared several recombinant mSin1 protein fragments and tested them as potential competitive inhibitors of mTORC2-catalyzed phosphorylation of Akt protein. These experiments revealed that both CRIM the extended CRIM-aRBD-PH multi-domain fragments are modest inhibitors of mTORC2 activity, with IC50 values of approximately 130 and 50 μ M, respectively; the slight increase in potency when aRBD-PH is added argues these domains may make a small contribution to recognition (**Fig. 3.18**). Deletion of the CRIM acidic loop (aa236-245), the 6xNQ mutations, or mutation of conserved aliphatic residues in the acidic loop ('AA' mutant) in CRIM²⁵ abolished CRIM's inhibition of mTORC2, further supporting a role of CRIM and its acidic loop in Akt recruitment (**Fig. 3.18**).

Remarkably, the basic Akt N-lobe patch (residues Lys158, Lys163, Lys182, and Arg222) required for CRIM binding were previously shown to form an intramolecular interaction with the phosphate of turn motif pThr450, stabilizing cellular Akt; this stabilizing interaction is proposed to be present in 26 of the 63 annotated AGC kinases³⁸. A crystal structure of PKC- ι shows pThr450-equivalent pThr555 bound in this basic pocket (3A8W)⁵⁶, highlighting the structural conservation with Akt. Overlaying this with our structure of CRIM-Akt reveals that this is a shared binding interface that could be occupied by either pThr450 or mSin1-CRIM, but not both (**Fig. 4.3d**). To test whether competition between these binding modes has functional consequences, we assayed cellular mTORC2 phosphorylation of Akt with mutations of Thr450 to alanine, aspartate, and glutamate. The acidic mutants increased pSer473, whereas the alanine mutation had little effect (**Fig. 4.3e**), similar to prior studies³⁸. The pThr450 antibody did not recognize any of these mutants, further validating its specificity. We interpret these data to mean that Thr450Asp/Glu mutants, which presumably have reduced coordination of the basic Akt N-lobe residues as compared to pThr450, facilitate CRIM binding by reducing competition with this intramolecular interaction, thereby increasing Ser473 phosphorylation. While, based on this model, the T450A mutant would be expected to ablate intramolecular N-lobe binding and further potentiate

Ser473 phosphorylation, reduced stability, increased phosphatase susceptibility,^{13,29} or binding to the mTOR active site as a dead-end substrate analog likely offsets any potential gains in CRIM recognition.

Prior combinatorial mutagenesis of Lys158, Lys163, Lys182, and Arg222 resulted in a steep reduction of both cellular Akt levels and Thr450 phosphorylation^{29,38}. Decreased pThr450 accompanying the loss of these four basic residues was understandably interpreted as enhanced susceptibility to cellular phosphatases. However, genetic studies show mTORC2 is required for Thr450 phosphorylation¹³, and phospho-Thr450 is generated in reactions with crudely immunoprecipitated mTOR²⁹. Thus, we hypothesized that mTORC2 could directly phosphorylate Thr450. To test this, we generated phosphate-free Akt C-tail by treatment of immobilized Akt with lambda protein phosphatase prior to expressed protein ligation. We observed that mTORC2 phosphorylated this phosphate-free C-tail on Akt at Thr450 at a similar rate to that of Ser473, and that the reaction was also inhibited by Torin1 (**Fig. 4.3f**, **Fig. 3.17f**). Overall, these experiments substantiate the importance of the Akt N-lobe-mSin1 CRIM interface, as observed in our cryo-EM structure, in mediating mTORC2 phosphorylation of both Ser473 and Thr450 of Akt.

3.3.7 Interface 2: Akt C-Lobe – mSin1 N-Mooring

The cryo-EM analysis also revealed a novel but flexible substrate docking site for Akt kinase domain C-lobe that involves residues near the N-terminus of mSin1 that we refer to as “mSin1 N-mooring” (**Fig. 4.4a**). Here, the Akt α G- α H loop containing residues 366-368 interacts directly with mSin1 Trp76 and Phe78, including hydrophobic interactions between (1) mSin1 Trp76 and Akt Phe368 and Ile367 and (2) mSin1 Phe78 with Akt Ile367, as well as (3) a potential interaction between mSin1 Phe78 with Akt Arg366 (**Fig. 4.4a**). To probe these interactions, we first mutated residues 366-368 on Akt to either glycine or glutamate, which showed sharply reduced Ser473 and Thr450 phosphorylation despite equal Akt expression (**Fig. 4.4b**). Mutation of mSin1 residues Trp76 and Phe78 to serine or glycine showed a

modest ~30% decrease in phosphorylation; curiously for the di-glycine mutant, the reduction in pThr450 was more pronounced than in pSer473, whereas for di-serine, the defect in pSer473 was larger but pThr450 did not reach statistical significance (Fig. 4.4c).

Figure 4

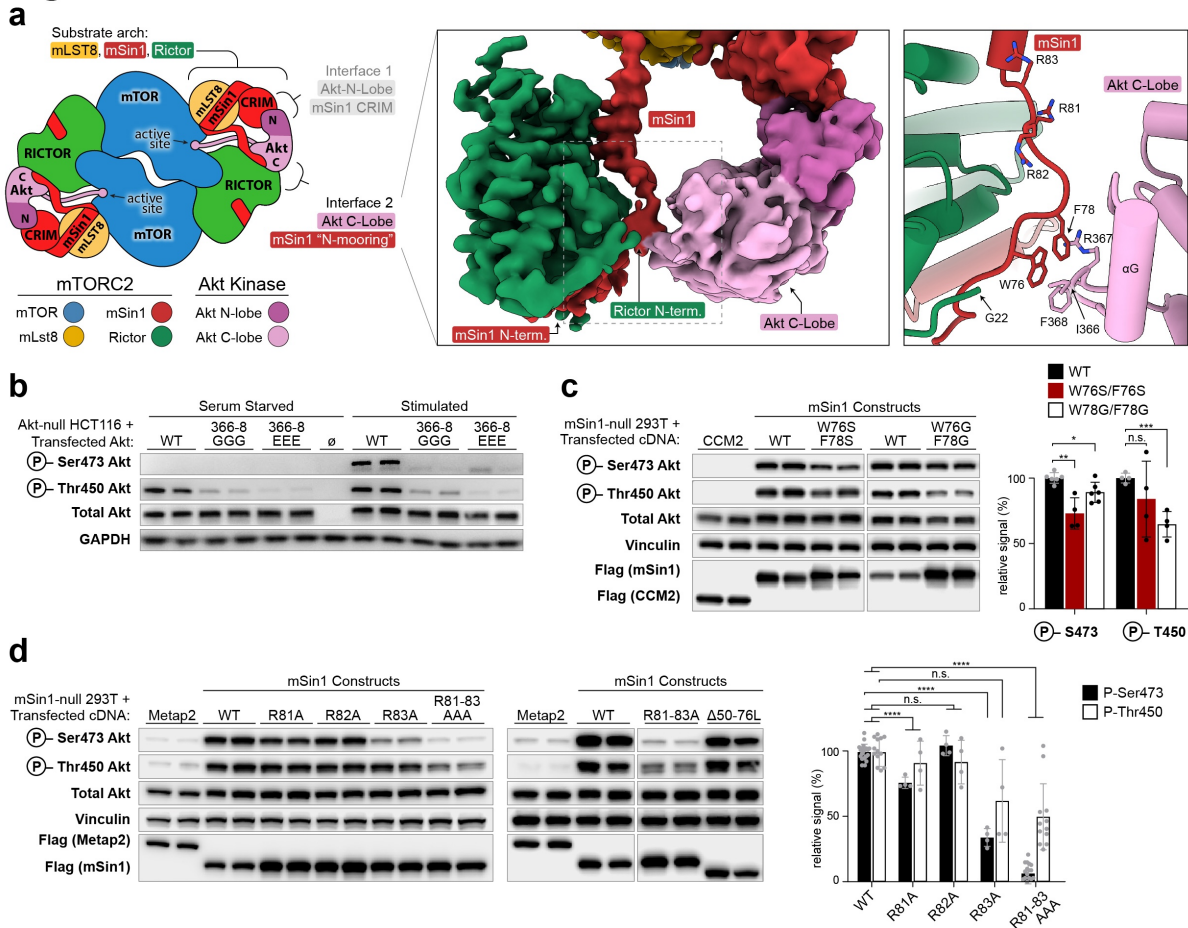


Figure 3.4: mTORC2-Akt Interface 2: Akt C-Lobe – mSin1 N-mooring. (Figure caption continued on next page.)

(Figure caption continued from previous page.) **a**, Cryo-EM structure of this flexible docking interface shows Akt N-terminus and Ile366-Phe368 abut the mSin1 N-terminal region, especially Trp76 and Phe78, with the unstructured N-terminus of Rictor in close proximity. **b**, Mutation of Akt residues 366-368 to glycine or glutamate ablates Ser473 phosphorylation and dramatically limits Thr450 phosphorylation in Akt-null HCT116 cells without affecting total protein levels. **c,d**, To assay contributions from mSin1, WT and mutant constructs were transfected into mSin1-null HEK293T cells. **c**, residues Trp76 and Phe78 were mutated to glycine or serine, which show modest effects on both Ser473 and Thr450 phosphorylation (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; n.s., not significant; $n=4$ points from $n=2$ independent experiments). **d**, Arginines 81-83, >15 Å from docked Akt, have been proposed to be important for SGK but not Akt phosphorylation by mTORC2. To test the importance of these residues, they were simultaneously mutated to Ala (R81-83A), which shows large impacts on Ser473 and Thr450 levels. Mutation of the individual residues show these effects are mostly due to R83, with lesser contribution from R81 and no significant contribution from R82 (**** $p < 0.0001$). To investigate the contribution of unstructured and bridging mSin1 residues 50-76, these were deleted and replaced with a 5-residue GGGGS linker ($\Delta 50-76L$), which includes Trp76. This shows no effect on Akt phosphorylation, which may be explained by backbone interactions. (Note, these gels are from the same experiment as figure 3c, left panels; all lanes shown in **Fig. 3.17e**).

Because of flexibility, the mSin1 N-mooring appears to also engage two helices in the Akt C-lobe and their connecting loop αH and αI , together spanning residues 374-408. Conservation in this region of Akt (**Fig. 3.19b**) suggested residues Ala376 and Lys386 might be involved in substrate recognition or local domain stability. Indeed, mutant A376G Akt and K386A Akt reduce mTORC2-catalyzed phosphorylation of Ser473 without impacting Thr450 phosphorylation (**Fig. 3.19a**). Because the highly conserved N-terminus of Rictor is also in close proximity to the Akt C-lobe in this region but residues 1-21 are not resolved structurally, we investigated its importance in Akt phosphorylation. Rictor constructs in which the first 16 and 26 residues are deleted (1-16, 1-26) complement Akt phosphorylation equally to WT in Rictor-null HEK293T cells and do not affect mTOR complex formation (**Fig. 3.19**), ruling out a significant contribution of these Rictor residues in Akt recognition.

The conserved mSin1 tri-arginine cluster R81-R83, situated at the edge of the N-mooring, was previously interpreted as important for mTORC2 phosphorylation of SGK1, but not for Akt⁵. To examine its function, we performed Akt phosphorylation assays with single and

triple Ala mutants. The triple RRR81-83AAA mutant nearly completely ablates Ser473 phosphorylation and causes around a 50% reduction in pThr450 levels; most of this can be explained by R83, with lesser contributions from R81 more than R82 (**Fig. 4.4d**). However, these arginine residues are positioned further than 15Å away from docked Akt. Together with data showing mTORC2 assembly with these mutants is grossly intact (**Fig. 3.17e**, **Fig. 3.23b**), arginine residues 81-83 are more likely required for local structural configuration of mTORC2 rather than direct contact with substrate.

The large effect of mSin1 R81-83 mutations contrasts with the modest effect of mutating Akt-abutting residues Trp76 and Phe78. To further probe the role of the N-terminal region of mSin1 in substrate recruitment, we replaced a large unstructured loop and extra bridging residues comprising residues 50-76 of mSin1 with a 5-residue GGGGS linker (50-76L, **Fig. 3.19d**). This mutant shows no effect on Akt phosphorylation by mTORC2 (**Fig. 4.4d**). A significant component of the mSin1 interaction in the N-mooring region may therefore be driven by or be complemented by interactions between the mSin1 backbone with the Akt C-lobe. Of note, both the mSin1-N-mooring - Akt C-lobe interaction and the mSin1 CRIM - Akt N-Lobe interaction are required in concert for mTORC2-Akt recognition and phosphorylation.

3.3.8 Integrative Modeling of the mTORC2:Akt co-Complex

To further describe the dynamic structure of mTORC2 in complex with Akt, we computed an integrative structure model of the complex using IMP⁵¹, based primarily on the cryo-EM and XL-MS data (**Fig. 4.5a**). We generated 5 million potential mTORC2-Akt structures through extensive conformational sampling, aiming to fit the structures in the map while satisfying as many cross-links as possible. While the model computed from the XL data alone results in two distinct solutions, discussed above, clustering of the mTORC2-Akt structures that sufficiently satisfied both the XLs and cryo-EM map revealed a unitary cluster (ensemble) satisfying 90% of the cross-links (181/201, **Fig. 4.5a**, **Fig. 3.20**, **Fig. 3.21**). Our integrative

model, for the first time, provides structural insights into the entire mTORC2 complex, including domains and loops that have been undetectable in cryo-EM structures: the Akt PH domains, mSin1 domains (CRIM, aRBD, and PH), and large, flexible loops within Rictor and mTOR. Together, these additional components account for approximately 26% of mTORC2 residues (Fig. 3.21b).

Figure 5

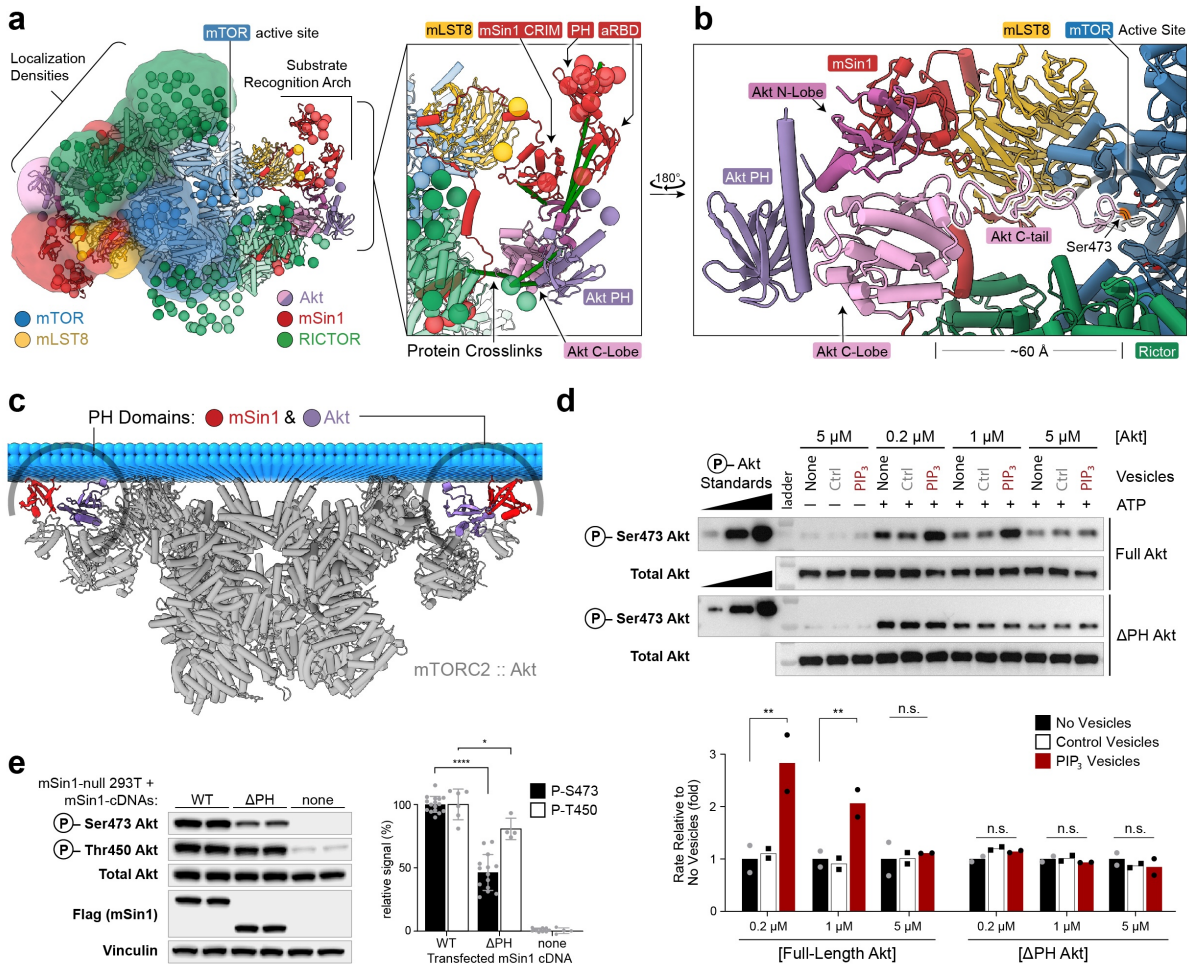


Figure 3.5: Integrative Modeling and Membrane Association of mTORC2-Akt. (Figure caption continued on next page.)

(Figure caption continued from previous page.) **a**, Integrative model centroid solution of mTORC2:Akt co-complex based on crosslinking mass spectrometry and biophysical constraints. Each sphere is a “flexible bead”, representing 10 residues, and localization densities, shown on one symmetric mTORC2 monomer, represent the variability found within the ensemble for each component. The modeling localizes the remaining components of mTORC2, including PH domains of both Akt and mSin1 and the mSin1 aRBD (inset). **b**, mTORC2:Akt interfaces from centroid structure; cryo-EM map (gray) and prv otein cross-links (navy blue pipes) shown. The modeling recapitulates the positioning of Akt in both Akt kinase N-lobe with mSin1 CRIM and the Akt C-lobe with the “mSin1 N-Mooring”. **c**, Model of Akt C-tail phosphorylation. The flexible C-tail (aa 430-480), modeled from the last resolved residue in the Akt C terminus (Ser422), extends $\sim 75\text{\AA}$ to the mTOR active site with abundant slack and without violating excluded volumes. The Akt PH domain (purple) is behind the kinase C-lobe in this view and does not encumber phosphorylation. Thr450 can be similarly accommodated (see **Fig. 3.21c**) **c**, Model of mTORC2 membrane engagement: the 4 PH domains (2 each mSin1 and Akt) in the mTORC2 homodimer can be coplanar within the ensemble, shown interacting with the phospholipid bilayer e.g. at the plasma membrane. **d**, *In vitro* vesicle kinase assay in which SUVs (1 mM total lipid) are pre-incubated with 5 nM mTORC2 and the indicated concentration of WT Akt or Akt lacking the PH domain ($\delta\text{PH-Akt}$, amino acids 144-480), and reacted for 10 min at 30°C with 1 mM ATP, 10 mM MgCl₂ (except where indicated) and analyzed by quantitative immunoblotting. Ctrl, control vesicles (PC:PE,1:1). PIP₃ containing vesicles (PC:PE:PS:PIP₃,30:50:19:1). $n=4$; **, $p<0.002$ for 2-fold rate increase at 0.2 μM and 1 μM , but not at 5 μM (two-way ANOVA). **e**, Effect of mSin1 PH domain on Akt phosphorylation. mSin1 lacking the C-terminal PH domain (ΔPH) or WT was transfected into mSin1-null HEK293T cells; $n=4-16$ from experiments done in duplicate.

The integrative model uniformly presents a fully engaged Akt substrate with its kinase domain N-lobe oriented towards CRIM and C-lobe oriented towards the mSin1 N-mooring. The Akt kinase domain is in an inactive conformation similar to allosterically inhibited and autoinhibited Akt crystal structures (PDB 3O96 and 7APJ), and the PH domain is positioned at the boundary of the kinase domain lobes, facing away from mTORC2, in a unique orientation distinct from either structure or the Alphafold model (**Fig. 4.5a,b**, **Fig. 3.21c**, **Fig. 3.22a**)⁵⁵. The simulated location of Akt and mSin1 CRIM is fairly consistent across models, but the positions of the mSin1 aRBD and PH domains show more flexibility. Next, we examined C-tail phosphorylation computationally. Based on previous crystal structures of Akt and disorder prediction algorithms^{55,57}, the Akt sequence C-terminal to residue ~ 430

is largely flexible. Accounting for the excluded volume of Akt and mTORC2, the Akt C-tail is sufficiently long to traverse the ~ 65 Å distance to the mTOR active site and position either Thr450 or Ser473 for phosphorylation with considerable residual slack in the C-tail (**Fig. 4.5b**, **Fig. 3.21c**). Finally, the model can place the PH domains of mSin1 and Akt in close proximity and coplanar with respect to the lipid bilayer, in a conformation that allows all four PH domains within the mTORC2-Akt dimer to simultaneously engage membrane phospholipids (**Fig. 4.5c**, **Fig. 3.22b**).

3.3.9 Membrane recruitment in mTORC2's phosphorylation of Akt

We hypothesized that the co-association of Akt and mTORC2 to PIP3-containing membranes might enhance mTORC2-mediated phosphorylation of Akt through increased relative concentration and preferential orientation relative to the membrane. To test this, we prepared vesicles containing PIP3 and control vesicles, which in an ultracentrifugation assay demonstrated PIP3-dependent recruitment of Akt (**Fig. 3.22c**). In addition, only vesicles containing PIP3 were able to stimulate mTORC2-catalyzed phosphorylation of Akt at low Akt concentrations (0.2 and 1 μM) but less so at higher Akt concentrations (5 μM). This stimulatory effect with full-length Akt was not observed with $\Delta\text{PH-Akt}$, and could be blocked by addition of soluble PIP3, indicating that membrane embedded PIP3-mediated recruitment to the vesicles was necessary (**Fig. 4.5d**)³⁸. These *in vitro* kinase results appear to correlate with how PIP3 elevation in cells can trigger Akt phosphorylation by mTORC2. To explore the role of the mSin1 PH domain, we employed the mSin1 knockout cells and observed that complementation with $\Delta\text{PH-mSin1}$ relative to WT mSin1 led to decreased pSer473 and slightly decreased pThr450 (**Fig. 4.5e**). Together, these experiments demonstrate a key role for membrane co-recruitment of mTORC2 and Akt by their respective PH domains for efficient Ser473 phosphorylation. Such membrane recruitment may be dispensable for Thr450 phosphorylation, perhaps due to stability of the modification conferred by

binding to the basic patch³⁸.

3.3.10 Mechanistic similarity in mTORC2 recognition and direct activation of Akt, PKC, and PKN

To test whether the mechanism by which mTORC2 recognizes the 3D structure of Akt is generalizable to other substrates, we assayed the phosphorylation of the PKC α C-tail at both hydrophobic (Ser657) and turn motifs (Ser638). We also evaluated AGC member PKN1, in which the hydrophobic motif Ser is replaced with a “phosphomimetic” Asp, and which has been reported to have mTORC2-regulated phosphorylation of the turn motif Ser916³⁰ (**Fig. 3.6a**). In transient expression experiments in HCT116 cells, following growth factor stimulation, PKC α Ser657 was robustly phosphorylated, PKC α S657A showed no phosphorylation, and Torin1 treatment prevented phosphorylation, confirming the expected mTORC2 dependence. As seen in Akt, catalytic null PKC α D463A showed similar or higher levels of phosphorylation as WT PKC α , together supporting direct phosphorylation of the PKC hydrophobic motif by mTORC2 rather than autophosphorylation (**Fig. 3.6b**, **Fig. 3.23a**). To test whether the mode of mTORC2 recognition of PKC is like that of Akt, we made analogous mutations in the kinase domain N-lobe as those that disrupt the Akt-mSin1 CRIM interaction: Akt K183A equivalent PKC α K372A and Akt T219A equivalent PKC α T409A, along with trialanine ‘3A’ mutants that include adjacent residues. These mutants show reduced phosphorylation and unperturbed expression levels (**Fig. 3.6b**, **Fig. 3.23a**).

Figure 6

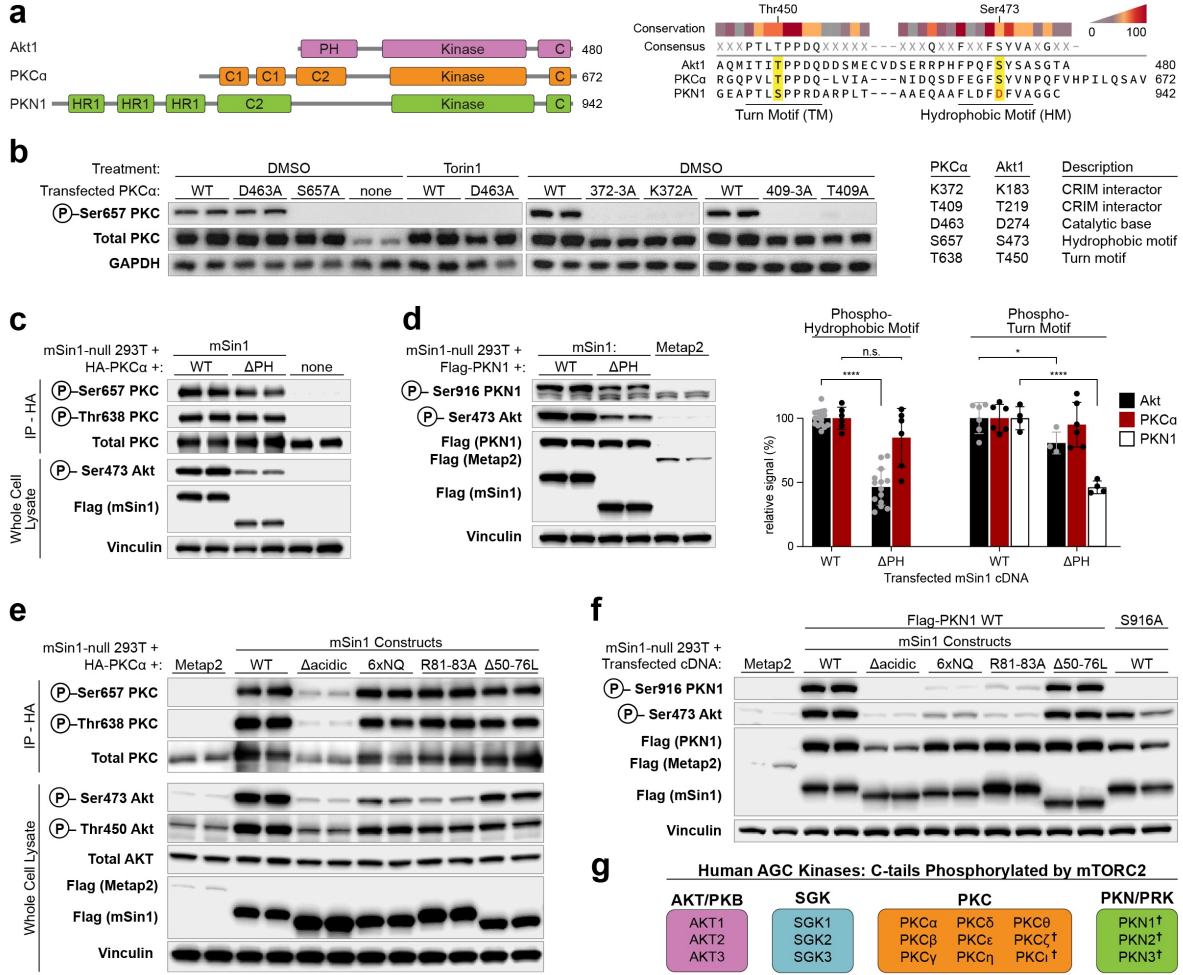


Figure 3.6: mTORC2 Similarly Phosphorylates AGC Kinases at Both Turn and Hydrophobic Motifs. (Figure caption continued on the next page.)

(Figure caption continued from previous page.) **a**, Domain diagram of Akt1, PKC α , and PKN1 and alignment of turn motif (TM) and hydrophobic motif (HM). The HM residue for PKN1 is instead a ‘phosphomimic’ aspartate. **b**, PKC phosphorylation assay in HCT116 cells mirrors Akt results. Cells were transfected with PKC α constructs, serum-starved, and treated with DMSO or Torin1 for 1 hour prior to stimulation with insulin and hIGF1 and immunoblotting. Catalytic-null D463A PKC α is indistinguishable from WT, and interfaces mirroring Akt K183 (PKC α K372) and Akt T219 (PKC α T409) are similarly required for HM phosphorylation at S657. **c,d**, Effect of mTORC2 PH domain on transfected substrate PKC α HM (Ser657) and PKN1 TM (Ser916) phosphorylation; comparison to endogenous Akt serves as an internal control. Quantified blots highlight that PKC α phosphorylation does not depend on the PH domain, whereas PKN1 TM shows a similar dependence on the PH domain as the Akt HM. n=4-16 points from independent experiments each done in duplicate. **e,f**, mSin1 and domain contribution to PKC α and PKN1 phosphorylation in mSin1-null HEK293T cells transfected with WT mSin1 or the indicated mSin1 mutants. Δ acidic lacks the acidic loop, 6xNQ mutant removes the six acidic residues from the acidic loop, and R81-83A and 50-76L constructs are in the mSin1 N-terminus. **e**, mSin1 presence and acidic loop are required for PKC phosphorylation; N-terminal constructs that affect Akt have less effect on PKC α . **f**, PKN1 TM Ser916 phosphorylation depends on mTORC2 status and is not detected in S916A mutant. mSin1 mutations have nearly identical effects on PKN1 as Akt. Domains: C1, C2, conserved regions; HR1, heptapeptide repeat 1; these domains of PKC α and PKN1 are involved in regulation and membrane targeting. **g**, AGC kinases phosphorylated on C-tail by mTORC2. Both TM and HM are directly modified, except for kinases marked †, in which HM Ser/Thr is Asp/Glu and only TM is targeted. * p<0.05; **** p<0.0001;

To test the importance of mTORC2 membrane association in these kinases, we assayed phosphorylation of transfected PKC α and PKN1 alongside endogenous Akt1 in mSin1-null HEK293T cells complemented with WT or Δ PH mSin1. We found that mTORC2 is required for phosphorylation of PKC α and PKN1 turn motifs and the PKC α hydrophobic motif. While the loss of the mSin1 PH domain had no impact on PKC α phosphorylation by mTORC2 (**Fig. 3.6c**), phosphorylation of the PKN1 turn motif Ser916 was reduced by approximately 50%, similar to the reduction seen in the Akt hydrophobic motif Ser473 (**Fig. 3.6d**).

To gain further mechanistic insight, we examined our panel of mSin1 CRIM mutations with Akt, PKC α and PKN1 in mSin1-null HEK293T cells (**Fig. 3.6e,f**). Phosphorylation of both PKC α hydrophobic and turn motifs requires the CRIM acidic loop, but no effect was seen with less disruptive mSin1 mutations 6xNQ and R81-83A. In contrast, PKN1 turn motif

pSer916 behaves nearly identically to Akt hydrophobic motif pSer473 in these experiments. Contradicting a prior hypothesis⁵, the consistent requirement for these arginine residues in Akt and PKN phosphorylation and the previous report of their importance in SGK phosphorylation argue against a role for these Arg residues in the specific recognition of these AGC kinase subsets. PKC α recognition by mTORC2 appears more robust and less sensitive to perturbation than Akt or PKN1. Strikingly, when transiently expressed, the stabilities of PKN1 and PKC α are dramatically reduced absent mTORC2 activity, as measured by total protein levels, whereas the effect of mTORC2 activity in stabilizing endogenous Akt in these cells is subtle but statistically significant over many experiments (**Fig. 3.23c**). Together, these data demonstrate that mTORC2 recognizes and activates PKC and PKN similarly to Akt, with long-range docking and a key interaction between the kinase domain N-lobe and mSin1 CRIM leading to direct phosphorylation of both hydrophobic and turn motifs by mTORC2.

3.4 Discussion

Using a combination of biochemical, structural, and cellular experiments, we have demonstrated that mTORC2 recognizes the three-dimensional structure of the Akt kinase domain at two distinct locations distal to the active site to facilitate direct phosphorylation of the Akt C-tail by mTOR. Our results support that Thr450 and Ser473 phosphorylation of Akt involve a similar Akt-mSin1 interface, although Ser473 phosphorylation is generally more sensitive to its disruption than Thr450 phosphorylation. This differential phosphorylation likely due in part to pThr450 protection from dephosphorylation by intramolecular interactions with the Akt N-lobe basic cluster Lys158, Lys163, Lys182, and Arg222^{29,38}. As this N-lobe basic cluster also appears to mediate mTORC2 interaction by binding the mSin1 CRIM acidic loop, we have identified an intricate interplay between pThr450 and pSer473. The fact that Ser473 phosphorylation is not significantly inhibited by the presence of pThr450

suggests that the interaction between pThr450 and the basic cluster is sufficiently weak to allow for efficient competition by the acidic loop of the CRIM domain, likely aided by additional CRIM residues Phe257 and Phe259 that together engage a broader surface area on Akt including Lys183 and Thr219.

Our data also show that the mode of recognition of Akt by mTORC2 is conserved among the family of AGC kinases including PKC, PKN, and likely others, where phosphorylation of hydrophobic motif, turn motif, or both are regulated. In addition, our vesicle and cellular assay data substantiate a role for PIP3-mediated membrane recruitment of Akt in mTORC2 catalyzed phosphorylation of Akt, and cellular data also show a similar role for membrane recruitment in PKN phosphorylation by mTORC2. More work is needed to understand these events and their regulation in physiology, however. While mTORC2 has been proposed to be regulated and mediate signal transduction through small G proteins including Ras and Rho isoforms in various contexts and across organisms^{27,28,58}, we have limited understanding of which signals are required, how they are integrated by mTORC2 to control downstream processes, and how this is dysregulated in diseases such as diabetes and cancer.

The data presented here could aid in the development of both broad-spectrum PIKK and selective mTORC2 inhibitors. By revealing the first high-resolution structure of a PIKK active-site pocket bound to a substrate (2.5 Å), we lay a strong foundation for the rational design of more potent and specific small molecule active site inhibitors. Even though our structure does not reveal a clear small molecule pocket that can be targeted to select between mTORC2 and mTORC1, we reveal several surfaces as key to such selective targeting, including mSin1 CRIM. The Akt kinase domain N-lobe could also be engineered as a dominant negative genetic tool, like the dominant negative mSin1 construct lacking CRIM that was recently reported to slow growth in a melanoma model⁵⁸.

Taking together the distinct sequences of turn and hydrophobic motifs mTORC2 recognizes (**Fig. 3.6g**, **Fig. 3.23d**), available binding information about mTORC1 with 4E-BP1²², the lack of specificity of mTOR for its known substrates in peptide array profiling

experiments⁵⁹, that mTORC2 can phosphorylate Akt Ser477 and Thr479⁴⁶, and our mass spectrometry data showing mTORC2 can doubly phosphorylate the Akt tail, we propose that mTOR displays limited local sequence specificity in phosphorylation², and that regulation is instead governed by remote docking interactions that control access to the mTOR active site. Structural insight into mTORC2 substrate recognition has been challenging and was unsuccessful in prior efforts^{4,5} for at least two key reasons. First, docking affinity in solution is low (mTORC2-Akt Kd > 10 μ M), necessitating membrane co-localization for efficient phosphorylation. Despite stoichiometric tethering of Akt-Torin to mTORC2 via the active site, only about 6% of all mTORC2 particles fully recruited Akt to the substrate arch. Second, the identified docking interfaces are degenerate and broad, accommodating a series of slightly different binding poses that blur the position of Akt in our cryo-EM maps to 5.5 Å resolution. Adjacent residues often compensate for point mutations, requiring larger deletions of entire loops or residue clusters to fully inhibit binding in two different modes: the mSin1 CRIM and Akt N-lobe interface employs complementary charged surfaces (interface 1), and the Akt C-lobe establishes redundant hydrophobic contacts with mSin1's N-mooring that may also involve backbone interactions (interface 2). From a technical perspective, the use of a small molecule ATP-competitive inhibitor site-specifically attached to a protein substrate as developed here suggests a generalizable strategy to facilitate structural determination of protein kinases in complex with protein substrates.

Acknowledgements

We thank K. Burns for support and discussions; B. Paschal and C.-S. Yang for reagents and protocols for studying PKN phosphorylation; S. Whedon for helpful advice on semisynthesis; H. Bae for Akt constructs; all members of the Cole and Sabatini labs for helpful insights; E. Brignole and S. Sterling for assistance in cryo-EM data collection at MIT.nano and Harvard Medical School; S. Andersson and A. Dastous for help with model building; Stanford Re-

search Computing Center and Whitehead Institute for providing computational resources; Daniel Rauh for helpful discussions on Akt allosteric inhibitors; B. Webb and I. Echeverria for help with integrative modeling. This work was supported by the National Institutes of Health grants R01 CA074305 (PAC, MC, and BP), K99/R00 CA255926, R35 GM150935, P30 CA124435 (KBR), K08 DK129824-01 (MST), K22 CA241105 (NC), F32 CA259214 (BAP), T32 CA009302 (KYL-G), R35GM137905, 5R21CA261737 (YS), R01 CA233800 (JAM), R01 GM083960, P41 GM109824 (AS), and R01 CA103866, R01 CA129105, R01 AI047389 (DMS); The Mark Foundation for Cancer Research (JAM), The Massachusetts Life Science Center (JAM). MW is supported by a Bridging Excellence Fellowship provided by the Life Science Alliance.

Author Contributions

MST, MC, NC, MW, KBR, DMS, and PAC formulated the research plan and interpreted experimental results with assistance from BDM, YX, YS, KYL-G, DF, SC, KM, MH, and KX. MST, MC, NC, BDM, KYL-G, DF, SC, KM, MH, and KX designed and generated constructs, purified proteins, and performed and analyzed biochemical experiments. KBR and MST prepared samples for cryo-EM. MW and KBR acquired data and determined the cryo-EM structures. MW and KBR analyzed cryo-EM data. MH performed modeling experiments with design input from MST, DM, and AS. BP and SF performed mass spectrometry experiments with assistance from BDM, MC, and MST and analyzed results with input and supervision from JAM. BJB and NSG designed Torin Acrylamide with input from PAC, NC, and KBR and BJB synthesized the compound. YX, MST, and YS designed and carried out XL-MS experiments. MST and PAC wrote the manuscript with contributions from MC, NC, MH, MW, DMS, and KBR. All authors edited and approved the manuscript. NC, KBR, and PAC jointly supervised the project.

Competing interests

PAC has received consulting fees from Scorpion Therapeutics and Nested Therapeutics. MST holds equity and has received consulting fees from ROME Therapeutics. JAM is a founder, equity holder, and advisor to Entact Bio, serves on the SAB of 908 Devices, and receives or has received sponsored research funding from Vertex, AstraZeneca, Taiho, Takeda, Springworks, and TUO Therapeutics. NG is a founder, science advisory board member (SAB) and/or equity holder in Syros, C4, Allorion, Lighthouse, Voronoi, Inception, Matchpoint, GSK, Larkspur (board member), Shenandoah (board member) and Soltego (board member). The Gray lab receives or has received research funding from Novartis, Takeda, Astellas, Taiho, Jansen, Kinogen, Arbella, Deerfield, Springworks, Interline and Sanofi.

3.5 Material and Methods

Kinase sequences compared were fetched from Uniprot and aligned using MUSCLE⁶⁰ package in Jalview 2.11.3.2 and visualized using Snapgene version 7.1.1.

3.5.1 Reagents

All chemicals and buffers were from Sigma unless otherwise stated. EDTA-free protease inhibitor tabs, EDC, Sulfo-NHS, Dynabeads M-270 Epoxy, and anti-HA magnetic beads were from Thermo Fisher Scientific. DSS was from ProteoChem. PhosSTOP was from Roche. PEI MAX 40K was from Polysciences. Lipids and PIP3 were from Avanti Polar Lipids. InstantBlue Coomassie Protein Stain was from Abcam; a similar reagent was prepared by dissolving 80 mg Coomassie Brilliant Blue G-250 (Goldbio), 50 g EtOH, 80 g phosphoric acid, and 5 g α -cyclodextrin in a final volume of 1 L H₂O.

3.5.2 Antibodies

Rabbit monoclonal antibody to Akt pS473 (ab81283, EP2109Y) and total PKC α (ab32376) were from Abcam; Rabbit monoclonal mTOR (2983), Rictor (9476), mLST8 (3274), Total Akt (4685), Akt pT450 (12178), total 4E-BP1 (9644), 4E-BP1 pT37/T46 (2855), PKC α pT638 (9375), PKC α pS665 (9371), Flag epitope tag (14793), HA epitope tag (3724), HRP-linked anti-mouse secondary antibody (7076) and HRP-linked anti-rabbit secondary antibody (7074) were from Cell Signaling Technology. Flag-M2 antibody used for preparing magnetic beads was from Millipore Sigma (F1804). Anti-Raptor (09-217) was from EMD Millipore. Anti-adenosine was from LSBio (LS-C347801). Mouse monoclonal to vinculin was from Santa Cruz (sc-73614). Mouse monoclonal to mSin1 (Sigma 05-1044, clone 1C7.2) was knockout validated; Rabbit polyclonal to mSin1 (Bethyl A300-910A) was used for insect overexpression but did not detect endogenous mSin1 in HEK-293T or HCT116. p-Ser916 PKN1 antibody and non-phospho PKN1 blocking peptide were gifts from Bryce Paschal and used as described³⁰. We were unable to identify any commercial reagent to SGK1 (pS422) that showed signal in untransfected or SGK-transfected HEK293T or HCT116 cells that depended on the presence of mTORC2.

3.5.3 Tissue Culture

Akt1/2-null HCT116 cells⁴⁴ were a gift from Bert Vogelstein. HEK293T cells were from ATCC. SF21, SF9, and ExpiSF cells, DMEM, FreeStyle 293 Medium, Grace's Insect Medium, SF-900 Medium, ExpiSF CD Medium, Lipofectamine 3000, Cellfectin II, and Expifectamine SF were from Thermo Fisher Scientific. FBS and penicillin/streptomycin were from Sigma. McCoy's 5A medium was from Quality Biological. Amino acid-free and potassium-free medium was from US Biologicals. Adherent mammalian cells were maintained in a humidified 37°C incubator with 5% CO₂ with 10% FBS and 100 IU/mL penicillin, and 100 μ g/mL streptomycin and passaged 2-3 times weekly. Suspension-adapted HEK293T cells were grown 125 ml – 5 L polycarbonate or glass Erlenmeyer flasks in Freestyle 293 medium

supplemented with 50 IU/mL penicillin, and 50 $\mu\text{g}/\text{mL}$ streptomycin and 1% FBS in a 37°C incubator with 8% CO₂, shaken at 90-125 rpm, and maintained at densities from 0.4-5 million/ml. SF9 cells were maintained at 27°C in glass spinner flasks or Erlenmeyer shake flasks at densities 0.5-4 million / ml in SF-900 II or SF-900 III medium. ExpiSF cells were maintained similarly in Erlenmeyer flasks at densities from 0.5-15 million/ml in ExpiSF CD medium.

3.5.4 Akt and PKC Cell Signaling Assay in HCT116 cells

Akt1/2-null HCT116 cells⁴⁴ were maintained in McCoy's 5A medium supplemented with 10% FBS and 500,000 cells per well were seeded in 6-well plates. The next day, cells were transfected with 1.5 μg per well of WT or mutant Akt pcDNA3 plasmid with duplicates using Lipofectamine 3000 reagent according to the manufacturer's instructions. After 24 hours, cells were washed in PBS, and starved for 18-20 hrs in McCoy's 5A with 0.5% FBS. Where noted "serum starved", the cells remain untreated until plates were collected. Where noted "stimulated", cells were treated with McCoy's 5A without FBS but with 200 ng/ml of insulin (ThermoFisher Scientific) and 60 ng/ml human IGF-1 for 40 minutes at 37°C. Cells were placed on ice, washed in ice cold PBS, and lysed in 100 μl RIPA buffer (CST) containing 1x complete protease inhibitor tablet and 1 mM phenylmethylsulfonylfluoride (PMSF), incubated on ice for 5 minutes before cells were scraped for harvest and cleared by centrifugation at 21,000 \times g for 10 minutes at 4°C. 30 μg of total protein (BCA assay) was assayed with anti-GAPDH (1:1000), total Akt (1:1000), pS473 Akt (1:5000), or pT450 Akt (1:1000) in 5% BSA in TBST. In PKC transfections, Akt1/2-null HCT116 cells were maintained, prepared, transfected, and starved following the pcDNA3 plasmid transfection procedure mentioned above. However, after 18-20 hrs serum starvation, cells were either treated with DMSO control or with 500nM Torin1 for 1hr at 37°C before stimulation with 200 ng/ml of insulin and 60 ng/ml human IGF-1 for 40 minutes at 37°C. Cells were harvested, lysed, and assayed similarly.

3.5.5 Akt, PKC, and PKN Cell Signaling Assays in HEK293T cells

WT, mSin1-null, or Rictor-null HEK293T cells maintained in supplemented DMEM. For Akt and PKN1 assays, 500,000 cells per well were seeded in duplicate 6-well plates; for immunoprecipitation (IP) experiments including PKC α assays and mTORC2 complex assessment, 3 million cells were plated on duplicate 10 cm dishes. The next day, cells were transfected using PEI-MAX 40k at a 3:1 ratio (w/w) to plasmid, with boiled salmon sperm DNA normalization; on 6-well wells, with 750 ng PKN1 plasmids, 500 ng mSin1 plasmids, 1 μ g Rictor plasmids, and/or control plasmids in indicated combinations; on 10-cm dishes, with 500 ng PKC α plasmids, 2.5 μ g mSin1 plasmids, 5 μ g Rictor plasmids, and/or control plasmids in the indicated combinations. Two days later cells were lysed in 200 μ L (6-well) or 800 μ L (10 cm) CHAPS lysis buffer (50 mM HEPES 7.5, 10 mM B-glycerol phosphate, 10 mM Sodium Pyrophosphate, 1% CHAPS, 150 mM NaCl, 1 mM TCEP, supplemented with protease inhibitor and PhosSTOP tablets, incubated on ice for 5 minutes before cells were scraped for harvest and cleared by centrifugation at 21,000 \times g for 10 minutes at 4°C. For IP, magnetic beads bound to antibody recognizing the Flag epitope tag were prepared in-house by coupling Dynabeads M-270 Epoxy (Thermo Fisher Scientific) to Flag M2 antibody (Millipore Sigma) as described⁶¹. Cleared lysates were added to 25 μ L washed HA 1 micron magnetic beads (Pearce) or 10 μ L Flag M2 Dynabeads and bound for 30 min at 4°C, separated and washed 3 times in 1 ml lysis buffer, and eluted by shaking for 10 min at 70°C in 50 μ L nonreducing 2x LDS loading dye (Thermo Fisher Scientific), separated, reduced with TCEP (20 mM final), and boiled for 2 min at 95°C.

3.5.6 Generation of cells with loss-of-function mutations of mTORC2 components Rictor or mSin1

Akt1/2-null HCT116 cells or HEK293T cells were transiently transfected with plasmids con-

taining guides targeting key coding exons of Rictor or mSin1 and encoding double-NLS Cas9-T2A-GFP and an optimized sgRNA scaffold⁶², viable GFPdim single cells were sorted into individual wells and expanded. Clones were assayed for loss of mSin1 and Rictor expression and Akt S473 phosphorylation by western blotting, followed by restoration of expression by addback of the missing genes; knockout was confirmed by NGS and loss of the plasmid demonstrated by fluorescence microscopy. The best guide for mSin1 (pMT906) targets exon 2, residue 59, sequence ACACATTCCCGTGTCATCAC. The best guide for Rictor (pMT921) targets exon 8, residue 204, sequence GGTGTTTAGTCCACCTCGAA.

3.5.7 Molecular Cloning and Plasmids

Flag-PKN1 (WT and S916A) expression plasmids were gifts from Bryce Paschal. cDNAs containing human Akt1 were in pcDNA3 for mammalian expression (Addgene, # 9021) or with C-terminal MxeIntein-CBD in pFastbac1 as described¹². To generate point mutants and truncations of cDNAs for structural validation, site-directed mutagenesis reactions were performed using a two-stage QuikChange protocol⁶³ using PfuUltra-II HS polymerase, by the Q5 Site-Directed Mutagenesis strategy (NEB), or by multichange isothermal assembly⁶⁴. Constructs for bacterial expression were cloned into pFloat.SUMO⁶⁵, with N-terminal His10-SUMO tag. mTORC2 components for mammalian expression of the complex were cloned into pDarmo.CMVT, with mTOR and Rictor in individual plasmids and mLST8 and mSin1 (variant 1.1, full length) assembled into two cassettes of a single plasmid⁶⁵. To express mTORC2 in insect cells, a simplified insect vector based on the MultiBac system⁶⁶ was developed called pDarmo.polH2.1 and all mTORC2 components were assembled using this system to generate a single 20.8 kB expression vector. pDarmo.polH2.1 contains minimal sequence required for baculovirus recombination, a Golden Gate acceptor site⁶⁷ for assembling multiple expression cassettes, and an expression cassette containing [5'Cis enhancer, polyhedrin promoter, ccdB death cassette, and polyhedrin polyadenylation signal] flanked by synthetic DNA barcodes. cDNAs for mTOR, Rictor, mSin1, and mLST8 were cloned into

pDarmo.polH2.1, and then Rictor, mSin1, and mLST8 (n-1) cassettes were amplified using the barcodes with primers containing ends for directional Golden Gate assembly (NEB) and assembled into the mTOR-containing vector to create the final plasmid and verified by next-generation sequencing (NGS). To generate mTORC1-expressing baculoviruses, mLST8 was first cloned into pLIB⁶⁸, and then sequential expression cassettes from pDARMO.polH2.1 with appropriately tagged mTOR and RAPTOR were cloned into the SpeI and AvrII sites, respectively, and verified, yielding 18.8 kB vectors. mTORC2/1 plasmids were recombined into DH10EMBacY (Geneva Biotech) and validated by PCR. All plasmids are listed in Supplementary Table 2 and all plasmids developed in this study are deposited in Addgene.

3.5.8 Bacterial Expression and Purification of mSin1 fragments and 4E-BP1

N-His10-SUMO tagged mSin1 protein fragments and 4E-BP1 proteins were transformed into in E.Coli LOBSTR⁶⁹ containing the Rosetta tRNA plasmid (Novagen), grown in Teriffic Broth in UltraYield Flasks (Thompson) at 220 rpm at 37°C with 1:5000 Antifoam 204 (Sigma) and 25 μ g/ml kanamycin to OD of 2-3, cooled on ice for 30-60 minutes, induced with 0.2 mM IPTG (Isopropyl β -D-1-thiogalactopyranoside), and grown overnight at 16°C. Pellets were resuspended in 4 volumes of lysis/wash buffer (50 mM HEPES, pH 8.0, 1 M NaCl, 0.5 mM TCEP, 0.1% Tween-20, 50 mM Imidazole), lysed with 2 passes through a Microfluidizer (Microfluidics) or French Press (Thermo), clarified by centrifugation, bound to Ni-NTA agarose (MCLAB), and cleaved overnight off the resin at 4°C using purified GST-Ulp1-R3 protease⁷⁰. Eluted proteins were concentrated and purified by gel filtration on a Superdex 200 column in a buffer containing 10 mM HEPES 7.4, 500 mM NaCl, and 0.5 mM TCEP.

3.5.9 mTORC2 Expression in Suspension HEK-293T Cells

HEK-293T cells were grown in suspension and log-phase cells at 3-4 million/ml were transfected using 1.1 mg/L cDNA as follows. Three plasmids expressing His-TEV-mTOR, Rictor-3C-3xFlag, and untagged mLST/mSin1.1 were mixed at a 1.5:2:1 molar ratio, respectively. cDNAs and 4.4 mg/L PEI-Max 40k (Polysciences) were separately diluted in 1/50th final culture volume unsupplemented DMEM (Thermo), mixed, incubated 15 minutes at RT, and added to cells. 1 day after transfection, valproic acid was added to 4 mM final⁷¹, and cells were harvested 3 days after transfection.

3.5.10 mTORC2 Expression in Insect Cells

“Quad” baculoviruses expressing all four components of mTORC2 on a single virus were generated with either Flag-TEV-mTOR, Rictor-3C-3xFlag, or Rictor-3C-Flag tags by transfection of adherent SF9 cells using Cellfectin II and subsequent amplification to generate P3 viruses according to the manufacturer’s instructions. 10-40 ml P3 virus per liter was added to SF9 cells at 3-4 million per mL and cells were harvested 72 hours after expression. For experiments with mutant mTORC2, wild type and mutant mTORC2s were generated by transient transfection of ExpiSF cells in suspension using the ExpiSF system (Thermo) according to the manufacturer’s instructions.

3.5.11 mTORC2 Purification

Screening of tag combinations revealed better complex stoichiometry with initial purification using Rictor with C-terminal 3xFlag tag. HEK293T- or SF9-expressed mTORC2 were purified similarly, except that for HEK-expressed mTORC2, glutamate and arginine were not included in the buffers. Cell pellets were resuspended in 4 volumes lysis/wash buffer containing 200 mM NaCl, 50 mM Bicine pH 8.5, 50 mM glutamate, 50 mM arginine, 2 mM MgCl₂, and 0.5 mM TCEP, supplemented with 1 protease inhibitor tablet per 100 ml and 30 U Uni-

versal Nuclease (Thermo) per ml, homogenized with 3 strokes in a Dounce homogenizer, and lysed with 2 passes through a Microfluidizer. Lysates were clarified at $40,000 \times g$ for 30 min, filtered through 1 micron glass fiber filters (Pall), and then bound to Flag M2 resin (1 ml per 12 g starting cell pellet) for 90 minutes. Resin was collected by centrifugation at $500 \times g$ for 5 min, transferred to a column, and washed using 10 column volumes (CV) wash buffer (lysis buffer with one protease inhibitor tab per 250 ml), 10 CV wash buffer supplemented with 5 mM ATP (Chem-Impex) and 10 mM $MgCl_2$, and 10 CV wash buffer, then eluted in 5 CV total wash buffer containing 0.25 mg/ml 3xFlag peptide (MIT Biopolymers Lab) and 5 U/ml Universal Nuclease in 3 rounds of 30 min each. Eluted mTORC2 was concentrated in pre-rinsed 100k centrifugal devices (Millipore Sigma) and purified by size exclusion on a TSKgel G4000SWXL silica matrix column²³ (Tosoh) in SEC buffer containing 150 mM NaCl, 10 mM HEPES pH 7.5, 50 mM glutamate, 50 mM arginine, 0.5 mM TCEP, 0.5 mM $MgCl_2$, and 0.1% CHAPS (Anatrace). Fractions containing monodisperse mTORC2 (~8.5 ml) were concentrated to 3-15 mg/ml and used directly for electron microscopy or aliquoted, snap-frozen, and stored at $-80^\circ C$. For experiments with Akt-Torin, $MgCl_2$ was omitted from the SEC buffer. For some experiments with Akt-ATP, $MnCl_2$ was used in place of $MgCl_2$.

3.5.12 mTORC1 Expression and Purification

“Triple” baculoviruses expressing all three components of mTORC1, mTOR, RAPTOR, and mLST8, with tag as either RAPTOR-3C-3xFlag or Flag-TEV-mTOR, were expressed and mTORC1 purified from SF9 insect cells analogously to mTORC2 as above. mTORC1 with RAPTOR-3C-3xFlag was used for kinetic experiments.

3.5.13 Small Uni-lamellar Vesicle (SUV) Preparation

Control (PC:PE,1:1) vesicles were prepared with 50% (by mass) 18:1 ($\Delta 9$ -Cis) PC (DOPC, Avanti) and 50% 18:1 ($\Delta 9$ -Cis) PE (DOPE, Avanti). PIP3 (PC:PE:PS:PIP3,30:50:19:1) vesicles⁷² were made with 50% DOPE, 30% 18:1 DOPC, 19% 18:1 PS (DOPS, Avanti) and

1% 18:1 PI(3,4,5)P3 (Avanti). Individual lipid stocks were made by dissolving powder in chloroform to make 10 mg/mL solutions, which were added to a round bottom flask to constitute control and PIP3 vesicle mixtures at above-stated proportions. Chloroform was removed by Rotovap and overnight high vacuum. Dried lipid film was resuspended in 50 mM HEPES to make aqueous vesicle solutions at 13 mM final lipid concentration. Vesicle solutions were thoroughly mixed by vortexing for 30 seconds every 10 minutes in 1 hour. Vesicle mixtures were freeze-thawed for five rounds using dry ice-methanol bath followed by 42 °C water bath. Lipid mixtures were then sonicated for 10 minutes before being passed 20 times through a 100 nm polycarbonate filter (Avanti) using a mini-extruder (Avanti). The final vesicle solutions were stored at 4 °C for up to seven days or until use.

3.5.14 Ultracentrifugation Vesicle Binding Assay

Control and PIP3 small uni-lamellar vesicles (SUVs) were incubated with full-length Akt or Δ PH Akt for 30 minutes at 25°C. As control, SUVs were simultaneously incubated with red fluorescent protein mScarlet in pulldown buffer (50 mM HEPES, 0.2 mg/mL BSA, 5 mM DTT, 1mM PMSF, 1x Roche protease inhibitor tablet). The vesicles and bound proteins were pelleted in thick-wall polycarbonate centrifuge tubes (Beckman Coulter) at 55,000 rpm using a TLA 120.1 rotor (Beckman, $\sim 130,000 \times g$) for 1.5 hours. Supernatant was carefully removed from vesicle pellets and then boiled in 10% SDS loading dye and run on SDS-PAGE. Pellets were washed once with 100 μ L pulldown buffer and boiled in 10% SDS loading dye.

3.5.15 *In Vitro* Akt Phosphorylation Assay with mTORC2 and mTORC1

In vitro kinase assays were carried out in 1.5 mL microcentrifuge tubes in a 20 μ L reaction mixture containing 50 mM HEPES pH 7.5, 10 mM MgCl₂, 5 mM DTT, 1 mM PMSF, 0.2 mg/mL BSA, 1x Roche Protease Inhibitor tablet (kinase reaction buffer), 1 mM ATP, 1 μ M

of Akt substrate and 5 nM mTORC1 or mTORC2, unless otherwise specified. Reactions were initiated by the addition of Akt and were performed at 30°C for 10 min (unless otherwise stated). The kinase reactions were quenched by adding 13 μ L quench solution (8 μ L of 4x SDS loading buffer and 5 μ L of 500 mM EDTA) and boiled for 5 min before loading 50 ng Akt on a 4-20% Tris-Glycine Gel. In addition to the reaction samples, standard semisynthetic pS473 Akt of known concentrations were also loaded on the same gel to create a calibration curve for data analyses. Samples separated on the gel were transferred onto a nitrocellulose membrane using iBlot2 (Thermo) and assayed with total Akt (1:5000) or pS473 Akt (1:5000) in 5% BSA in TBST. Western blot images were taken and processed using a CCD camera. Band intensities (area under the curve) were calculated with ImageJ. Calibration curves were constructed using the pS473 Akt standards.

3.5.16 *In Vitro* Akt Phosphorylation Assay with γ -³²P-ATP

In vitro kinase assays were carried out in 1.5 mL microcentrifuge tubes in a 20 μ L reaction mixture containing 50 mM HEPES pH 7.5, 10 mM MgCl₂, 5 mM DTT, 1 mM PMSF, 0.2 mg/mL BSA, 1x Roche Protease Inhibitor tablet (kinase reaction buffer), 1 mM cold ATP, 1 μ Ci γ -³²P-ATP, 5 μ M of Akt substrate and 10 nM mTORC2. Reactions were initiated by the addition of Akt and were performed at 30°C for 10 min. The kinase reactions were quenched by adding 13 μ L quench solution (8 μ L of 4x SDS loading buffer and 5 μ L of 500 mM EDTA) and boiled for 5 min before loading 1 μ g Akt on a 4-20% Tris-Glycine gel (Fisher Scientific). The gel was rinsed in ddH₂O, Coomassie stained and washed again in ddH₂O. It was then laid out on a sheet of filter paper, covered with Saran wrap, and dried under mild heat and vacuum for 2 hours in the Model 583 gel dryer (Bio-Rad). A sheet of BAS-IP storage phosphor screen (GE Healthcare) was photobleached for 2 hours on a white light transilluminator (Fisher Scientific). The phosphor screen was then placed face down on top of the dried gel in a BAS cassette (Fujifilm) for overnight development in the dark. The phosphor screen was scanned the next day for radioactivity read out in a Typhoon

Biomolecular Imager (Amersham).

3.5.17 *In Vitro* 4E-BP1 Phosphorylation Assay with mTORC1 and mTORC2

In vitro kinase assays were carried out in 1.5 mL microcentrifuge tubes in a 20 μ L reaction mixture containing 25 mM HEPES pH 7.4, 10 mM MgCl₂, 0.5 mM TCEP, 100 mM NaCl, 1 mM ATP, 1 μ M of 4E-BP1 substrate, and 5 nM mTORC1 or mTORC2. Enzyme calculations are for the protomer (monomeric) unit; in other words, 5 nM mTORC2 contains 5 nM total mTOR. Reactions were initiated by the addition of ATP and were performed at 30°C for 10 min (unless otherwise stated). The kinase reactions were quenched by adding 6 μ L of 4x SDS loading buffer and boiled for 5 min before loading 50 ng 4E-BP1 on a 4-20% Tris-Glycine Gel. In addition to the reaction samples, standard phospho-4E-BP1 concentrations were also loaded on the same gel to create a calibration curve for data analyses. Those standards were generated by reacting 1 μ M 4E-BP1 with 100 nM mTORC1 for 2 hrs at 30°C, as we tested and observed plateauing of phospho-signal with both more enzyme and longer incubation time. Samples separated on the gel were transferred onto a nitrocellulose membrane using iBlot2 (Thermo) and assayed with total 4E-BP1 (1:2000) or pT37/46 4E-BP1 (1:5000) in 5% BSA in TBST. Western blot images were taken and processed as mentioned in Akt phosphorylation assay section.

3.5.18 mTORC2-Akt reactions for mass spectrometry analysis

Triplicate reactions were set up with 10 μ M Akt, 50 mM HEPES pH7.5, 10 mM MgCl₂, 2.5 mM DTT, 5 mM ATP, 1 mM EGTA in 40 μ L total, and initiated by adding 50 nM mTORC2; mTORC2 was omitted from control reactions. Reactions were incubated for 2 hours at 30°C and quenched by adding an equal volume of 10% SDS. Samples were reduced, alkylated, acidified, and processed for mass spectrometry using S-Trap Micro columns (Pro-

tiFi) with digestion using Sequencing Grade Modified Trypsin (Promega) according to the manufacturer's instructions with the exception that Trypsin was dissolved in 50 mM ammonium bicarbonate pH 8.0 for digestion, with elution in three sequential volumes of 40 μ L each 50 mM ammonium bicarbonate pH 8, 0.2% formic acid, 50% acetonitrile in water. Samples were then dried using a SpeedVac (Thermo). For phosphopeptide enrichment, an additional set of triplicate experimental and control reactions were next processed using Cell Signaling PTMScan phospho-enrichment IMAC Fe-NTA magnetic beads were used according to the manufacturer's instructions, with two elutions in 40 μ L water/20% trifluoroacetic acid and subsequently dried using a SpeedVac (Thermo).

3.5.19 nanoLC-MS analysis of AKT reactions

Seven standard peptides corresponding to the last 15 residues of the Akt C-tail (R466-A480) were purchased from Genscript: unmodified, pSer473, pTyr474, pSer475, pSer477, pThr479, double pSer473-pSer475. Synthetic standards and experimental tryptic peptides were reconstituted in 3% acetonitrile with 0.5% formic acid and loaded onto EVO tips (EVOSEP, Odense, Denmark) according to the manufacturer's instructions. Peptides were analyzed by nano-LC/MS using an EVOSEP HPLC pump interfaced to a timsTOF Pro2 mass spectrometer (Bruker, Billerica, MA). Peptides were resolved with a 20 SPD method with a self packed analytical column⁷³ (15 cm of 1.9 μ m Reprosil packed into 75 μ m I.D. fused silica). The mass spectrometer executed 10 cycles of PASEF MS/MS (charge 2 to 5, target intensity of 14500, intensity threshold 1750, mobility range 0.6-1.6). Active exclusion was enabled with a release time of 0.4 minutes. Data files were searched against a custom database containing AKT and mTORC2 using FragPIPE (version 20.0, with MSFragger 3.8 and Philopsher 5.0.0)⁷⁴. Search parameters specified variable phosphorylation of S, T, Y residues, variable oxidation of methionine, and fixed carbamidomethylation of cysteine. Peak areas for peptides were determined using mzStudio software version 1.3⁷⁵. To confirm site localization, reactions were analyzed by targeted EAD experiments on a 7600 ZenoTOF

mass spectrometer (ABSciex, Framingham, MA). Peptides were injected with a Waters M-Class UHPLC onto a heated (50°C) self packed column (15 cm of 1.9 μ m Reprosil C18, Dr. Maisch) and eluted with an HPLC gradient (1-28% B in 45 minutes, A=0.1% formic acid in water, B=0.1% formic acid in acetonitrile). Peptides were introduced from the top port at a flow rate of 0.5 μ L/min using an Opti-flow source equipped with a low flow needle (Sciex) operated at 4kV. EAD was performed with an electron beam current of 7500 nA, a kinetic energy of 5eV and a reaction time of 25 ms. Synthetic peptides were analyzed in a similar fashion, except that 5 μ M solutions were prepared in 30% acetonitrile, 0.1% formic acid and directly infused at a flow rate of 0.5 μ L/min into the top port of the Optiflow source. Figures were rendered in mzStudio software.

3.5.20 Peptide Synthesis

Peptides corresponding to residues 460-480 of Akt1 (CVDSERRPHFPQFSYSASGTA) and derivatives were synthesized using solid phase peptide synthesis on a Prelude automated synthesizer (Protein Technologies) using the Fmoc strategy as previously reported¹². Briefly, residues were double coupled (1.5 h for standard amino acids and 3 h for phospho-amino acids) and Fmoc groups were removed with 20% (v/v) piperidine in DMF over three 10 min cycles. Proline residues were triple coupled. The peptides were deblocked and cleaved from resin with trifluoroacetic acid: water: triisopropylsilane (95:2.5:2.5, v/v/v) for 3 hours, then precipitated with chilled diethyl ether and purified using reverse-phase C18 HPLC using a gradient of water:acetonitrile containing 0.05% trifluoroacetic acid. Pure fractions were combined, concentrated in vacuo and lyophilized. Peptide structures were confirmed using MALDI or electrospray mass spectrometry.

3.5.21 Semisynthesis of Akt Proteins

Both homogenous Akt protein substrate and Akt phospho-Ser473 standard used for mTORC2 kinase assays and quantitative western blots were produced using expressed pro-

tein ligation with Akt-intein fusions expressed in SF9 insect cells as described¹². Briefly, 5 mg of intein-mediated N-terminal Akt (2-459) thioester was reacted with 2 mM of the synthetic Ser473 or phospho-Ser473 Akt(460-480) peptides in ligation buffer (50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP, 0.5 mM PMSF, 10% glycerol) for 5 hours at RT and then maintained overnight at 4 ° C. The semisynthetic Akt proteins were purified by gel filtration on a Superdex 200 column (GE Healthcare) with Akt storage buffer (50 mM HEPES 7.5, 150 mM NaCl, 0.5 mM TCEP, 10% glycerol). The pure fractions of monodisperse protein were combined, concentrated to ~5 mg/ml, aliquoted and then stored at -80 ° C.

3.5.22 Preparation of Lambda Phosphatase Dephosphorylated Akt

10 mg Akt(2-459)-Intein-CBD (chitin binding domain) was immobilized on chitin resin, washed, and divided equally into two tubes in a buffer containing 50 mM HEPES pH 7.5, 100 mM NaCl, 2 mM DTT, 0.01% Brij 35, 1 mM MnCl₂. One was treated for two hours with 4000 units lambda protein phosphatase (New England Biolabs) and the other with no enzyme as a control. The resins were drained and thoroughly washed to remove any residual phosphatase with 40 CV wash buffer, 20 CV wash buffer supplemented with 1M NaCl, and 40 CV wash buffer. Akt (2-459) thioester was cleaved from the column overnight with 300 mM MESNA in 50 mM HEPES pH 7.5, 150mM NaCl, 10% glycerol, 1 mM PMSF, then ligated with non-phosphorylated S473 Akt (460-480) tail peptide and purified by gel filtration as above. Immunoblotting against total Akt (pan) and phospho-Akt (Thr450) antibodies confirmed complete dephosphorylation of Akt at Thr450, and there was no difference in yield of dephosphorylated monodisperse protein vs control mock-treated protein.

3.5.23 Semisynthesis of Akt-ATP Protein Bisubstrate

The strategy for site-specific Akt-ATP peptide conjugation was modified from⁷⁶ and is similar

to production of ATP-GSK3 peptide reported in¹². Briefly, Akt (460-480) peptide was synthesized with Ser473 replaced with alloc-protected diaminopropionic acid (DAP). The DAP Alloc was orthogonally deprotected with 4.4 equivalents of palladium dissolved in 20:1:0.5- chloroform:dichloromethane:acetic acid for 3 h, and then reacted with 10 equivalents of bromoacetic anhydride dissolved in 5% NMM in DMF for 3 h to generate bromo-Akt peptide. Bromo-Akt peptide was cleaved from the resin, purified using reverse phase C18 HPLC, and lyophilized as described above. 10 mg of bromo-Akt peptide was reacted with 2 equivalents of ATP-gamma-S (Lithium salt, Roche) in 100 mM NH₄HCO₃ and purified using reverse phase HPLC with basic conditions (30-55% gradient of acetonitrile to water containing 10 mM NH₄HCO₃). 2 mM of Akt-ATP peptide was ligated to 5 mg of Akt(2-459) thioester as above for 2 days at 4°C followed by purification by gel filtration on a Superdex 200 column in Akt storage buffer. Monodisperse protein was pooled, concentrated, ATP conjugation verified with anti-adenosine antibody (1:1000, Extended Data Figure), and used for further analysis.

3.5.24 Synthesis of Torin Acrylamide

Synthesis of Ethyl 4-((4-(4-(tert-butoxycarbonyl)piperazin-1-yl)-3-(trifluoromethyl)phenyl)amino) -6-chloroquinoline-3-carboxylate (Fig. 3.12, SI-1):

A suspension of ethyl 4,6-dichloroquinoline-3-carboxylate (772 mg, 2.87 mmol, 1.0 equiv) and tert-butyl 4-(4-amino-2-(trifluoromethyl)phenyl)piperazine-1-carboxylate (994 mg, 2.87 mmol, 1.0 equiv) in anhydrous dioxane (10 mL) was heated to 110 °C for 24 hours. UPLC-MS analysis showed roughly 50% conversion to the desired product; the reaction was returned to heating at 110 °C for an additional 18 hours. UPLC-MS analysis showed no additional conversion. The reaction was poured into saturated aqueous NaHCO₃ and extracted with EtOAc (3x50 mL) and 3:1 CHCl₃/iPrOH (2x50 mL). The combined organics were washed twice with water and then with brine, dried over MgSO₄, filtered, and concentrated to

provide a dark red oil. The crude material was carried forward to the next step. LRMS (ESI) calculated for C₂₈H₃₁ClF₃N₄O₄ [M+H]⁺ 579.1980, found 579.17.

Synthesis of tert-butyl 4-(4-(9-chloro-2-oxobenzo[h][1,6]naphthyridin-1(2H)-yl)-2-(trifluoromethyl)phenyl)piperazine-1-carboxylate (Fig. 3.12, SI-2):

The crude SI-1 was dissolved in absolute ethanol (20 mL) and cooled on ice. Sodium borohydride (636 mg, 17.22 mmol, 6.0 equiv) was added and the reaction was stirred overnight, slowly warming to room temperature. UPLC-MS analysis showed full conversion to the alcohol (observed M+H = 537.17). The reaction was poured into saturated aqueous NH₄Cl and extracted with EtOAc (3x50 mL). The combined organics were washed twice with water and then with brine, dried over MgSO₄, filtered, and concentrated.

The crude alcohol intermediate was dissolved in DCM (25 mL). MnO₂ (1.50 g, 17.22 mmol, 6.0 equiv) was added and the reaction was stirred at room temperature for 48 hours. The reaction was then filtered through Celite, washing with DCM, and concentrated. UPLC-MS analysis showed approximately 80% conversion to the aldehyde (observed M+H = 535.07).

The crude aldehyde was dissolved in absolute ethanol (20 mL) in a 100-mL round-bottom flask. Triethyl phosphonoacetate (1.120 mL, 5.6 mmol, 2.0 equiv) and K₂CO₃ (1.161 g, 8.4 mmol, 3.0 equiv) were added. The flask was fitted with a reflux condenser and heated to 100 °C for 16 hours. The reaction was cooled to room temperature and volatiles were removed by rotary evaporation. The residue was dissolved in EtOAc (50 mL) and poured into water. The aqueous component was extracted with EtOAc (3x30 mL). The combined organics were washed with water and brine, dried over MgSO₄, filtered, and concentrated. ISCO flash chromatography (24 g silica, 0 to 100% EtOAc/hexanes, 9 min gradient to elute impurities, followed by a solvent switch; 0 to 10% MeOH/DCM, 6 min gradient and 6 min hold) provided the title compound as a rust-colored solid (94.6 mg, 6% yield over 4 steps). LRMS (ESI) calculated for C₂₈H₂₇ClF₃N₄O₃ [M+H]⁺ 559.1718, found 559.17.

Synthesis of tert-butyl 4-(4-(2-oxo-9-(quinolin-3-yl)benzo[h][1,6]naphthyridin-1(2H)-yl)-2-(trifluoromethyl)phenyl)piperazine-1-carboxylate (Fig. 3.12, SI-3):

A suspension of SI-2 (94.6 mg, 0.17 mmol, 1.0 equiv), 3-quinolineboronic acid (58.1 mg, 0.34 mmol, 2.0 equiv), Pd(PPh₃)₂Cl₂ (12.1 mg, 0.017 mmol, 0.10 equiv), and XPhos (12.3 mg, 0.026 mmol, 0.15 equiv) in dioxane (3.4 mL) and saturated aqueous Na₂CO₃ (0.85 mL) was sparged with N₂ for 10 minutes and heated to 80 °C for 18 hours. The reaction was cooled to room temperature and filtered through Celite, washing with EtOAc and 10% MeOH/DCM. ISCO flash chromatography (12 g silica, 0 to 10% MeOH/DCM; requires a second round of purification at a 5% MeOH/DCM isocratic run) provided the title compound as an amber oil (86.8 mg, 78% yield). LRMS (ESI) calculated for C₃₇H₃₃F₃N₅O₃ [M+H]⁺ 652.2530, found 652.26.

Synthesis of 1-(4-(4-acryloylpiperazin-1-yl)-3-(trifluoromethyl)phenyl)-9-(quinolin-3-yl)benzo [h][1,6]naphthyridin-2(1H)-one (Fig. 3.12, BJG-06-035, Torin-acrylamide):

To a solution of SI-3 (86.8 mg, 0.13 mmol, 1.0 equiv) in DCM (1.0 mL), added TFA (0.25 mL). The reaction was stirred at room temperature; UPLC-MS analysis at 3 hours showed roughly 75% deprotection. An additional 0.2 mL TFA and 0.5 mL methanol were added, and the reaction was heated to 50°C for 4 hours, at which point UPLC-MS analysis showed roughly 85% deprotection. Solvents were removed in vacuo.

The crude deprotection product was dissolved in DCM (1.5 mL). Acryloyl chloride (1 M solution in DCM, 0.15 mL, 0.15 mmol, 1.2 equiv) and triethylamine (54.4 μL, 0.39 mmol, 3.0 equiv) were added dropwise. THF (0.5 mL) was added as a co-solvent. The reaction was stirred at room temperature for 90 minutes. The reaction was poured into dilute aqueous HCl and extracted with DCM (3x5 mL). The combined organics were washed with saturated aqueous NaHCO₃, water, and brine, dried over MgSO₄, filtered, and concentrated. The material was purified by ISCO flash chromatography (12 g silica, 0 to 10% MeOH/DCM,

7 min gradient) and further purified by reverse-phase HPLC (0.035% TFA, 100 to 20% H₂O/MeCN, 30 min gradient, 20 mL/min). Lyophilization from H₂O/MeCN provided the title compound as a light-yellow powder (7.1 mg, TFA salt). LC-MS analysis showed >95% purity; ¹H NMR suggests ~9:1 mixture of rotamers. ¹H NMR (500 MHz, DMSO-d₆) 9.24 (s, 1H), 8.60 (d, J = 2.4 Hz, 1H), 8.37 (d, J = 9.5 Hz, 1H), 8.30 (d, J = 2.3 Hz, 1H), 8.23 (d, J = 8.6 Hz, 1H), 8.19 (dd, J = 8.6, 1.9 Hz, 1H), 8.07 (d, J = 8.4 Hz, 1H), 8.01 (dd, J = 8.8, 1.8 Hz, 2H), 7.89 – 7.79 (m, 2H), 7.77 – 7.68 (m, 2H), 7.14 (d, J = 2.4 Hz, 1H), 6.98 (d, J = 9.4 Hz, 1H), 6.78 (dd, J = 16.6, 10.5 Hz, 1H), 6.15 (dd, J = 16.6, 2.4 Hz, 1H), 5.75 (dd, J = 10.5, 2.4 Hz, 1H), 3.65-3.42 (m, 4H), 2.81 – 2.63 (m, 4H). LRMS (ESI) calculated for C₃₅H₂₇F₃N₅O₂ [M+H]⁺ 606.2111, found 606.17.

3.5.25 Protein Cross-linking and Mass Spectrometry (XL-MS)

mTORC2 preparations (apo or pre-assembled with Akt-ATP or Akt-Torin) were purified by gel filtration, diluted to 1.2-3 μ M in HEPES buffered saline (50 mM HEPES pH 7.4, 150 mM NaCl), and cross-linked with 0.5-5 mM disuccinimidyl suberate (DSS) for 25 min or with 5-30 mM 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) and 2.5 molar equivalents of N-hydroxysulfosuccinimide (sulfo-NHS) for 60 min at 23°C with agitation at 1000 rpm and quenched with 50 mM Tris pH 8.0. Protein samples were reduced in 8M urea, 5mM TCEP, 5mM DTT, alkylated with 30mM iodoacetamide, and trypsin digested in-solution into peptides. The peptides were then desalted using C18 resin and subjected to analysis by nanoflow liquid chromatography coupled to Thermo Q Exactive HF-X mass spectrometry. All the samples were analyzed at least twice and the raw data was searched using pLink software version 2.3.9⁷⁷ with 5% FDR and manually checked to remove potential false-positive identifications essentially as described^{50,78-80}. Cross-linking results were then visualized by CX-Circos (<https://cx-circos.vercel.app/>).

3.5.26 Integrating Modeling of mTORC2-Akt

Integrative structure modeling of the mTORC2-Akt complex proceeded through five stages⁸¹: (1) gathering input information, (2) representing subunits, (3) translating input information into spatial restraints, (4) configurational sampling to produce an ensemble of structural models that satisfy input information, and (5) analyzing and validating the ensemble. The integrative structure modeling protocol (stages 2, 3, 4, and 5) was implemented using our open-source Integrative Modeling Platform (IMP) package version 2.17 (<https://integrativemodeling.org>)⁵¹. The current procedure is an updated version of previously described protocols^{79,80,82}. All input files, modeling scripts, and output files are available at <https://integrativemodeling.org/examples/mTORC2-Akt> and in the nascent wwPDB archive for integrative structures.

Stage 1: Gathering input information

The subunit structures were determined previously by X-ray crystallography, cryo-EM, and nuclear magnetic resonance (NMR) spectroscopy as well as by de novo prediction using AlphaFold2. 201 intra- and inter-molecular DSS and EDC cross-links were identified using XL-MS (Extended Data Table 1), which informed both the spatial proximities and conformations of the subunits of mTORC2 and Akt. The density map of the entire mTORC2-Akt complex was determined by single particle reconstruction cryo-EM at 3-10Å resolution, which informed the morphology of the complex.

Information about the modeled system ('Stage 1: gathering input information') can in general be used for representing the system, defining the scoring function that ranks alternative structures, limiting the structural sampling, filtering good-scoring structures, and validating the resulting model. Here, the mTORC2-Akt representation relies on X-ray structures, cryo-EM structures, and AlphaFold2 models of the subunits. The scoring function relies on chemical cross-links, the cryo-EM density map, excluded volume, and sequence connectivity. The sampling benefits from knowledge of mTORC2-Akt's symmetry as well

as spatial proximities from previous experimental structures. The model was validated, in part, through mutagenesis experiments as well as phosphorylation assays.

Stage 2: Representing subunits

To increase the efficiency of structural sampling, the subunits of mTORC2-Akt were represented using rigid bodies and flexible beads consisting of multiple residues. A rigid body consists of beads whose relative distances, defined by an experimental structure or de novo prediction, are constrained during conformational sampling. A rigid body is represented in a multiscale fashion, by both 1-residue and 10-residue beads. A 1-residue bead represents the position of a corresponding C α atom. A 10-residue bead represents the center of mass of up to 10 residues. The remaining $\sim 27\%$ of residues in subunits not in rigid bodies were represented using flexible strings of 10-residue beads.

Representations of mTOR, Rictor, mLST8, and mSin1 N-terminus (1-157) relied primarily on previous atomic cryo-EM structures of apo mTORC2 (PDB ID: 6ZWM). Representation of mSin1 Crim (158-279) relied on an atomic NMR structure (PDB ID: 2RVK). Representation of mSin1 Rbd (280-380) and mSin1 PH domain (381-522) relied on a crystal structure (PDB ID: 7LC1). Representation of the Akt PH domain (1-144) relied on high-confidence regions (pLDDT > 60) of the AlphaFold2 model. Representation of the Akt Kinase Domain C-lobe (145-231) and Akt1 Kinase Domain N-lobe (232-480) relied on the crystal structure (PDB ID: 3o96).

Stage 3: Translating input information into spatial restraints

After specifying the model representation, we defined a differentiable pseudo-Bayesian scoring function based on a subset of the gathered input information (Stage 1). The scoring function was used to quantify the consistency of a model with the input information. The scoring function consisted of the following spatial restraints.

- (1) Cross-link restraint: A spatial restraint based on each of the 200 DSS and 20 EDC

cross-links is the Bayesian data-likelihood for the cross-link observation given an evaluated structure. The restraint is similar to an upper bound on the distance between the pair of cross-linked residues. Additionally, the restraint takes into account the assignment ambiguity for cross-links involving different residues of the same subunit type, which can be intra- or inter-molecular, by considering both assignments, with only the least violated distance contributing to the score.

(2) Cryo-EM density restraint: A spatial restraint based on the cryo-EM density map is the correlation coefficient between the Gaussian Mixture Model (GMM) representations of the map and an evaluated structure⁸³. To reduce the cost of the restraint's evaluation, we omitted high-resolution regions of the map for which a model was unambiguously built (mTOR, Rictor, mLST8). The arch region of the map with local resolution ~ 10 Å was fit by 100 Gaussians, while the N-terminal lobe of Akt, C-terminal lobe of Akt, and mSin1 Crim domain were fit with 1 Gaussian per 5 residues, resulting in 76 model Gaussians for these 3 rigid bodies⁸⁴.

(3) Excluded volume restraint: A spatial restraint based on the excluded volume is a lower distance bound on all pairs of 10-residue beads. The radius of a bead is determined by its volume, which is estimated from the number of residues in the bead and the statistical relationship observed from a large number of protein structures. To accommodate attempting fitting rigid bodies into a noisy cryo-EM density map, we removed the excluded volume restraint between Akt KD N-lobe and Akt KD C-lobe/mSin1 Crim.

(4) Sequence connectivity restraint: A spatial restraint based on sequence connectivity is an upper harmonic distance bound on the distance between two consecutive beads in a single subunit, with the mean distance equal to twice the sum of the radii of the two connected beads.

Stage 4: Sampling structures to produce an ensemble of structures that satisfy input information

To maximize the precision at which the sampling of good-scoring structures is exhaustive, we improved the efficiency of structural sampling in the following two ways. First, we reduced the number of degrees of freedom by considering the known spatial proximities of mTORC2 subunits. The configuration of the mTOR, Rictor, mLST8, and mSin1 N-terminus rigid bodies was constrained to that in the input cryo-EM structure. Second, we reduced the number of degrees of freedom by considering the C2 symmetry of mTORC2-Akt. The C2 symmetry was applied to the system around the z -axis, resulting in a symmetry unit consisting of one copy of mTOR, Rictor, mLST8, mSin1 N-terminus, mSin1 Crim, mSin1 Rbd, mSin1 PH, Akt PH, Akt KD N-lobe, and Akt KD C-lobe; as a result, we only sampled the structure of a single symmetry unit, although the scoring function evaluated the model of the entire complex including the interface between the symmetry units.

The starting positions and orientations of all rigid bodies and the starting positions of all beads were assigned randomly. The positions of all beads were first optimized via 100 steps of conjugate gradients. The positions and orientations of the rigid bodies as well as the conformations of the flexible strings of beads were then sampled using Replica Exchange Monte Carlo with 8 replicas with a temperature range between 1.0 and 2.5. Each step of sampling proposed 20 Monte Carlo moves including random translation and rotation of rigid bodies (up to 5 Å and 5 radians, respectively) and random translation of 10-residue beads in the flexible strings of beads (up to 5 Å). The acceptance of a Monte Carlo Move is conditioned on the scoring function (Stage 3). 100 independent runs of sampling were performed for 50,000 steps. The sampling produced a total of 5,000,000 models from 100 independent runs, requiring ~4 days on 800 computational cores. For the most detailed specification of the sampling procedure, see the IMP modeling script (<https://integrativemodeling.org/examples/mTORC2-Akt>).

Stage 5: Analyzing and validating the ensemble

Input information and obtained structures need to be analyzed to estimate the precision and accuracy of the model, detect potentially inconsistent and missing information, and suggest future experiments. We used an analysis and validation protocol published earlier, as follows. First, we estimated the precision at which sampling is exhaustive. Second, we clustered good-scoring structures and estimated the model precision from their variability. Third, we quantified the fit of the model to input information that was used to construct the model. Fourth, we assessed the fit of the model to data that was not used to compute it.

A previously developed test estimated the precision at which the stochastic sampling of our mTORC2-Akt model is exhaustive (sampling precision) to be ~ 14 Å. The sampling precision provides a lower bound on the size of interpretable model features.

We obtained a sample of good-scoring mTORC2-Akt structures by performing density-based clustering on the score of all sampled structures from which we filtered 2000 of the best-scoring structures. The structures were clustered based on root-mean-square fluctuations (RMSF). Over 80% (1616/2000) of the structures clustered in a single cluster (RMSF) to yield the final ensemble of good-scoring structures (the ensemble). The precision of the ensemble is defined by the structural variability (quantified by RMSD) among the constituent structures. The precision of the ensemble is ~ 12 Å, close to the estimate of sampling precision and sufficient for characterizing the molecular architecture of the system. The local fluctuations, as quantified by RMSF, vary in different parts of the model (SI Figure 5), and are particularly high in the flexible regions. We were encouraged to see, however, that spatial proximities of Akt are relatively consistent between ensemble structures, with the PH domain presenting the most significant fluctuations.

The mTORC2-Akt model sufficiently satisfies all input information that was used to construct the model. We consider a cross-link as satisfied by the ensemble if the cross-linked distance in any individual structure in the ensemble is less than the maximum threshold

(16 Å for EDC and 35 Å for DSS). The mTORC2-Akt ensemble satisfies 90% (181/201) of the input cross-links of the inter-subunit and intra-subunit cross-links (**Fig. 3.21a**). 61% (11/18) of the unsatisfied cross-links can be attributed to the Akt-ATP bisubstrate binding and cross-linking to the Rictor ATP binding pocket ('A site'), ~80 Å from the active site (**Fig. 3.11**). The remaining unsatisfied cross-links may be attributed to false positives in data collection, the complex existing in multiple states in solution, and the finite structural sampling during modeling.

The localization densities for the dominant cluster overlap well with the cryo-EM density map. The cross-correlation coefficient of the ensemble density with the half of the map containing substrate density is 0.85 (**Fig. 3.20c**). The cross-correlation is negatively impacted by areas of the model lacking any density in the cryo-EM map, particularly the entirely absent mSin1 and Akt domains and flexible regions of mTOR and Rictor. For visualization, the localization densities are typically contoured at the threshold that results in approximately twice the protein volume estimated from its sequence (**Fig. 3.20b**).

The remaining connectivity and excluded volume restraints are harmonic, with a specified force constant related to the standard deviation of the corresponding Gaussian distribution. A restraint is satisfied by the mTORC2-Akt ensemble if the restrained distance in any structure in the ensemble (considering restraint ambiguity) is violated by less than 3 standard deviations. The connectivity restraints were either satisfied or their violations can be rationalized without the need to change the interpreted features of the model, as follows. First, a quarter of the ensemble structures violated connectivity restraints between the 4 beads spanning mTOR residues 1222-1261. These persistent violations are attributed to the distance between L1222 and S1261 in the cryo-EM mTOR model being too long to be spanned by coarse-grained beads that assume a Van der Waals sphere based on statistics computed from globular proteins. Second, the vast majority of the structures (99.9%) satisfy each remaining connectivity restraint. These violations can be rationalized by the coarse-graining and limited structural sampling of the flexible and rigid regions. Similarly, the excluded

volume restraints were also satisfied by the ensemble. The vast majority of the structures (99.9%) satisfy all excluded volume restraints.

The model sufficiently satisfies input information that was not used to construct it, as follows. First, our model is consistent with the orientation of Akt of the manually built EM structure (**Fig. 4.3**, **Fig. 4.4**). Our model is consistent with the mutagenesis experiments at the interface of mTORC2 and the N-terminal lobe of Akt Kinase domain. The Akt triple mutant KKE182-184AAA and adjacent T219A significantly disrupt mTOR phosphorylation of Akt (**Fig. 4.3g**, **Fig. 4.4d**), suggesting a coordinated role of these residues in the interface of the N-terminal lobe of the Akt Kinase domain with mTORC2. Our model is consistent with the role of these residues in the recognition and engagement of Akt as the Cα positions of these charged residues are in direct spatial proximity of potentially constituent negatively charged residues in the mTORC2 “acidic loop” (**Fig. 4.3a**). In concert with the EM structure, our model identified a secondary interface between the C-lobe of the Akt KD and the disordered portions of Rictor/mSin1. We validated this interface with the A376G and K386A mutations, which similarly ablated phosphorylation of Akt (**Fig. 4.4d**).

3.5.27 mTORC2 and mTORC2:Akt co-Complex Preparation for cryo-EM

For mTORC2:Akt-Torin co-complex, purified mTORC2 was mixed with Akt-Torin (present with excess unligated Akt 1-459) and incubated for 60 min on ice. The co-complex was injected on either a TSKgel UP-SW Aggregate column (Tosoh) or a TSKgel G4000SWXL column (Tosoh) and fractions containing monodisperse mTORC2+Akt-Torin were combined. For all mTORC2 preparations, freshly purified mTORC2 was concentrated to 3-15 mg/ml and cleared 10 minutes at $200,000 \times g$ in a TLA-100 rotor (Beckman). Supernatant concentration was re-measured by A280 and held at high concentration on ice with dilution immediately before grids were spotted. When used, Akt-ATP was added at between 1.2 and 10 molar equivalents and incubated on ice for at least 60 minutes prior to vitrification. Spec-

imens were prepared with a Vitrobot Mark IV (Thermo). 3.5 μL of 0.6-1.0 μM mTORC2, mTORC2:Akt-ATP, or mTORC2:Akt-Torin was applied to glow-discharged gold 300 square mesh Quantifoil R 1.2/1.3 holey carbon grids (Quantifoil). The grids were then blotted from both sides for 3 s at 90% humidity and plunge-frozen in liquid ethane.

The resulting mTORC2 grids were screened for favorable ice thickness and particle distribution on a Talon Arctica electron microscope (Thermo Fisher Scientific) at MIT.nano, operated at 200 kV. High-quality mTORC2 datasets (Apo, +Akt-ATP, +Akt-Torin) were collected using two different Titan Krios electron microscopes (Thermo Fisher Scientific) with Gatan K3 direct detection cameras and BioQuantum energy filters (slit width of 20 eV) at: (1) Harvard University, and (2) MIT.nano. Specific microscope settings used for data collection are listed in Extended Data Table 1. Fully automated data collection was carried out using either SerialEM34 (Harvard) or EPU (MIT.nano) operating software.

3.5.28 Cryo-EM Image Processing

Image processing was carried out as previously reported, with modification⁸². Data processing workflow of the final mTORC2 reconstitutions are illustrated in **Fig. 3.31**. Large movie datasets recorded with a Titan Krios G3i microscopes at MIT.nano (apo-mTORC2: 17,072 movies, mTORC2-Akt-Torin: 22,619 movies) were corrected for drift using MotionCor2 implementation in Relion^{83,84}, and CTF parameters were determined using GCTF⁸⁵. For each motion-corrected micrograph dataset, we applied two particle search models in Topaz: (1) a default pre-calculated model provided by the software developers, and (2) a trained model, based on a manually picked subset of 400 mTORC2 particles⁸⁶. The resulting particle sets (apo-mTORC2: (1) 17,448,682 and (2) 4,142,679; mTORC2-Akt-Torin: (1) 15,361,038 and (2) 3,631,574) were extracted and downscaled for further processing in Relion. Referenced 2D classifications (based on the preliminary 5 Å map of mTORC2) eliminated contaminating non-mTORC2 particles⁸⁷. The remaining particles were reclassified in 3D, and the best classes from Topaz sets ('default' and 'trained') were combined using a strict distance cut-off

(100 Å) to discard duplicates. The resulting clean sets of particles (apo-mTORC2: 765,077 and mTORC-Akt-Torin: 2,256,242) were used as input for a refined search model picking in Topaz. Picked particles (apo-mTORC2: 5,153,640 and mTORC2-Akt-Torin: 5,819,505) were extracted and given as input for referenced 2D classification to eliminate non-mTORC2 particles (apo-mTORC2: 1,460,017 and mTORC2-Akt-Torin: 3,192,975). The best classes in 3D classification were combined with the particles from the previous rounds and a strict distance-cutoff (100 Å) applied to discard duplicates. Overall, a total number of 1,046,398 particles were determined for apo-mTORC2 and 3,169,869 for mTORC2-Akt-Torin.

Particles were re-extracted at full size and used for initial refinement and reconstruction. Iterative cycles of CTF and aberration refinements in Relion⁸⁸ (per-particle defocus, per-micrograph astigmatism, beam tilt and higher-order aberrations) improved the maps substantially. Further alignment of the mTORC2 map to C2 symmetry and correction of per-particle motion in Relion⁸⁹ produced maps extending to a resolution of 3.52 Å for apo-mTORC2 and 2.76 Å for mTORC2-Akt-Torin. The particle sets were further refined in cryoSPARC v.3.1.0⁹⁰ using iterative cycles of non-uniform refinement and CTF refinement—until convergence. The final map obtained from these particles was estimated at 3.12 Å resolution for apo-mTORC2 and 2.55 Å for mTORC2-Akt-Torin according to a gold-standard Fourier shell correlation of 0.143 (**Fig. 3.30**, **Fig. 3.31**). Local resolution was estimated in cryoSPARC and is visualized in **Fig. 3.30**, **Fig. 3.31**.

3.5.29 Cryo-EM Model Building and Refinement

All model building tasks were performed using the C2-symmetric composite map of mTORC2. Coordinates from PDB entry 6ZWM and AlphaFold predictions were docked into the cryo-EM reconstruction with UCSF Chimera⁹¹ and rebuilt by hand using Coot⁹²; The mSin1 CRIM domain as well as the Akt N and C lobes were built independent of each other. Note that only one protomer unit of mTORC2 was built manually, and after its refinement with phenix.real_space_refine⁹³, the second protomer model was generated with C2

symmetry operators. Model refinement was carried out with restraints applied to secondary structure elements. The final model contains 8380 amino acids. Figures of the map and the final model were rendered in UCSF ChimeraX⁹⁴.

3.5.30 Data availability

The coordinates for the apo-mTORC2 and mTORC2-Akt-Torin have been deposited in the Protein Data Bank (PDB ID: 9B08 and 9B09, respectively). The single particle cryo-EM maps and composite maps have been deposited in the EMDB under accessions EMD-44036 (apo), EMD-44037 (mTORC2-Akt-Torin). The mass spectrometry proteomics data for *in vitro* phosphorylation of Akt by mTORC2 have been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository with the dataset identifier PXD048719. The proteomics data from XL-MS analysis was deposited at MassIVE: MSV000090707 and will be made publicly upon publication. Files containing the input data, scripts, results of integrative modeling are available at <https://integrativemodeling.org/examples/mTORC2-Akt> and the nascent integrative modeling section of the worldwide Protein Data Bank (wwPDB) PDB-Dev repository for integrative structures. Novel plasmids are deposited in Addgene.

Extended Data/Supplement

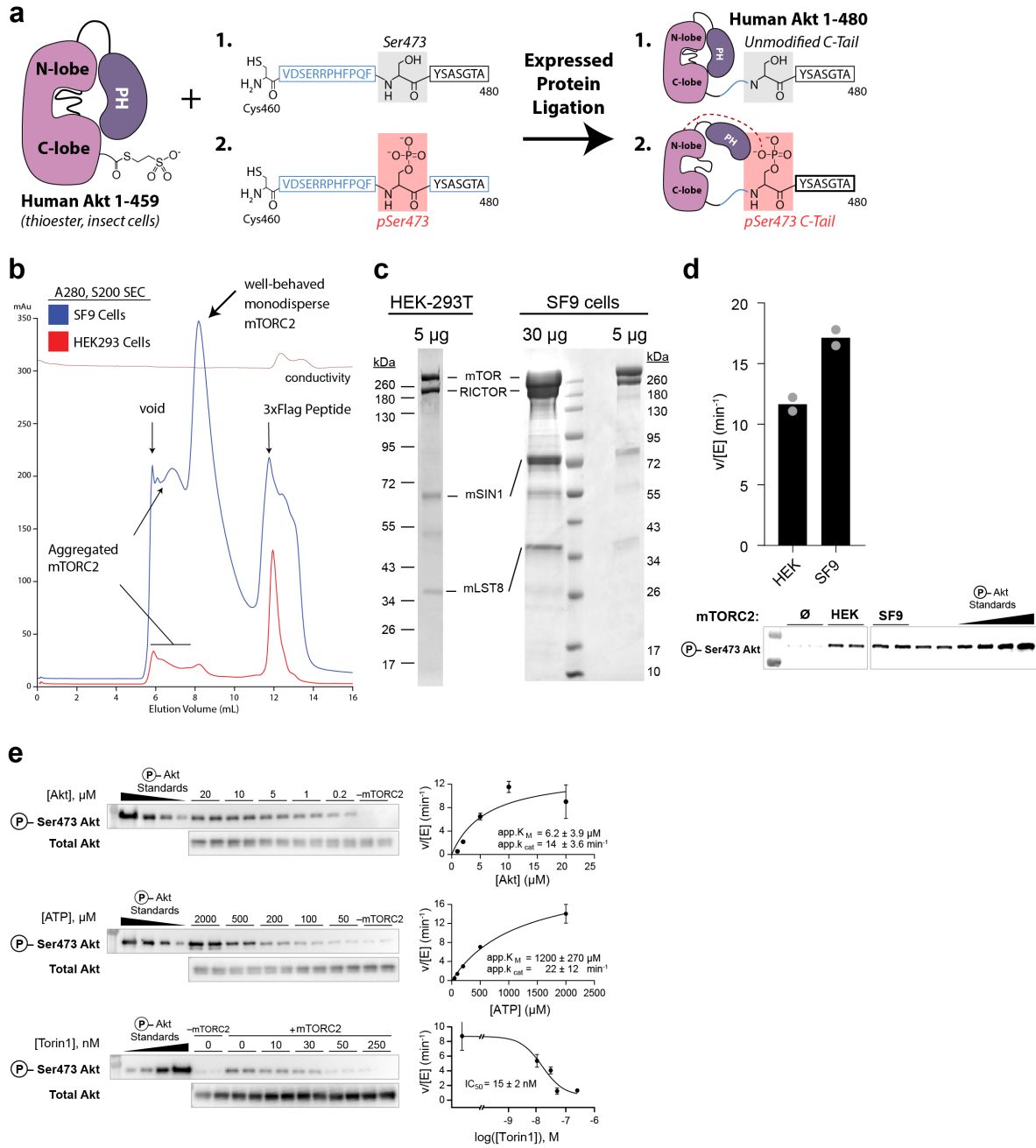
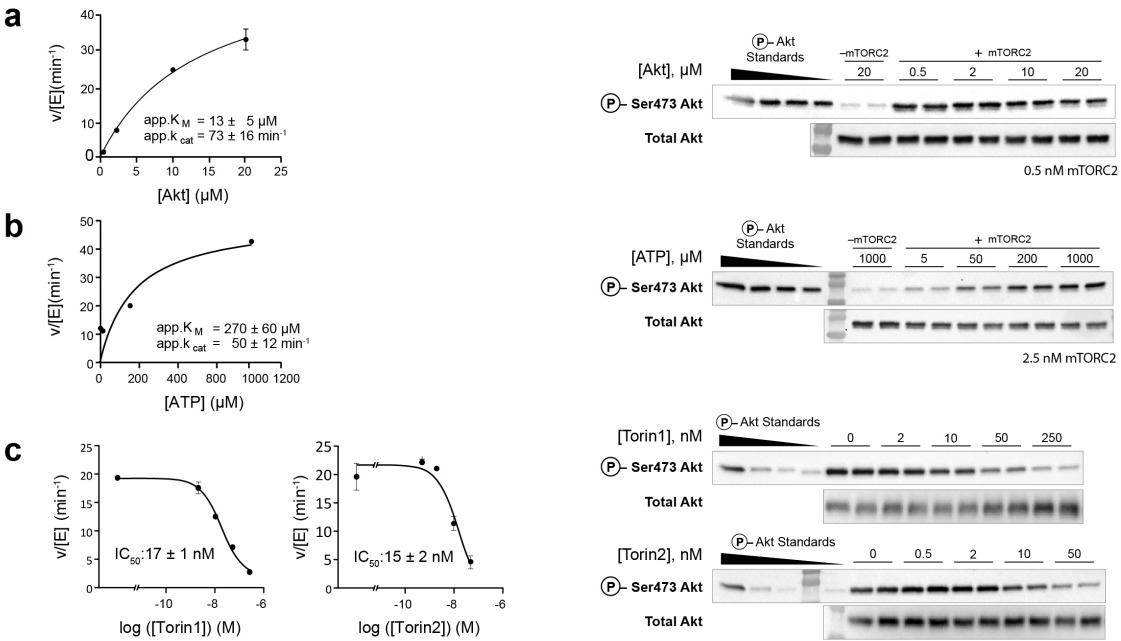


Figure 3.7

HEK-293T Expressed mTORC2



Catalytic Null Akt (D274A), "TIM" motif mutant Akt (T443A), and Akt Inhibitors

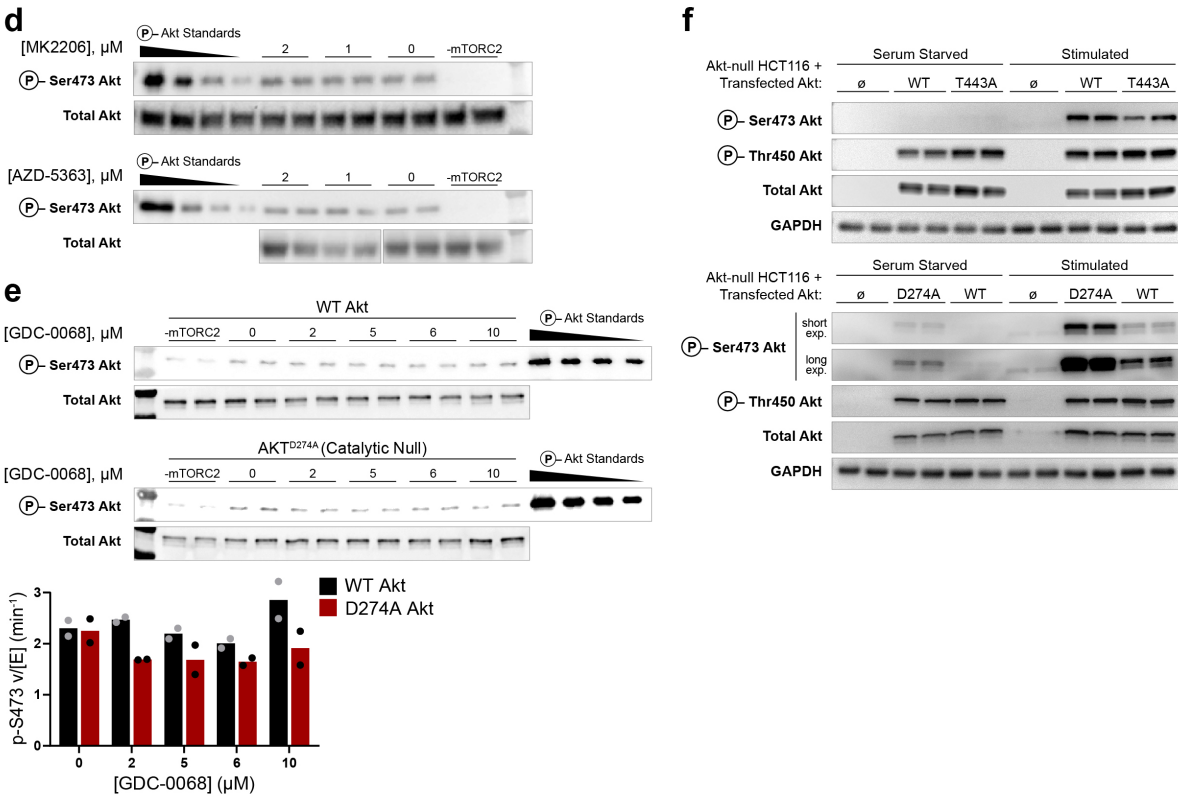


Figure 3.8

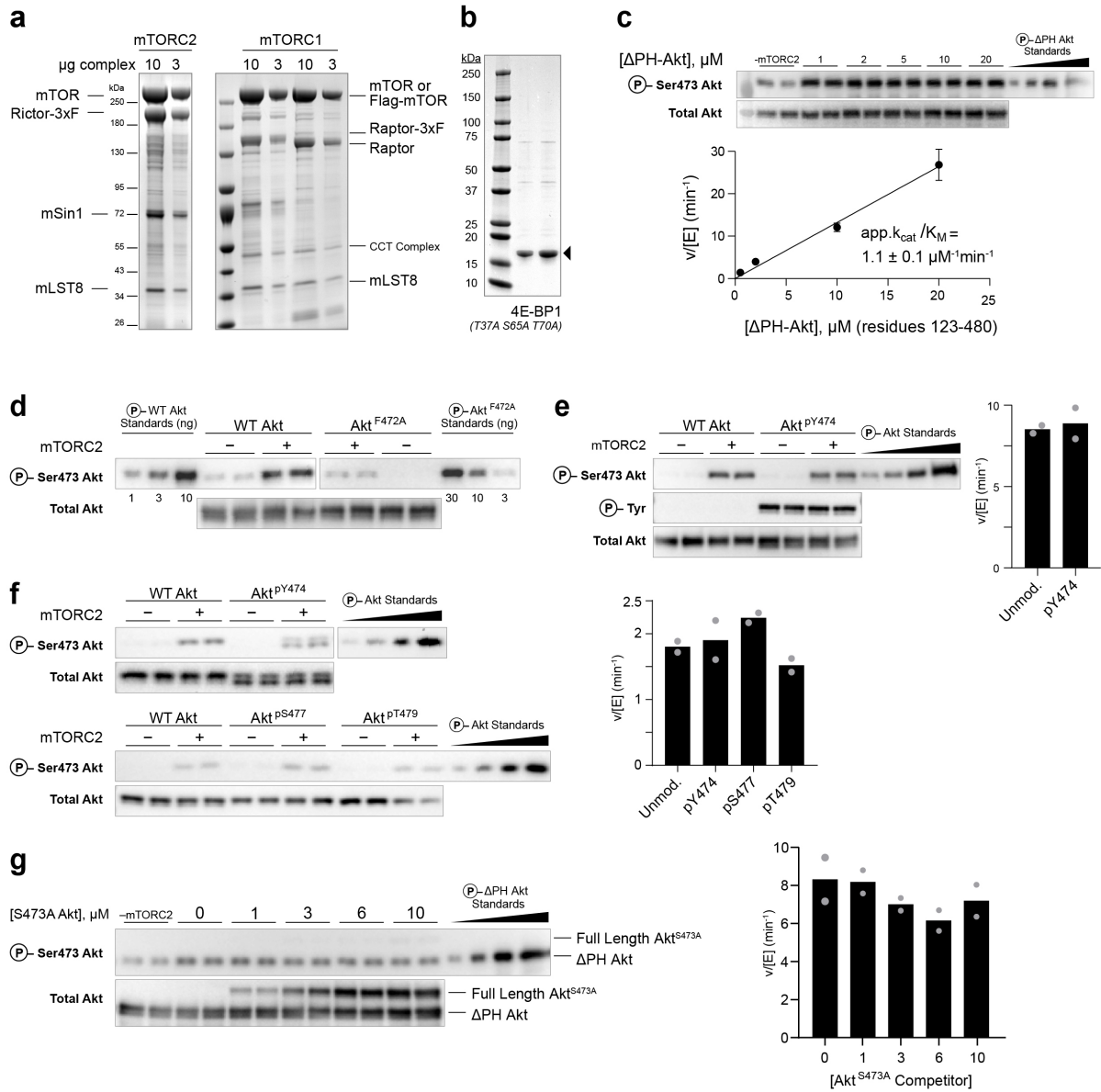


Figure 3.9

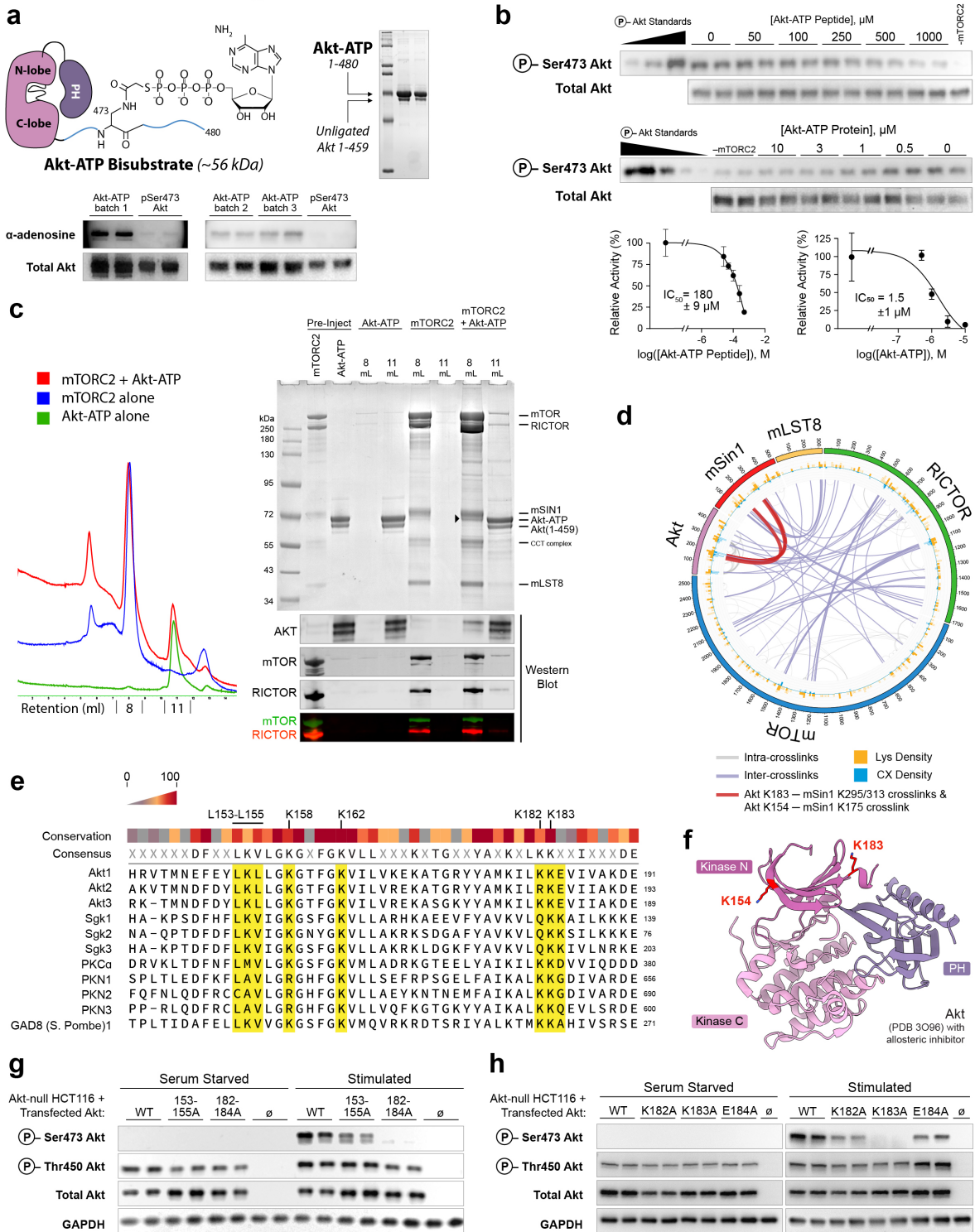


Figure 3.10

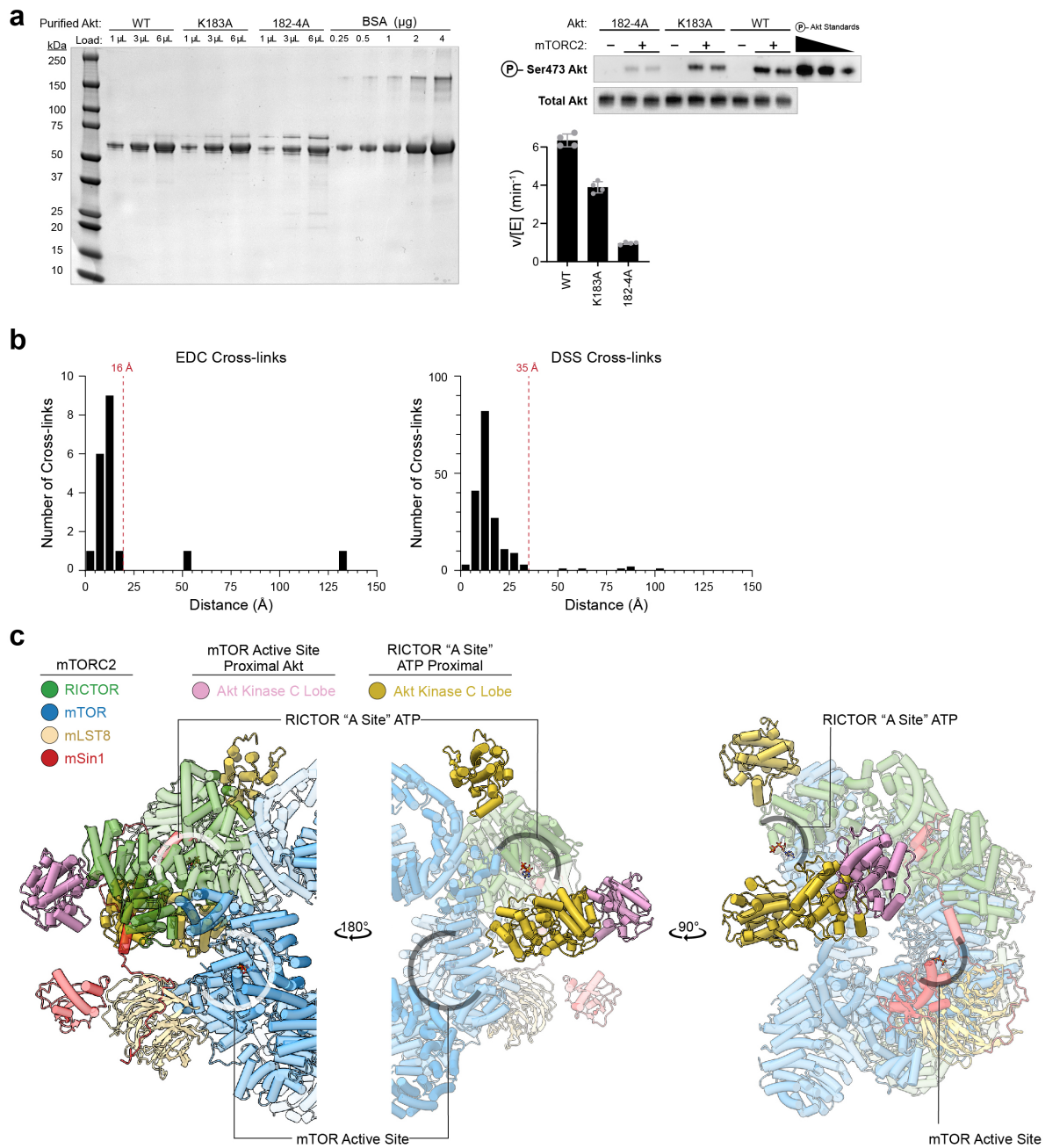


Figure 3.11

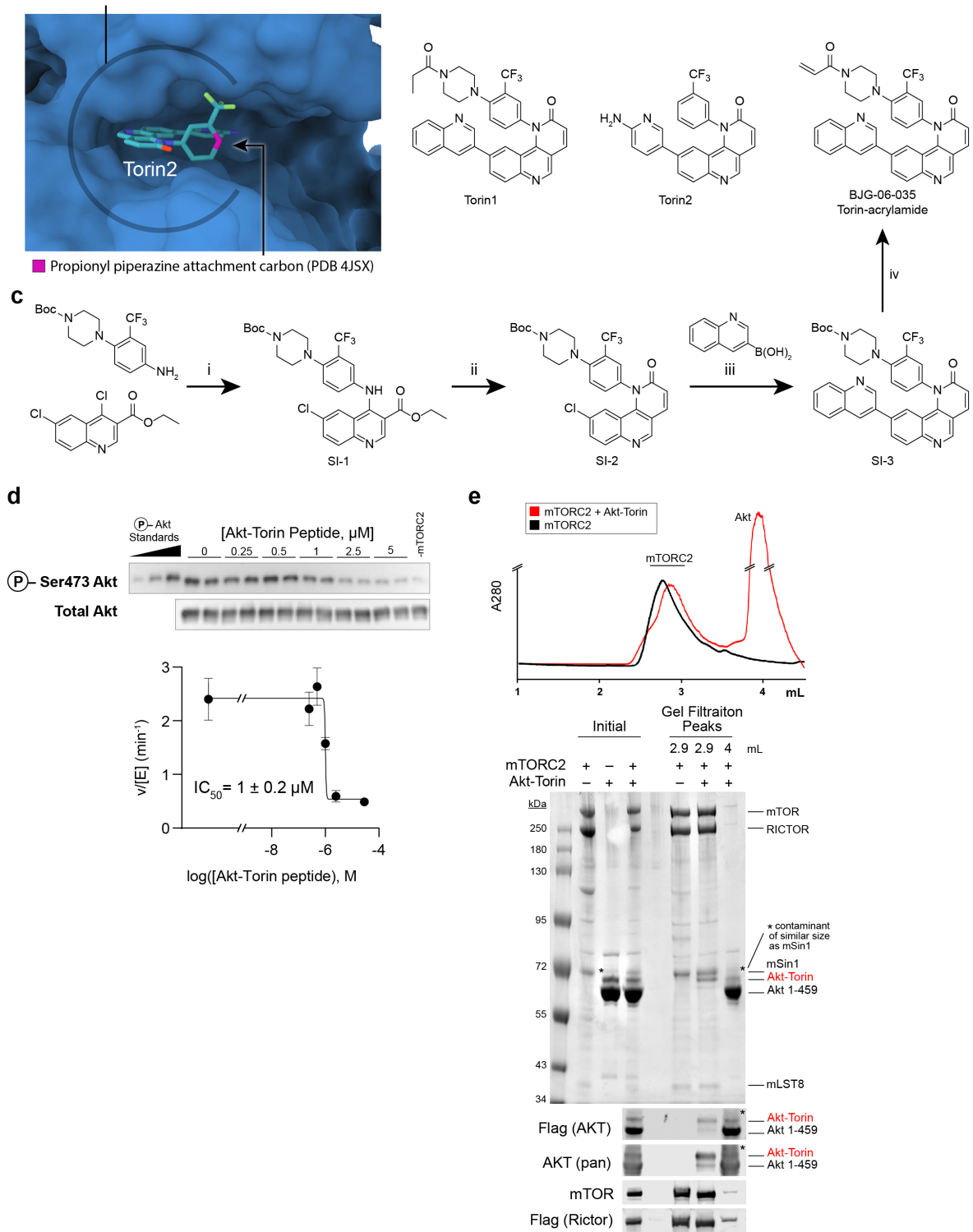


Figure 3.12

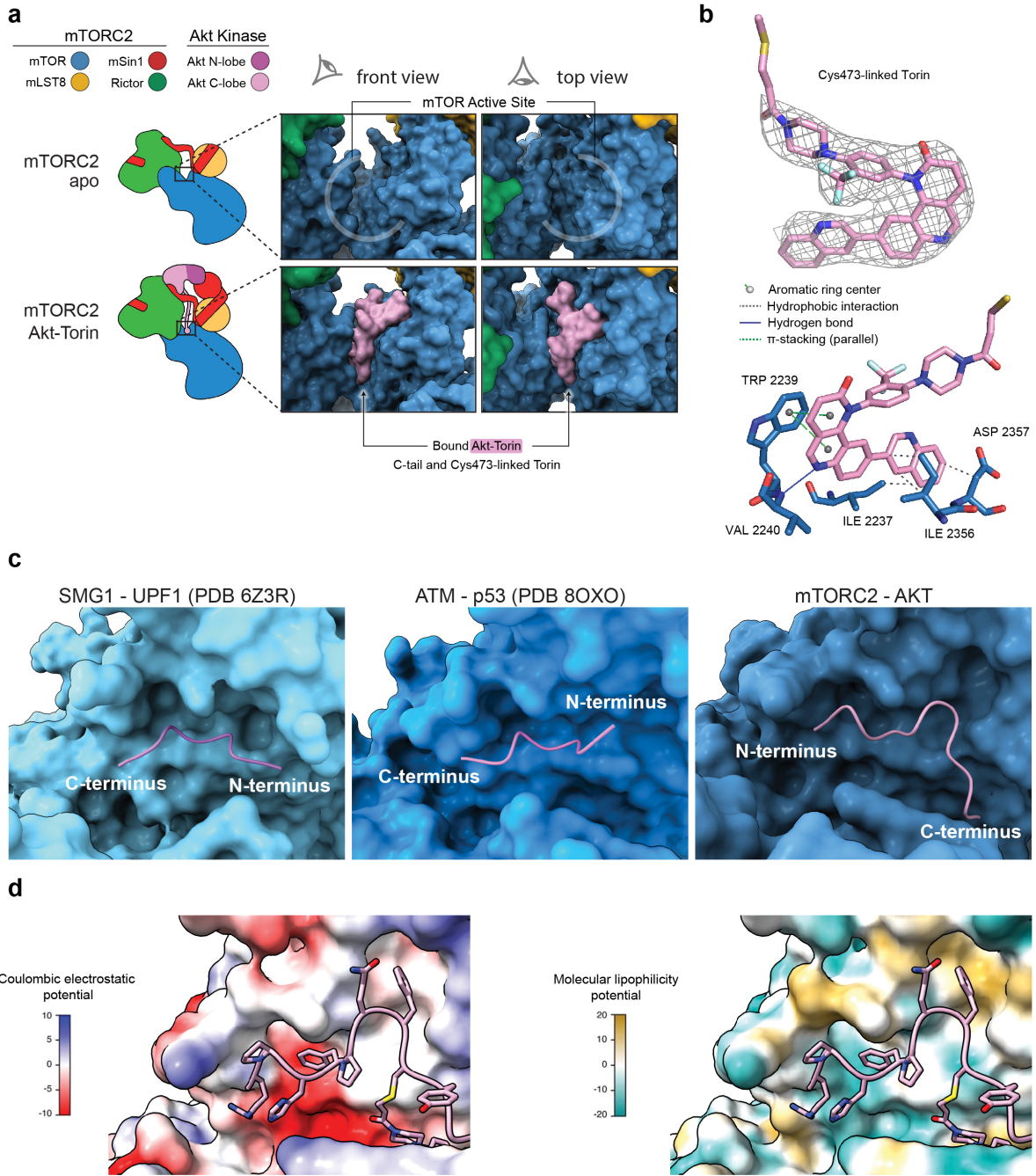
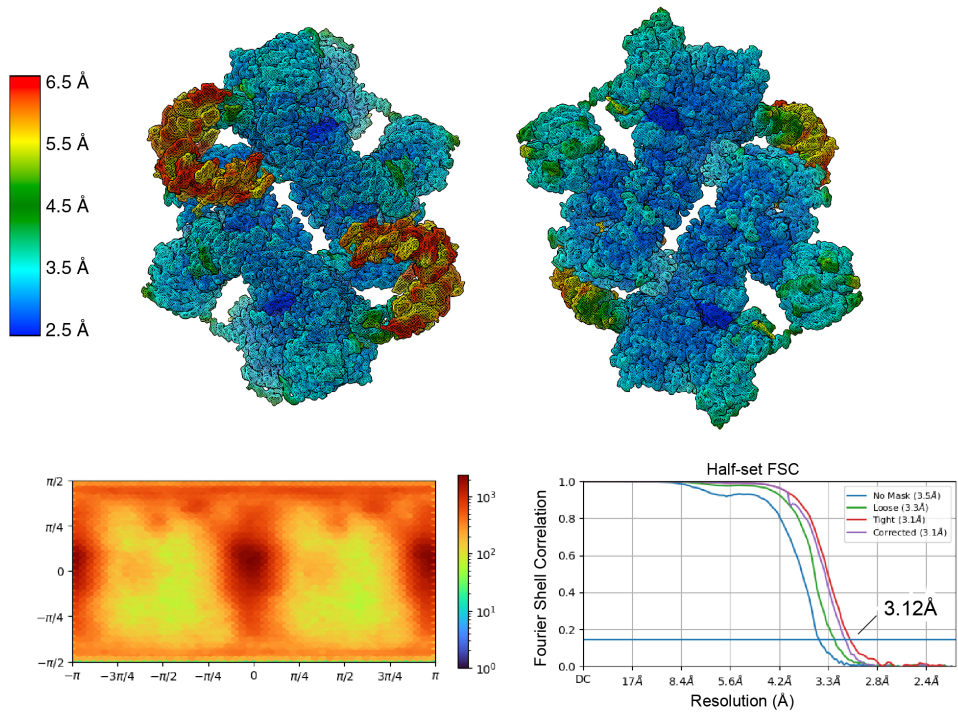


Figure 3.13

apo-mTORC2



mTORC2-Akt-Torin

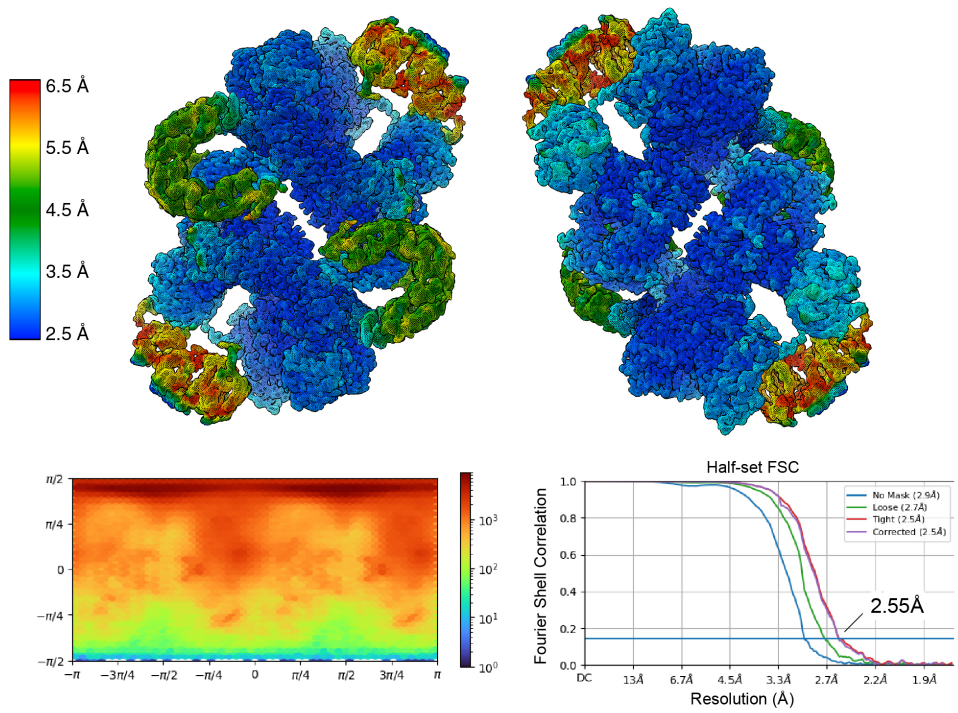


Figure 3.14

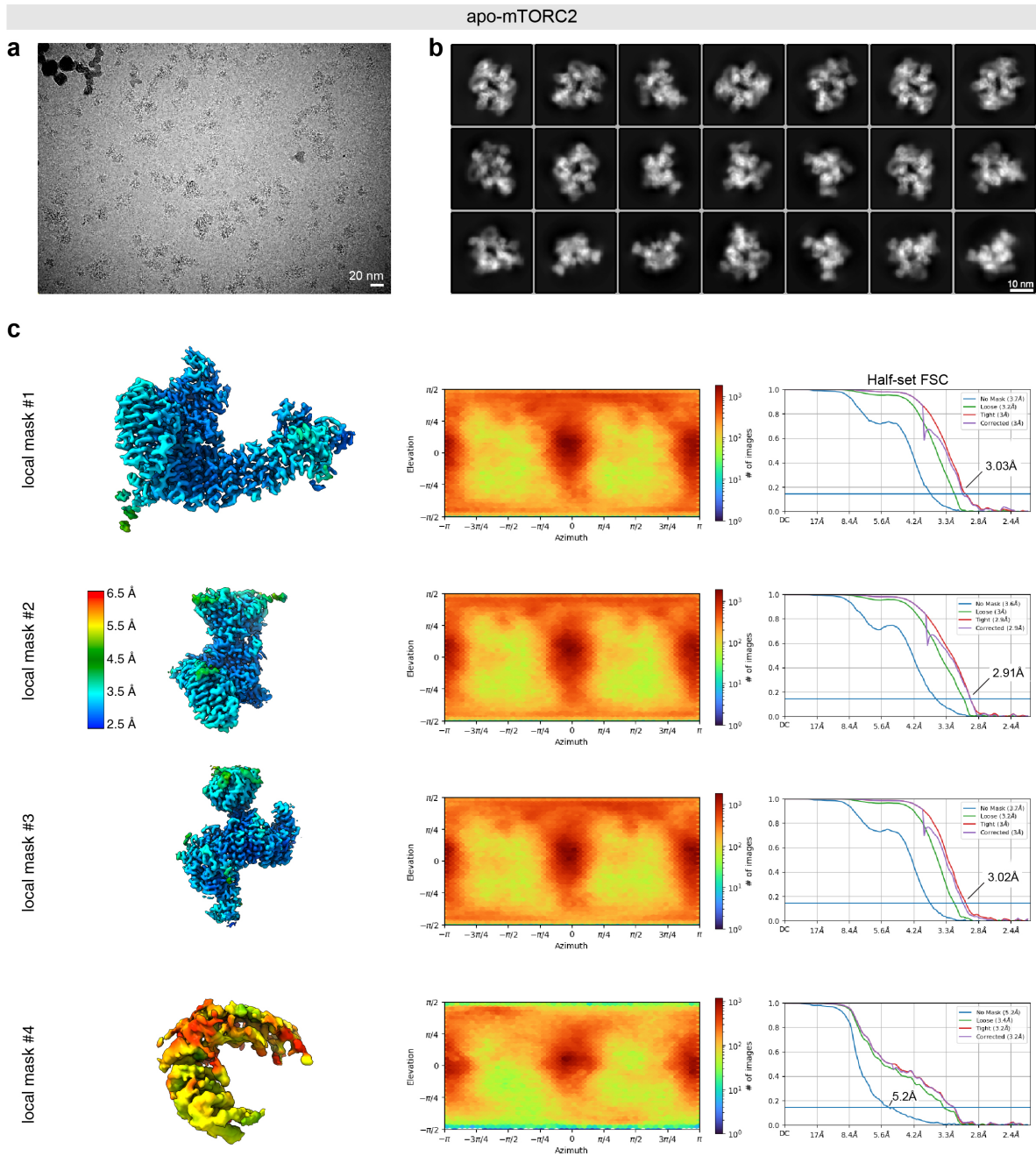


Figure 3.15

mTORC2-Akt-Torin

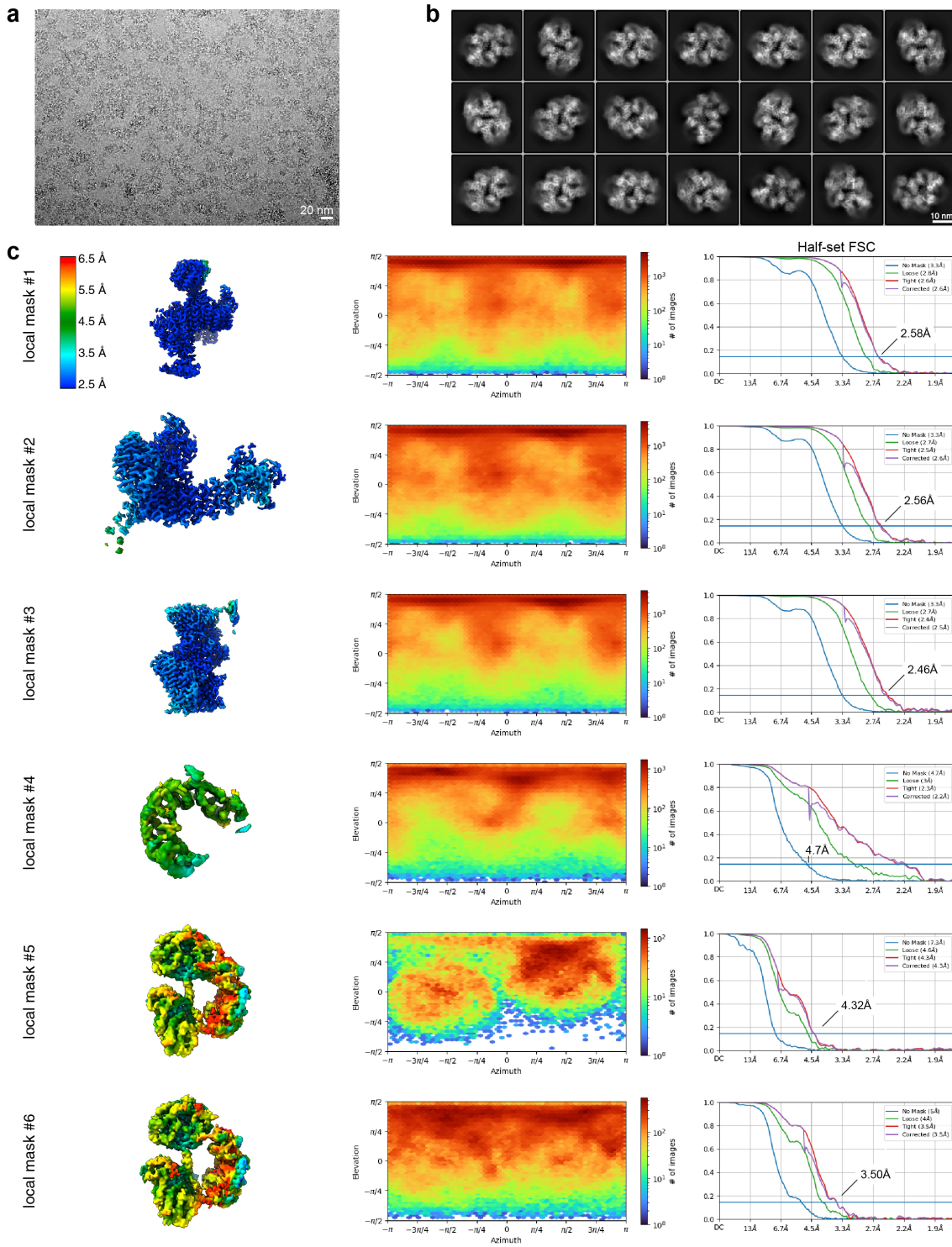


Figure 3.16

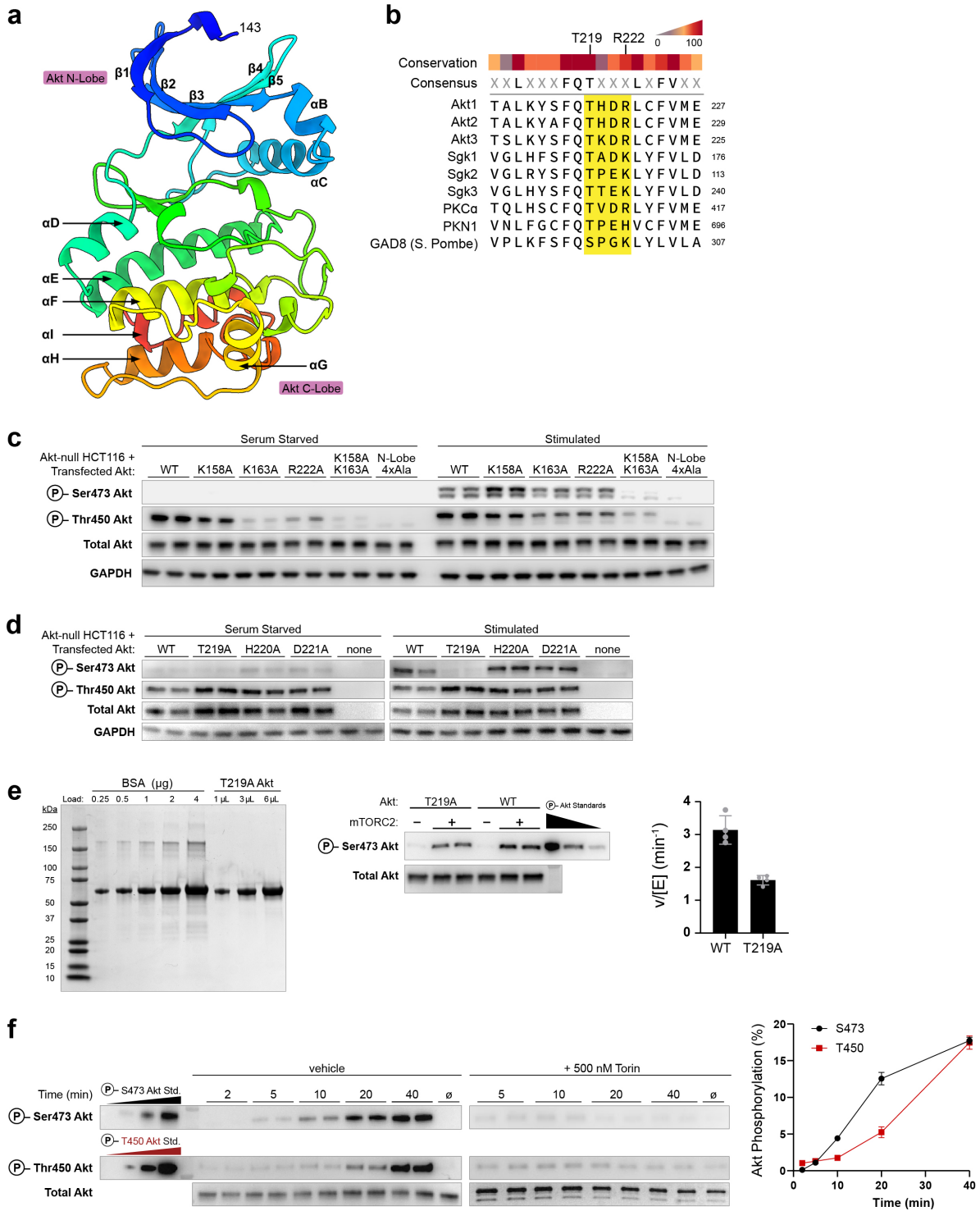


Figure 3.17

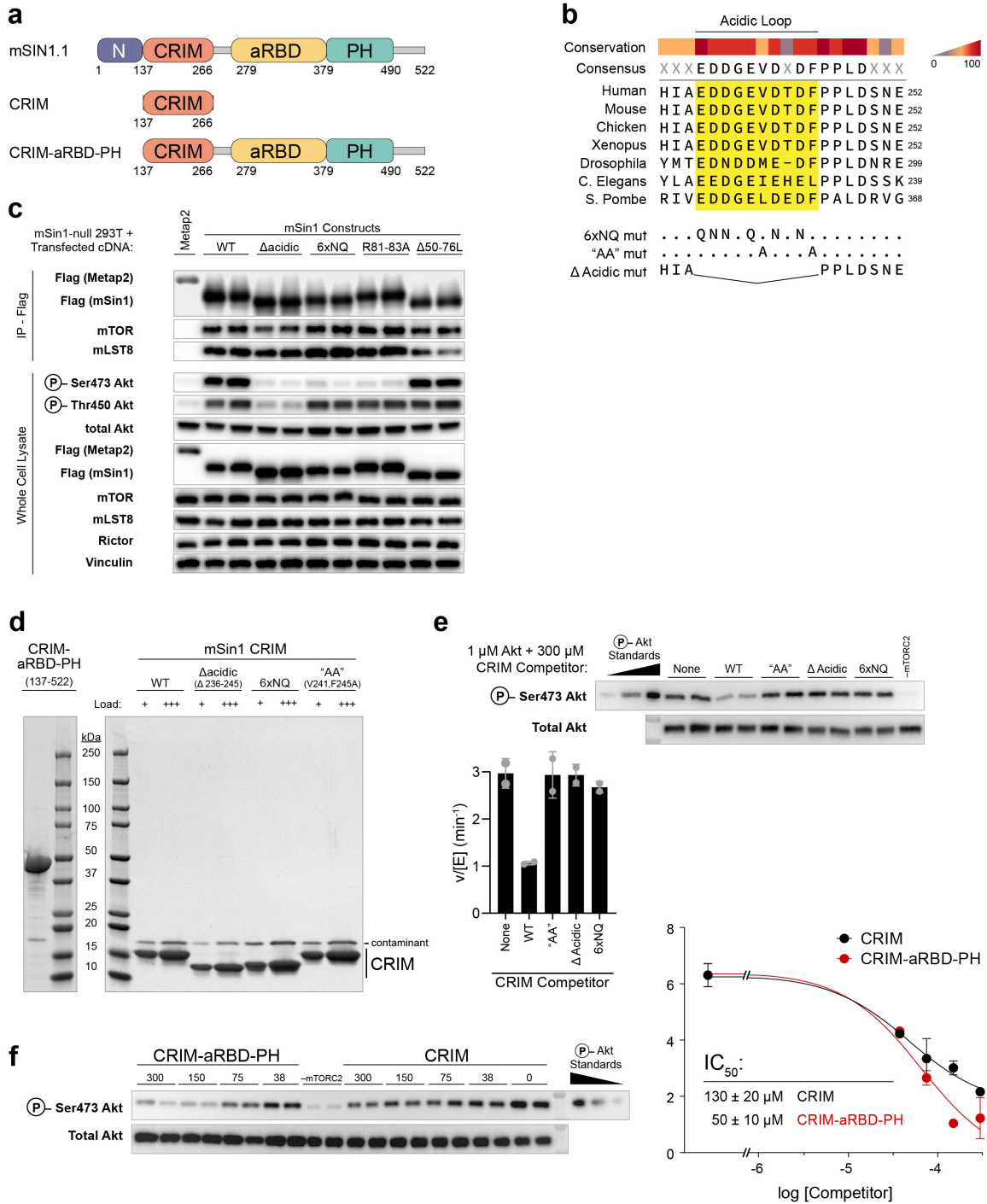


Figure 3.18

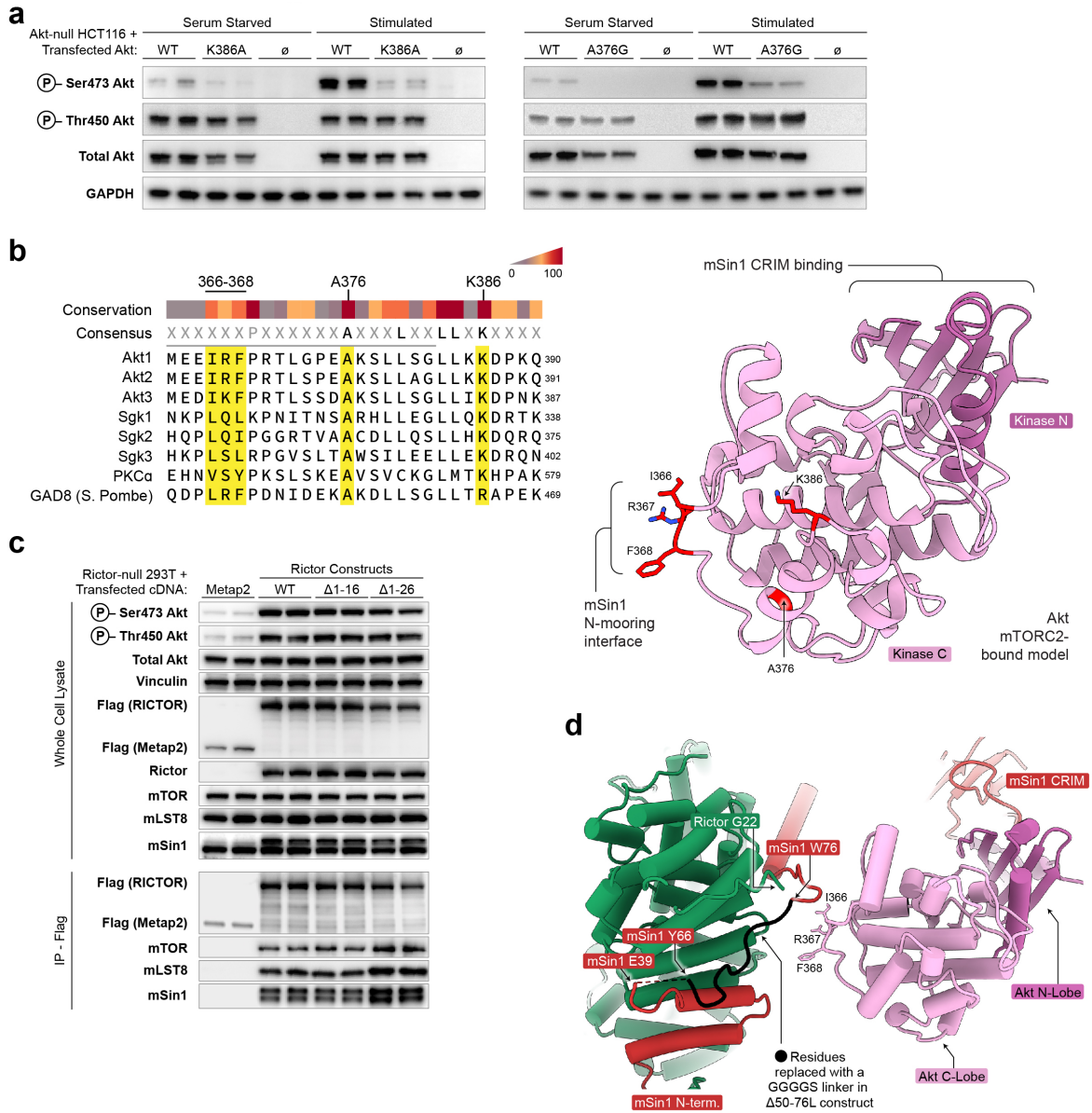


Figure 3.19

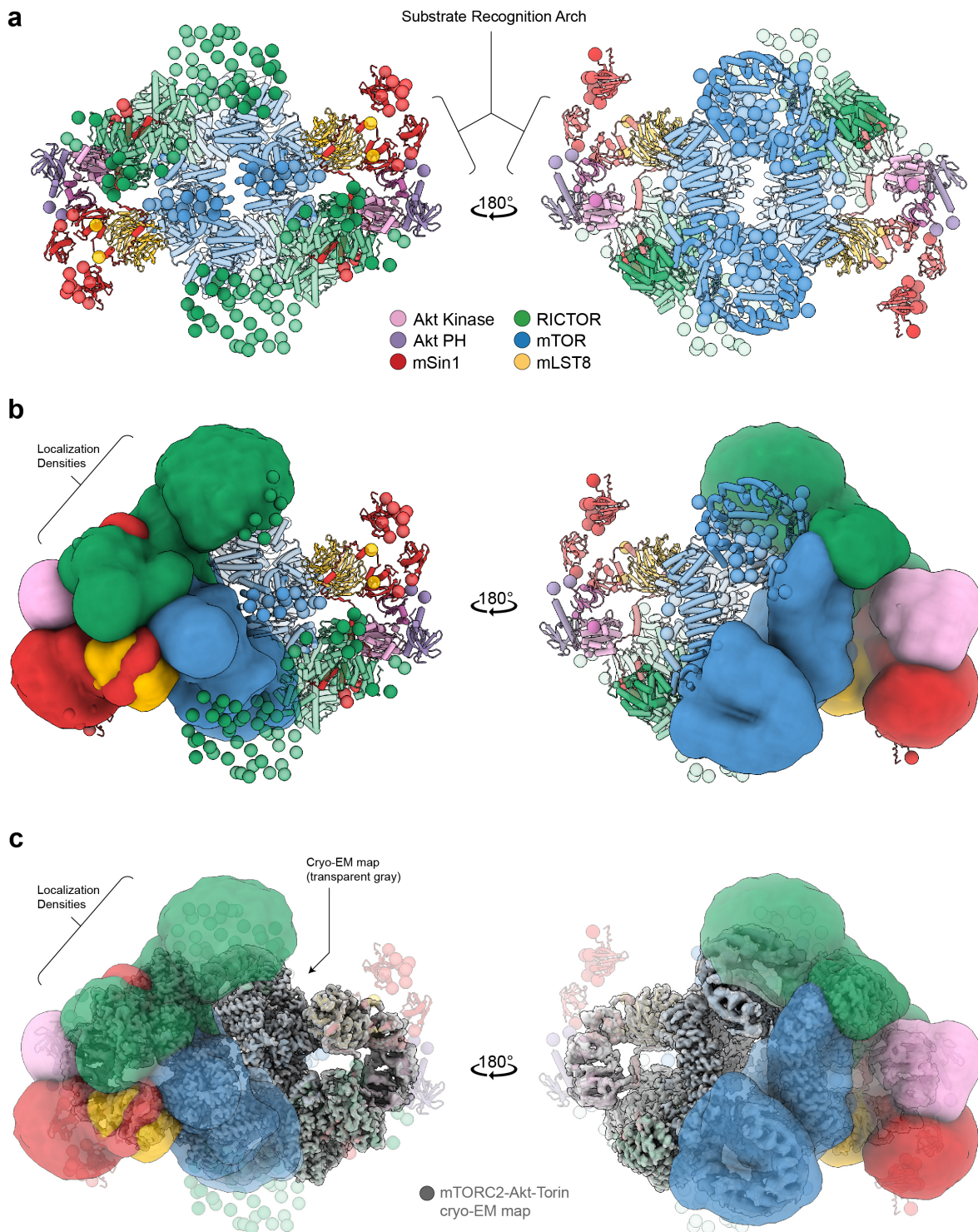


Figure 3.20

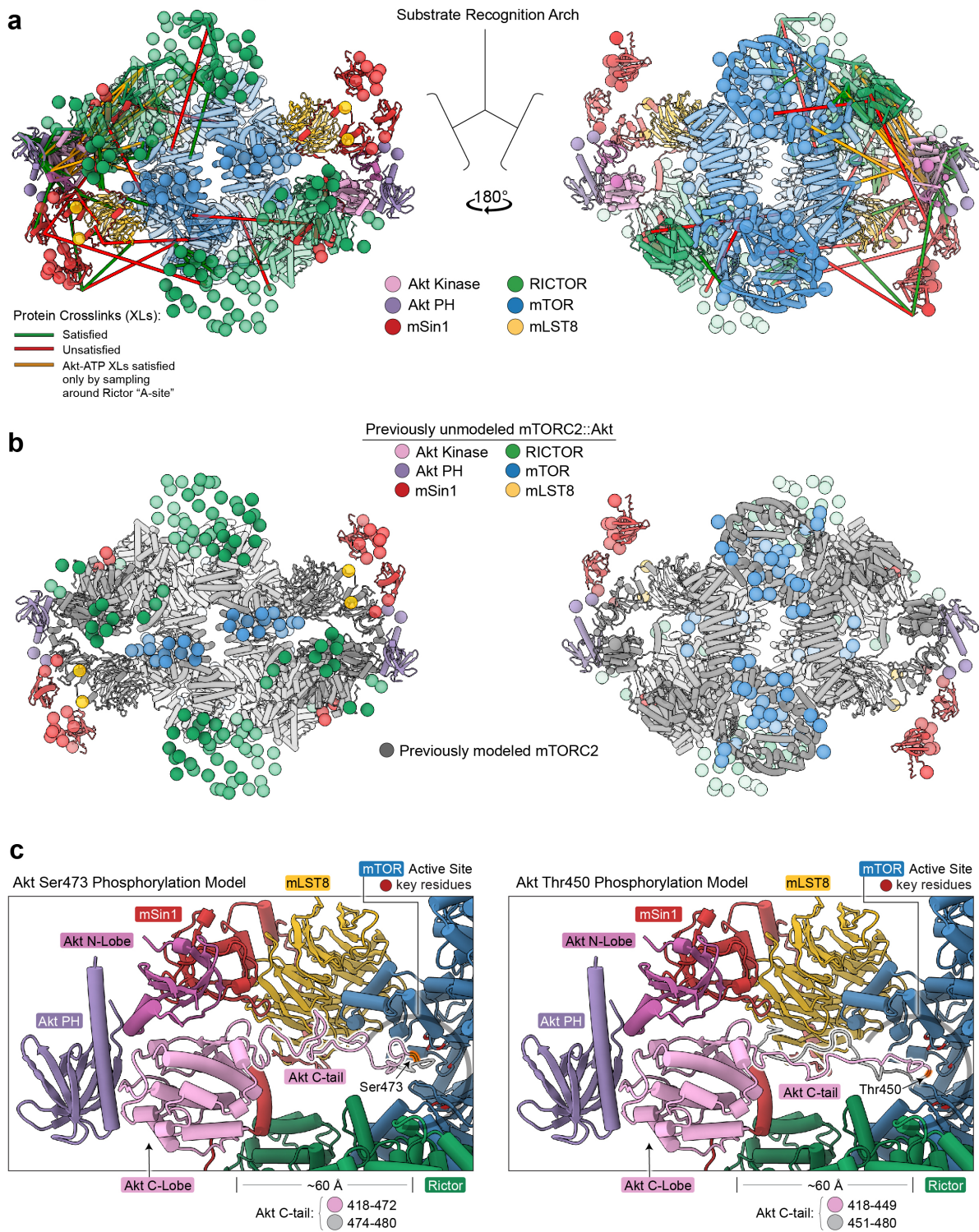


Figure 3.21

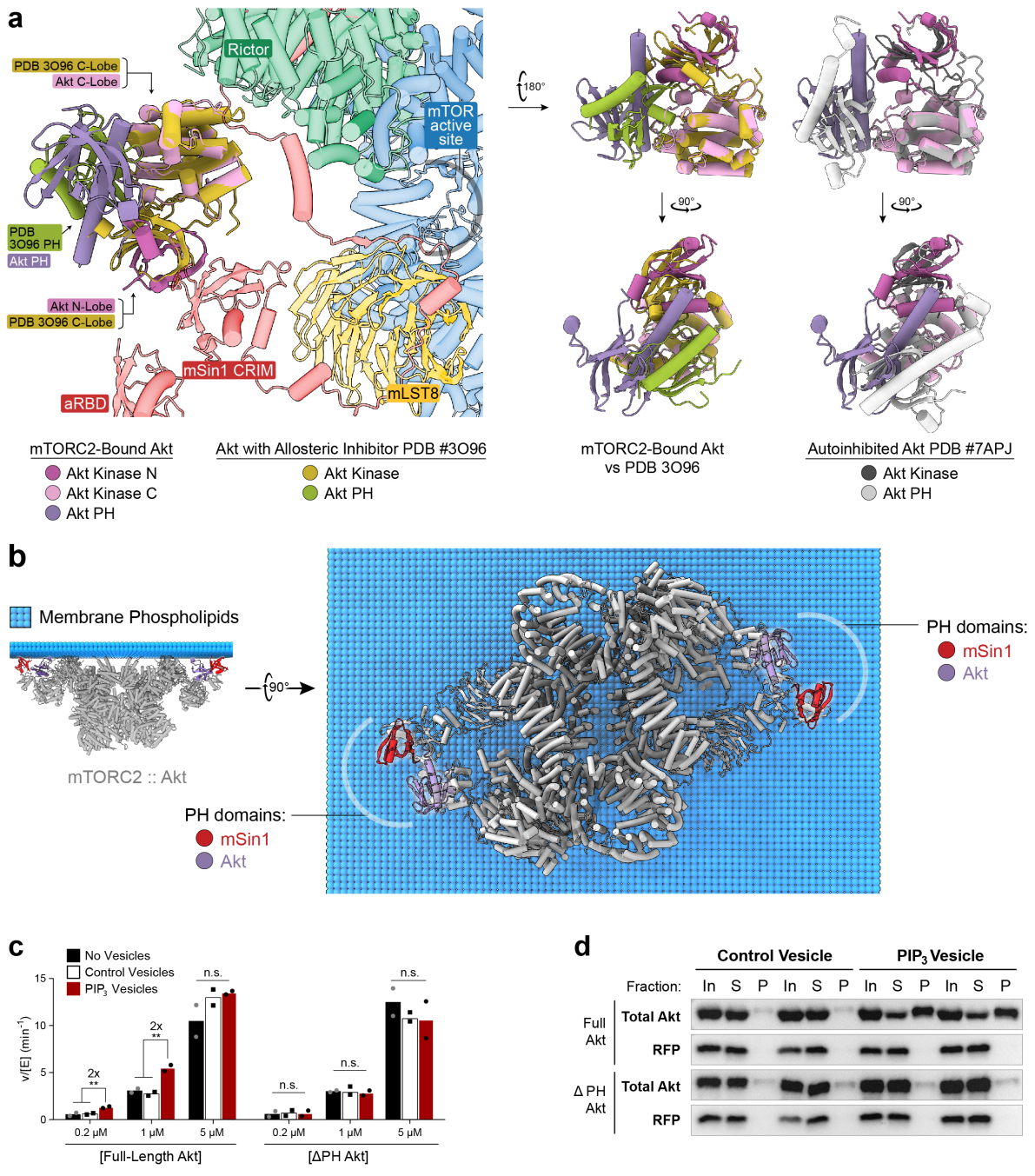


Figure 3.22

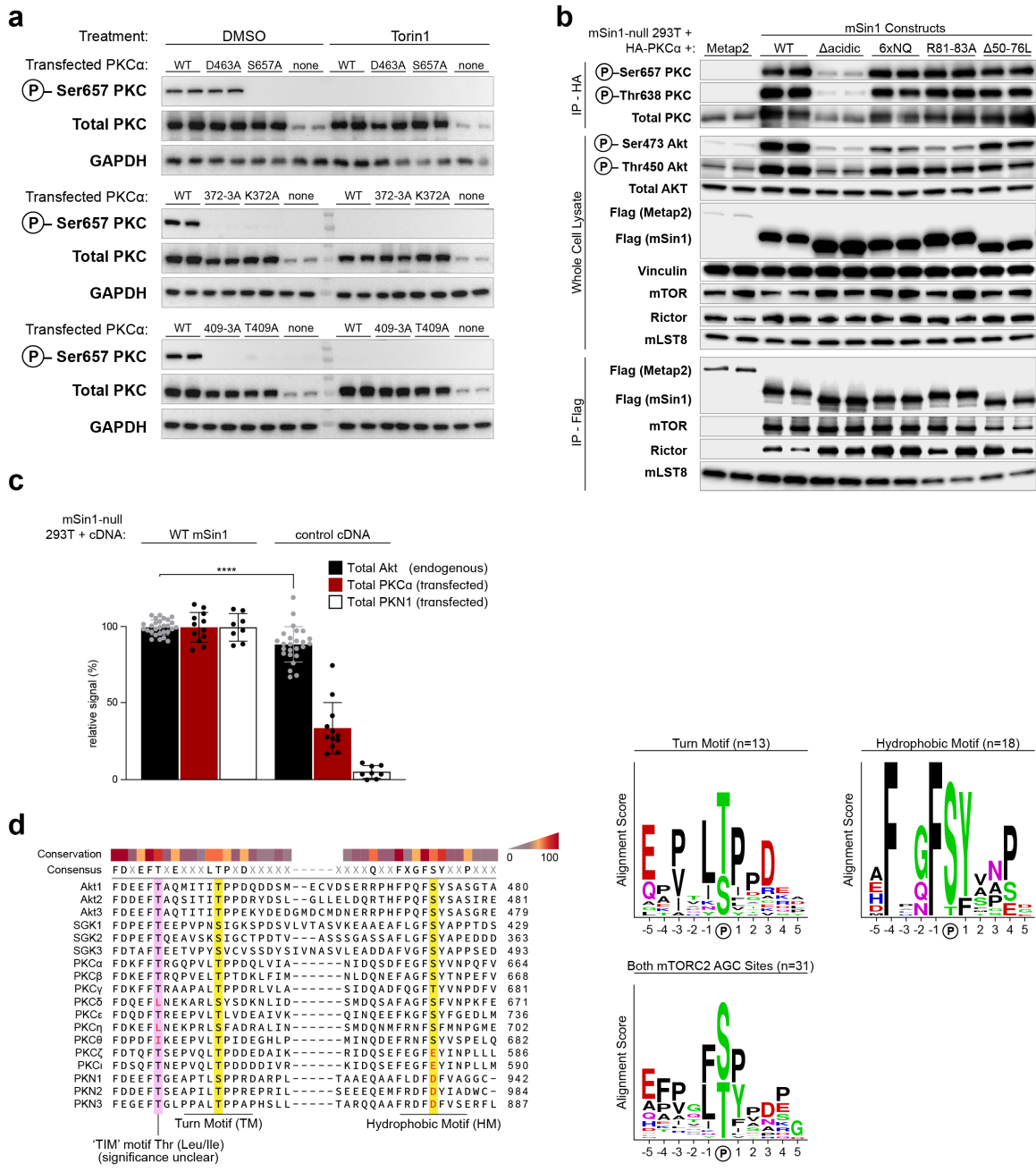


Figure 3.23

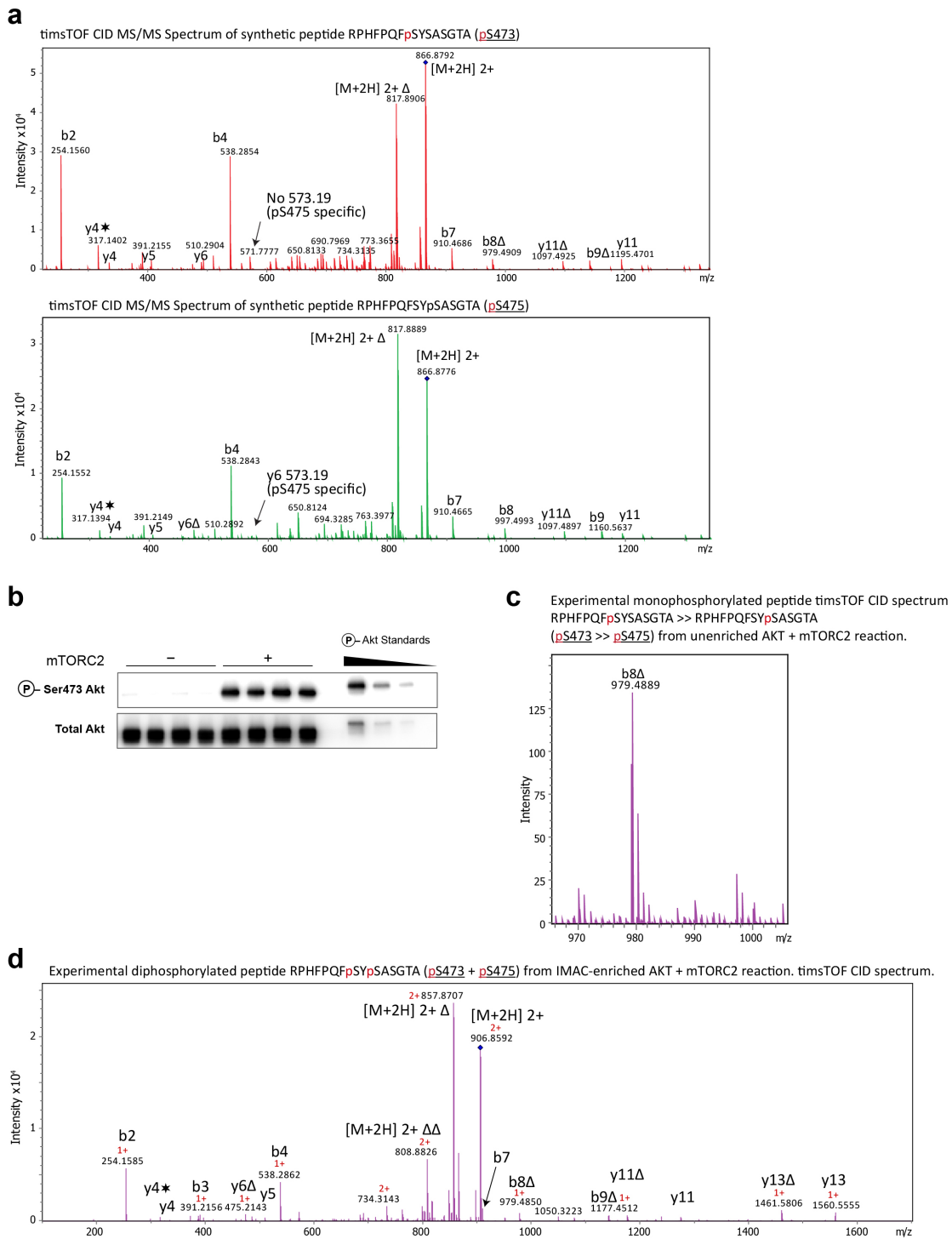
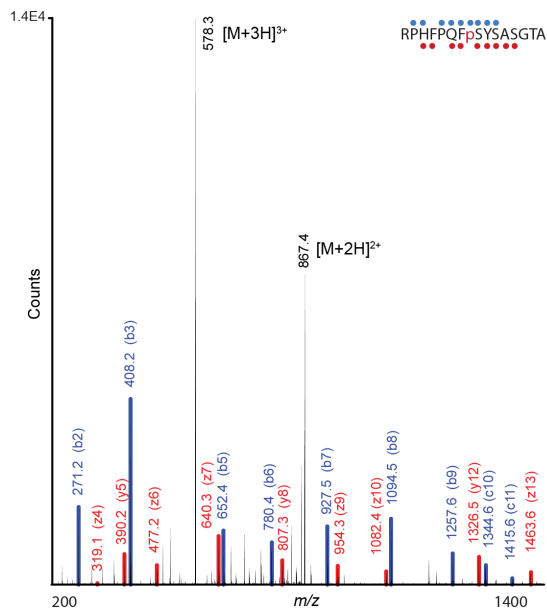


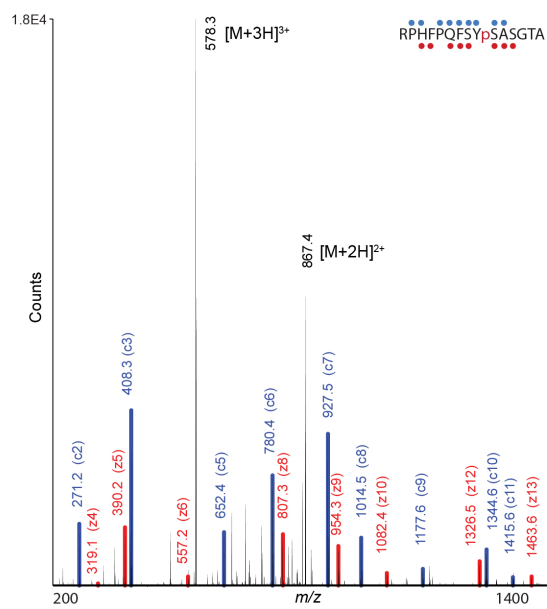
Figure 3.24

ZenoTOF EAD MS/MS Spectra

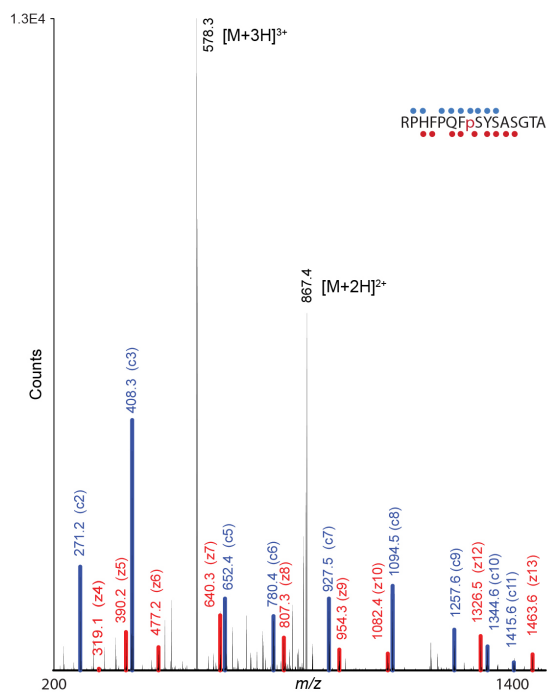
Synthetic peptide RPHFPQFpSYSASGTA (pS473)



Synthetic peptide RPHFPQFSYpSASGTA (pS475)



Experimental spectrum RPHFPQFpSYSASGTA (pS473) from unenriched AKT + mTORC2 reaction



Experimental spectrum RPHFPQFSYpSASGTA (Double pS473 + pS475) from IMAC-enriched AKT + mTORC2 reaction

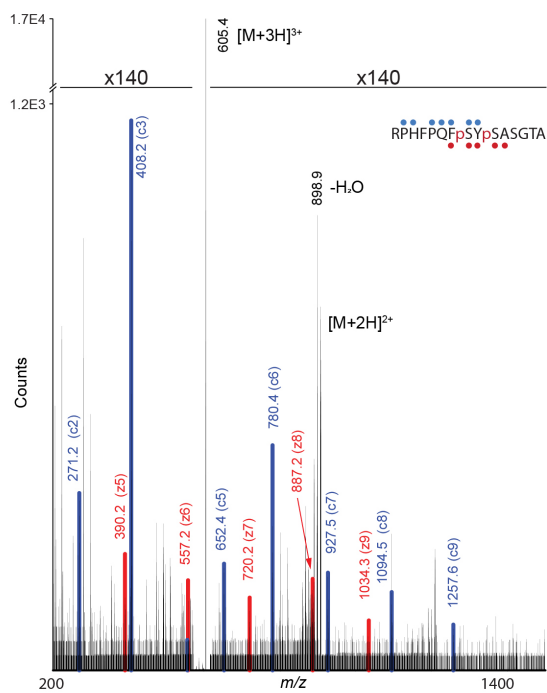


Figure 3.25

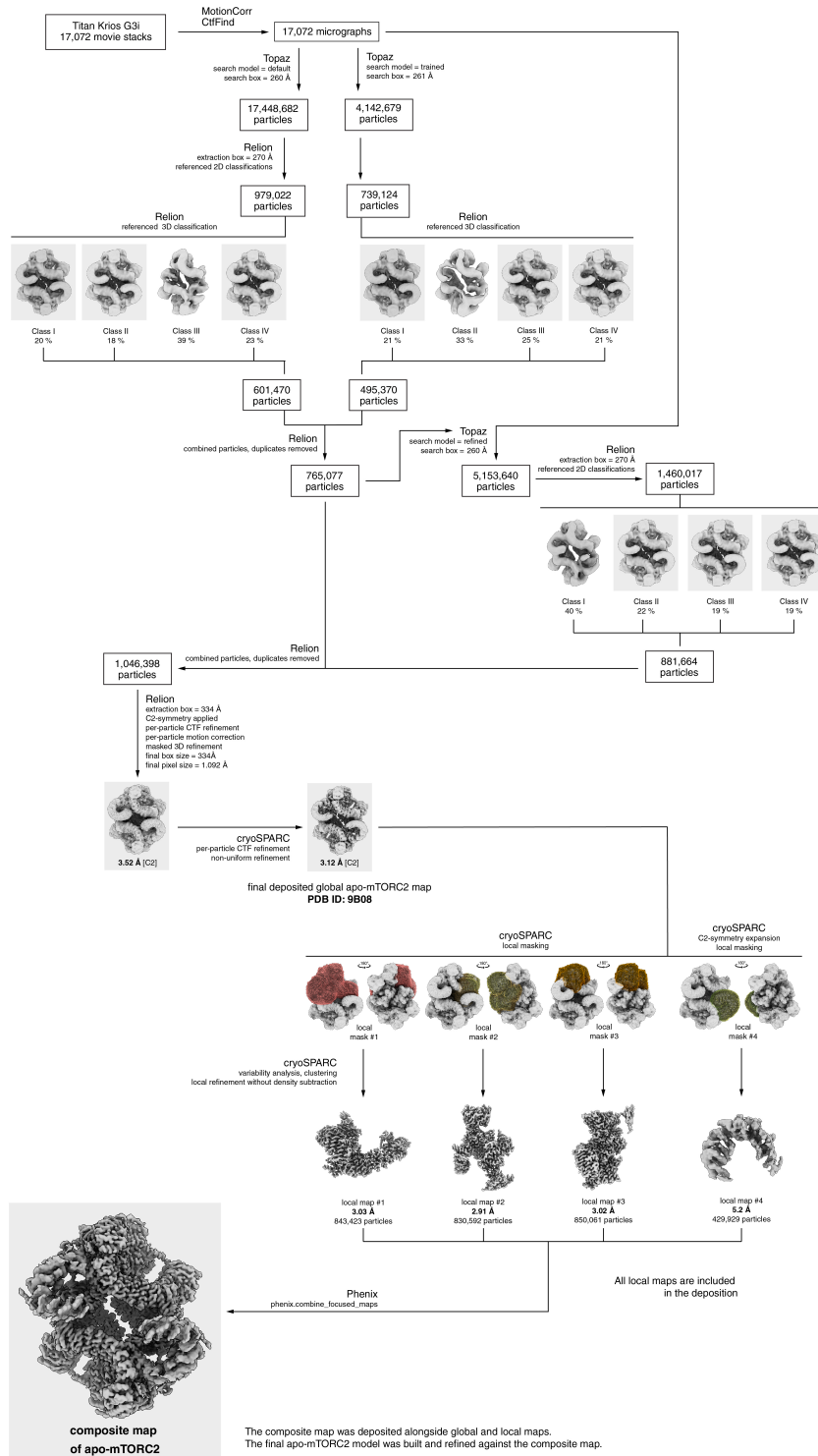


Figure 3.26

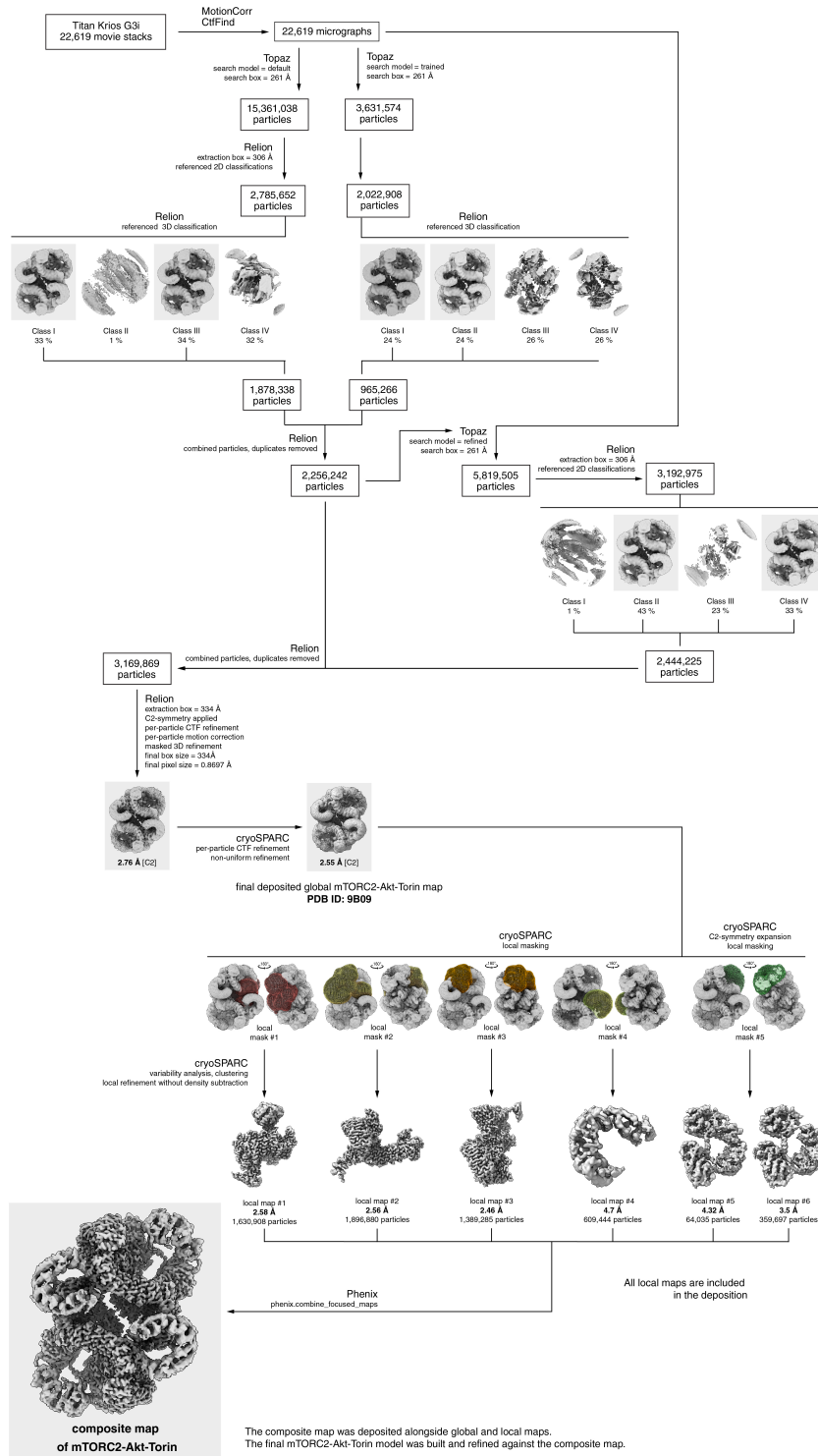


Figure 3.27

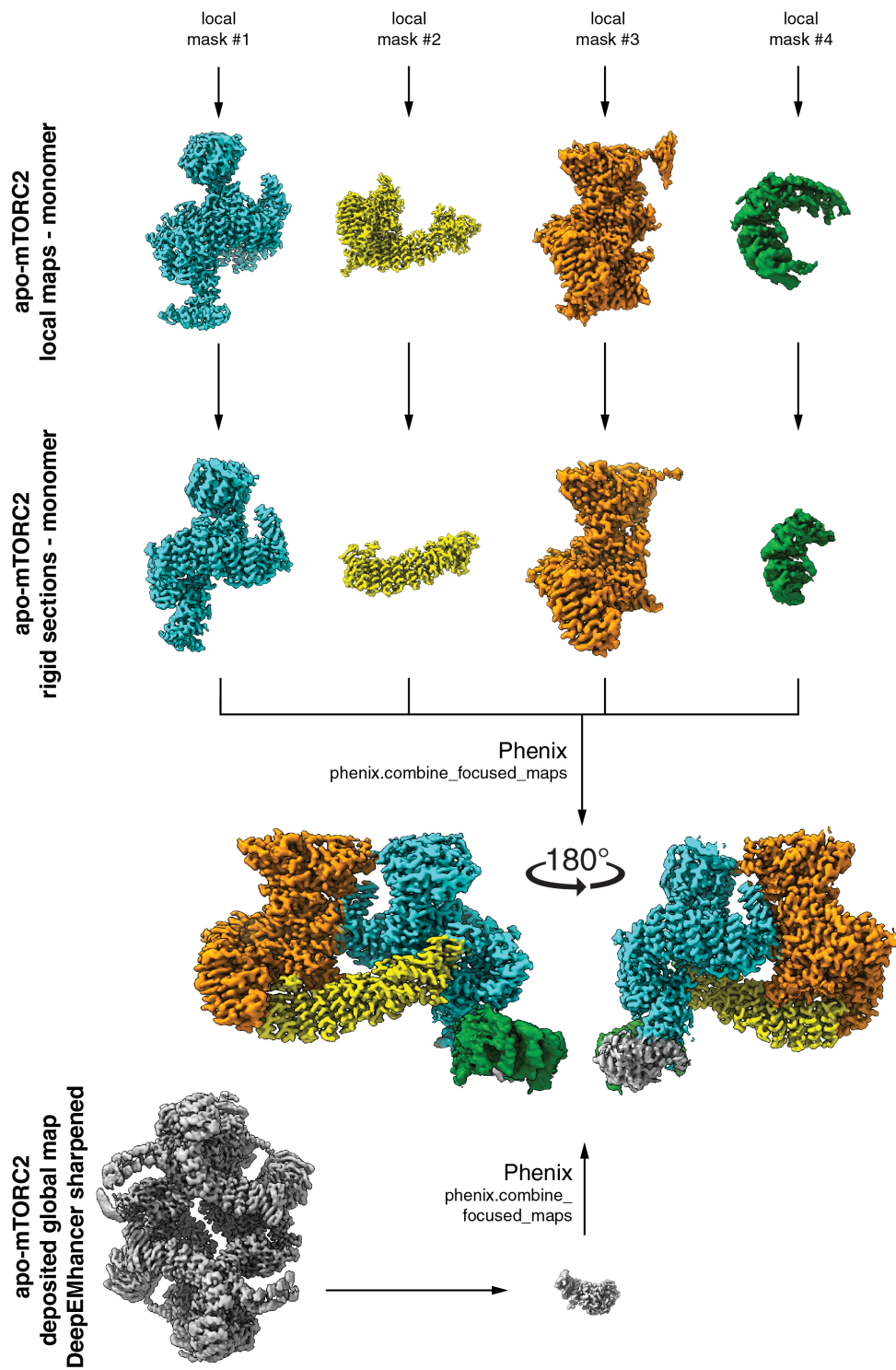


Figure 3.28

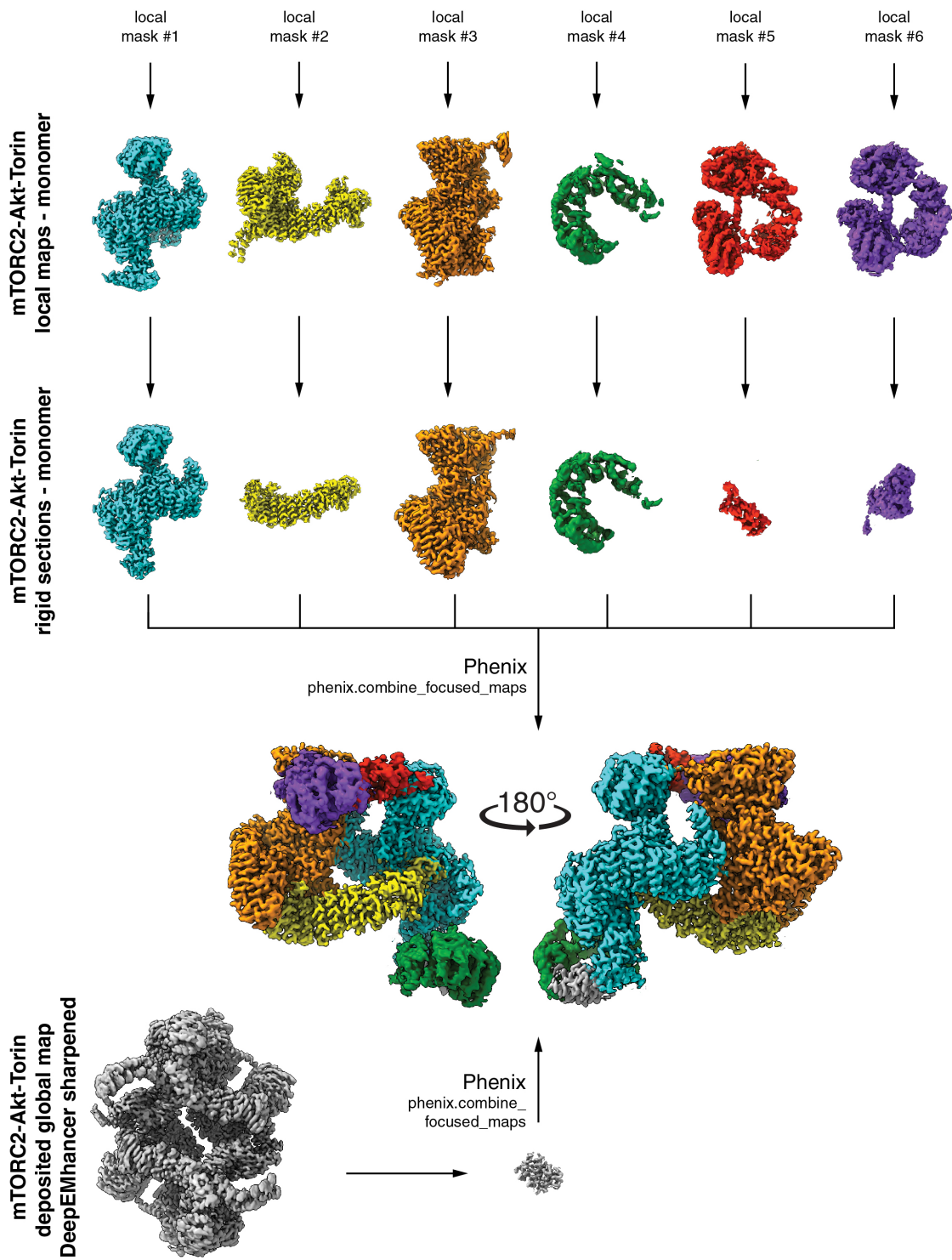


Figure 3.29

Supplementary Figure 2.5. Cryo-EM Particle Distribution and Resolution

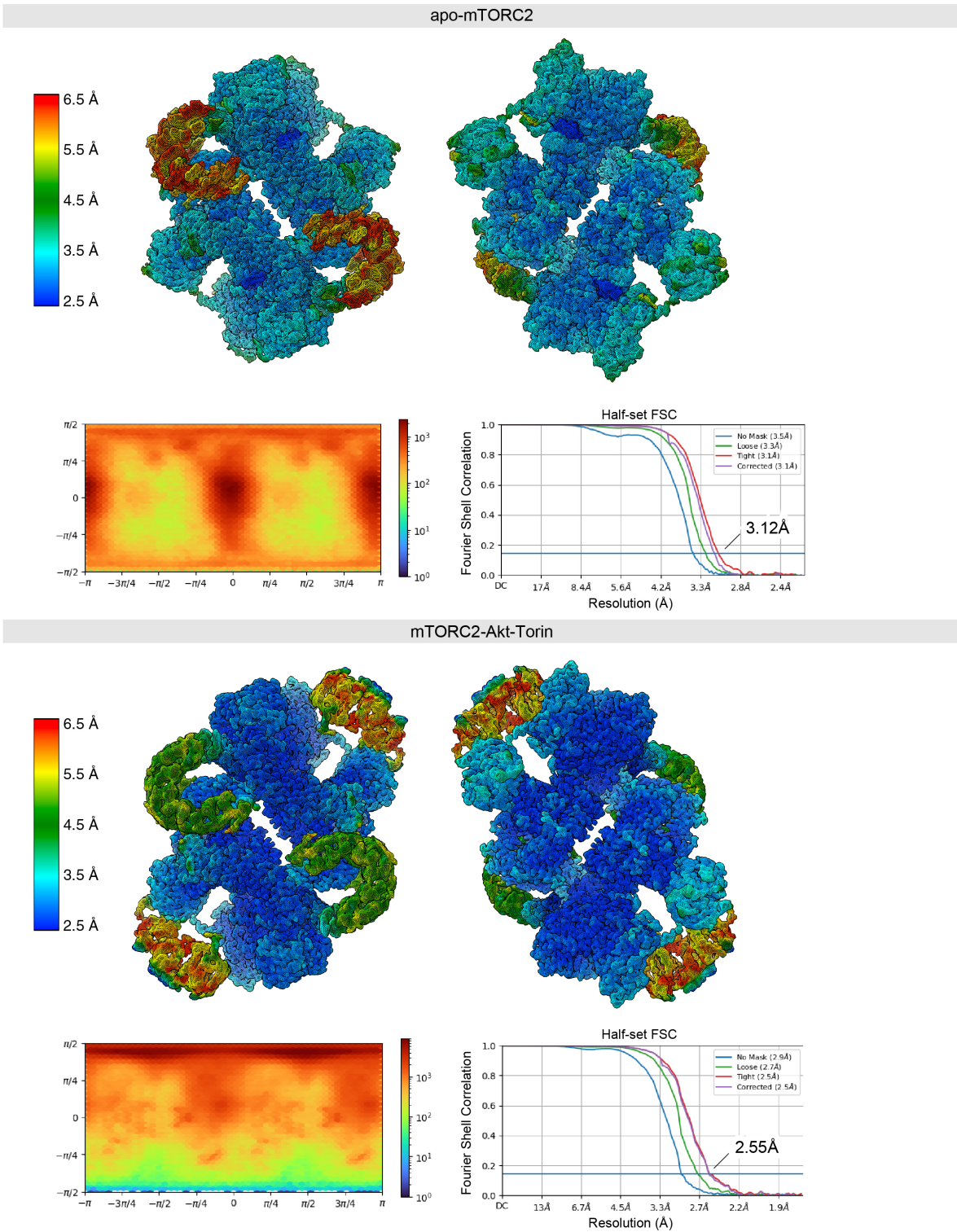


Figure 3.30

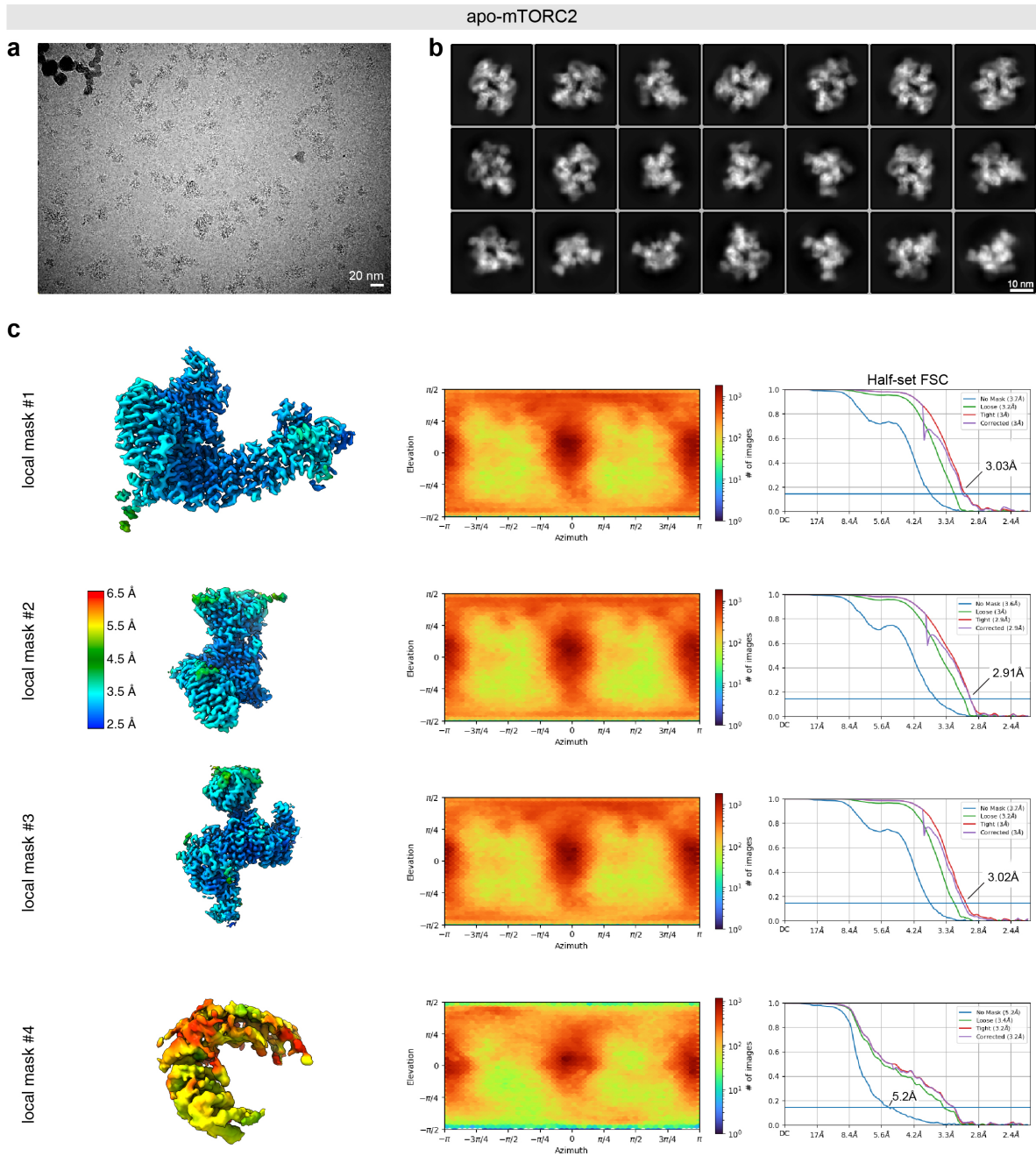


Figure 3.31

mTORC2-Akt-Torin

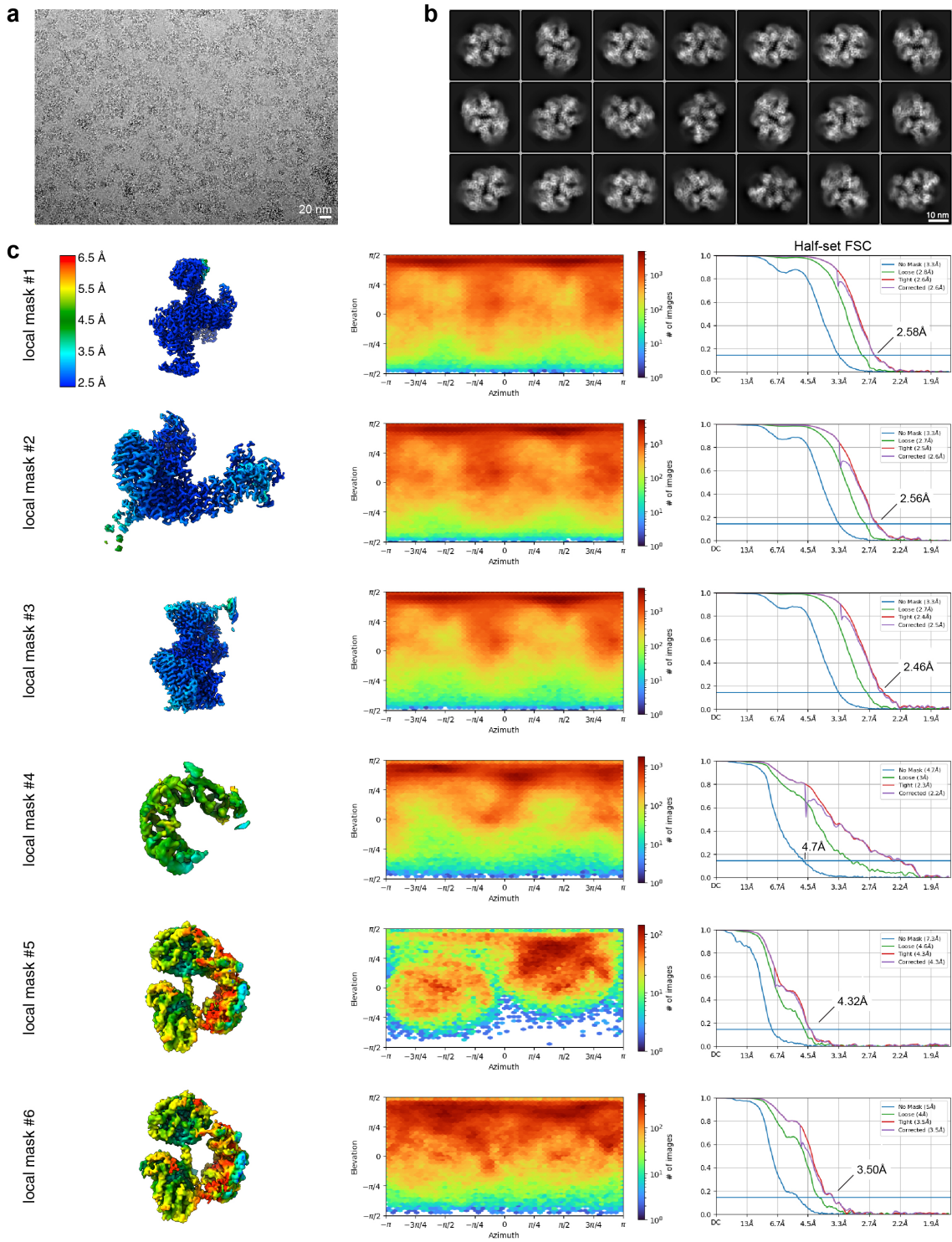


Figure 3.32

References

1. Sarbassov, D. D., Guertin, D. A., Ali, S. M. & Sabatini, D. M. Phosphorylation and regulation of Akt/PKB by the rictor-mTOR complex. en. *Science* **307**, 1098–1101 (Feb. 2005) (cit. on pp. 37, 38, 41).
2. Battagioni, S., Benjamin, D., Wälchli, M., Maier, T. & Hall, M. N. mTOR substrate phosphorylation in growth control. en. *Cell* **185**, 1814–1836 (May 2022) (cit. on pp. 37, 38, 66).
3. Ragupathi, A., Kim, C. & Jacinto, E. The mTORC2 signaling network: targets and cross-talks. en. *Biochem. J* **481**, 45–91 (Jan. 2024) (cit. on p. 37).
4. Scaiola, A., Mangia, F., Imseng, S., Boehringer, D., Berneiser, K., Shimobayashi, M., Stutfeld, E., Hall, M. N., Ban, N. & Maier, T. The 3.2-Å resolution structure of human mTORC2. en. *Sci Adv* **6** (Nov. 2020) (cit. on pp. 37, 38, 40, 41, 44, 48, 66).
5. Yu, Z., Chen, J., Takagi, E., Wang, F., Saha, B., Liu, X., Joubert, L.-M., Gleason, C. E., Jin, M., Li, C., Nowotny, C., Agard, D., Cheng, Y. & Pearce, D. Interactions between mTORC2 core subunits Rictor and mSin1 dictate selective and context-dependent phosphorylation of substrate kinases SGK1 and Akt. en. *J. Biol. Chem.* **298**, 102288 (Sept. 2022) (cit. on pp. 37, 40, 41, 44, 56, 66).
6. Baffi, T. R., Lordén, G., Wozniak, J. M., Feichtner, A., Yeung, W., Kornev, A. P., King, C. C., Del Rio, J. C., Limaye, A. J., Bogomolovas, J., Gould, C. M., Chen, J., Kennedy, E. J., Kannan, N., Gonzalez, D. J., Stefan, E., Taylor, S. S. & Newton, A. C. mTORC2 controls the activity of PKC and Akt by phosphorylating a conserved TOR interaction motif. en. *Sci. Signal.* **14** (Apr. 2021) (cit. on pp. 37, 40, 41).
7. Manning, B. D. & Toker, A. AKT/PKB Signaling: Navigating the Network. en. *Cell* **169**, 381–405 (Apr. 2017) (cit. on pp. 37, 41, 43).

8. Liu, G. Y. & Sabatini, D. M. mTOR at the nexus of nutrition, growth, ageing and disease. en. *Nat. Rev. Mol. Cell Biol.* **21**, 183–203 (Apr. 2020) (cit. on pp. 38, 43).
9. Ebner, M., Sinkovics, B., Szczygieł, M., Ribeiro, D. W. & Yudushkin, I. Localization of mTORC2 activity inside cells. en. *J. Cell Biol.* **216**, 343–353 (Feb. 2017) (cit. on p. 38).
10. Alessi, D. R., Andjelkovic, M., Caudwell, B., Cron, P., Morrice, N., Cohen, P. & Hemmings, B. A. Mechanism of activation of protein kinase B by insulin and IGF-1. en. *EMBO J.* **15**, 6541–6551 (Dec. 1996) (cit. on pp. 38, 41).
11. Alessi, D. R., James, S. R., Downes, C. P., Holmes, A. B., Gaffney, P. R., Reese, C. B. & Cohen, P. Characterization of a 3-phosphoinositide-dependent protein kinase which phosphorylates and activates protein kinase B α . en. *Curr. Biol.* **7**, 261–269 (Apr. 1997) (cit. on pp. 38, 41).
12. Chu, N., Salguero, A. L., Liu, A. Z., Chen, Z., Dempsey, D. R., Ficarro, S. B., Alexander, W. M., Marto, J. A., Li, Y., Amzel, L. M., Gabelli, S. B. & Cole, P. A. Akt Kinase Activation Mechanisms Revealed Using Protein Semisynthesis. en. *Cell* **174**, 897–907.e14 (Aug. 2018) (cit. on pp. 38, 41, 44, 72, 80–82).
13. Ikenoue, T., Inoki, K., Yang, Q., Zhou, X. & Guan, K.-L. Essential function of TORC2 in PKC and Akt turn motif phosphorylation, maturation and signalling. en. *EMBO J.* **27**, 1919–1931 (July 2008) (cit. on pp. 38, 41, 54).
14. García-Martínez, J. M. & Alessi, D. R. mTOR complex 2 (mTORC2) controls hydrophobic motif phosphorylation and activation of serum- and glucocorticoid-induced protein kinase 1 (SGK1). en. *Biochem. J* **416**, 375–385 (Dec. 2008) (cit. on p. 38).
15. Oleksak, P., Nepovimova, E., Chrienova, Z., Musilek, K., Patocka, J. & Kuca, K. Contemporary mTOR inhibitor scaffolds to diseases breakdown: A patent review (2015–2021). en. *Eur. J. Med. Chem.* **238**, 114498 (Aug. 2022) (cit. on p. 38).

16. Linde-Garelli, K. Y. & Rogala, K. B. Structural mechanisms of the mTOR pathway. en. *Curr. Opin. Struct. Biol.* **82**, 102663 (Oct. 2023) (cit. on p. 38).
17. Nojima, H., Tokunaga, C., Eguchi, S., Oshiro, N., Hidayat, S., Yoshino, K.-I., Hara, K., Tanaka, N., Avruch, J. & Yonezawa, K. The mammalian target of rapamycin (mTOR) partner, raptor, binds the mTOR substrates p70 S6 kinase and 4E-BP1 through their TOR signaling (TOS) motif. en. *J. Biol. Chem.* **278**, 15461–15464 (May 2003) (cit. on p. 40).
18. Schalm, S. S. & Blenis, J. Identification of a conserved motif required for mTOR signaling. en. *Curr. Biol.* **12**, 632–639 (Apr. 2002) (cit. on p. 40).
19. Yang, H., Jiang, X., Li, B., Yang, H. J., Miller, M., Yang, A., Dhar, A. & Pavletich, N. P. Mechanisms of mTORC1 activation by RHEB and inhibition by PRAS40. en. *Nature* **552**, 368–373 (Dec. 2017) (cit. on pp. 40, 41).
20. Tee, A. R. & Proud, C. G. Caspase cleavage of initiation factor 4E-binding protein 1 yields a dominant inhibitor of cap-dependent translation and reveals a novel regulatory motif. en. *Mol. Cell. Biol.* **22**, 1674–1683 (Mar. 2002) (cit. on p. 40).
21. Cui, Z., Napolitano, G., de Araujo, M. E. G., Esposito, A., Monfregola, J., Huber, L. A., Ballabio, A. & Hurley, J. H. Structure of the lysosomal mTORC1-TFEB-Rag-Ragulator megacomplex. en. *Nature* **614**, 572–579 (Feb. 2023) (cit. on p. 40).
22. Böhm, R., Imseng, S., Jakob, R. P., Hall, M. N., Maier, T. & Hiller, S. The dynamic mechanism of 4E-BP1 recognition and phosphorylation by mTORC1. en. *Mol. Cell* **81**, 2403–2416.e5 (June 2021) (cit. on pp. 40, 65).
23. Tao, Z., Barker, J., Shi, S. D.-H., Gehring, M. & Sun, S. Steady-state kinetic and inhibition studies of the mammalian target of rapamycin (mTOR) kinase domain and mTOR complexes. en. *Biochemistry* **49**, 8488–8498 (Oct. 2010) (cit. on pp. 40, 75).

24. Lu, M., Wang, J., Ives, H. E. & Pearce, D. mSIN1 protein mediates SGK1 protein interaction with mTORC2 protein complex and is required for selective activation of the epithelial sodium channel. en. *J. Biol. Chem.* **286**, 30647–30654 (Sept. 2011) (cit. on p. 40).
25. Tatebe, H., Murayama, S., Yonekura, T., Hatano, T., Richter, D., Furuya, T., Kataoka, S., Furuita, K., Kojima, C. & Shiozaki, K. Substrate specificity of TOR complex 2 is determined by a ubiquitin-fold domain of the Sin1 subunit. en. *Elife* **6** (Mar. 2017) (cit. on pp. 40, 49, 52, 53).
26. Liu, P., Gan, W., Chin, Y. R., Ogura, K., Guo, J., Zhang, J., Wang, B., Blenis, J., Cantley, L. C., Toker, A., Su, B. & Wei, W. PtdIns(3,4,5)P3-Dependent Activation of the mTORC2 Kinase Complex. en. *Cancer Discov.* **5**, 1194–1209 (Nov. 2015) (cit. on p. 40).
27. Castel, P., Dharmiah, S., Sale, M. J., Messing, S., Rizzuto, G., Cuevas-Navarro, A., Cheng, A., Trnka, M. J., Urisman, A., Esposito, D., Simanshu, D. K. & McCormick, F. RAS interaction with Sin1 is dispensable for mTORC2 assembly and activity. en. *Proc. Natl. Acad. Sci. U. S. A.* **118** (Aug. 2021) (cit. on pp. 40, 45, 65).
28. Senoo, H., Murata, D., Wai, M., Arai, K., Iwata, W., Sesaki, H. & Iijima, M. KARATE: PKA-induced KRAS4B-RHOA-mTORC2 supercomplex phosphorylates AKT in insulin signaling and glucose homeostasis. en. *Mol. Cell* **81**, 4622–4634.e8 (Nov. 2021) (cit. on pp. 40, 65).
29. Facchinetti, V., Ouyang, W., Wei, H., Soto, N., Lazorchak, A., Gould, C., Lowry, C., Newton, A. C., Mao, Y., Miao, R. Q., Sessa, W. C., Qin, J., Zhang, P., Su, B. & Jacinto, E. The mammalian target of rapamycin complex 2 controls folding and stability of Akt and protein kinase C. en. *EMBO J.* **27**, 1932–1943 (July 2008) (cit. on pp. 41, 54, 64).
30. Yang, C.-S., Melhuish, T. A., Spencer, A., Ni, L., Hao, Y., Jividen, K., Harris, T. E., Snow, C., Frierson Jr, H. F., Wotton, D. & Paschal, B. M. The protein kinase C

- super-family member PKN is regulated by mTOR and influences differentiation during prostate cancer progression. en. *Prostate* **77**, 1452–1467 (Nov. 2017) (cit. on pp. 41, 61, 69).
31. Levinson, N. M., Seeliger, M. A., Cole, P. A. & Kuriyan, J. Structural basis for the recognition of c-Src by its inactivator Csk. en. *Cell* **134**, 124–134 (July 2008) (cit. on p. 41).
 32. Park, E., Rawson, S., Li, K., Kim, B.-W., Ficarro, S. B., Pino, G. G.-D., Sharif, H., Marto, J. A., Jeon, H. & Eck, M. J. Architecture of autoinhibited and active BRAF–MEK1–14-3-3 complexes. en. *Nature* **575**, 545–550 (Oct. 2019) (cit. on p. 41).
 33. Haar, E. T., Prabakhar, P., Liu, X. & Lepre, C. Crystal structure of the p38 alpha-MAPKAP kinase 2 heterodimer. en. *J. Biol. Chem.* **282**, 9733–9739 (Mar. 2007) (cit. on p. 41).
 34. Hoxhaj, G. & Manning, B. D. The PI3K–AKT network at the interface of oncogenic signalling and cancer metabolism. en. *Nat. Rev. Cancer* **20**, 74–88 (Nov. 2019) (cit. on p. 41).
 35. Chu, N., Viennet, T., Bae, H., Salguero, A., Boeszoermyeni, A., Arthanari, H. & Cole, P. A. The structural determinants of PH domain-mediated regulation of Akt revealed by segmental labeling. en. *Elife* **9** (Aug. 2020) (cit. on p. 41).
 36. Bae, H., Viennet, T., Park, E., Chu, N., Salguero, A., Eck, M. J., Arthanari, H. & Cole, P. A. PH domain-mediated autoinhibition and oncogenic activation of Akt. en. *Elife* **11** (Aug. 2022) (cit. on p. 41).
 37. Guertin, D. A., Stevens, D. M., Thoreen, C. C., Burds, A. A., Kalaany, N. Y., Moffat, J., Brown, M., Fitzgerald, K. J. & Sabatini, D. M. Ablation in mice of the mTORC components raptor, rictor, or mLST8 reveals that mTORC2 is required for signaling to Akt-FOXO and PKCalpha, but not S6K1. en. *Dev. Cell* **11**, 859–871 (Dec. 2006) (cit. on p. 41).

38. Hauge, C., Antal, T. L., Hirschberg, D., Doehn, U., Thorup, K., Idrissova, L., Hansen, K., Jensen, O. N., Jørgensen, T. J., Biondi, R. M. & Frödin, M. Mechanism for activation of the growth factor-activated AGC kinases by turn motif phosphorylation. en. *EMBO J.* **26**, 2251–2261 (May 2007) (cit. on pp. 41, 53, 54, 60, 61, 64).
39. Collins, B. J., Deak, M., Arthur, J. S. C., Armit, L. J. & Alessi, D. R. In vivo role of the PIF-binding docking site of PDK1 defined by knock-in mutation. en. *EMBO J.* **22**, 4202–4211 (Aug. 2003) (cit. on p. 41).
40. Muir, T. W., Sondhi, D. & Cole, P. A. Expressed protein ligation: A general method for protein engineering. *Proceedings of the National Academy of Sciences* **95**, 6705–6710 (1998) (cit. on p. 42).
41. Blake, J. F., Xu, R., Bencsik, J. R., Xiao, D., Kallan, N. C., Schlachter, S., Mitchell, I. S., Spencer, K. L., Banka, A. L., Wallace, E. M., Gloor, S. L., Martinson, M., Woessner, R. D., Vigers, G. P. A., Brandhuber, B. J., Liang, J., Safina, B. S., Li, J., Zhang, B., Chabot, C., Do, S., Lee, L., Oeh, J., Sampath, D., Lee, B. B., Lin, K., Liederer, B. M. & Skelton, N. J. Discovery and Preclinical Pharmacology of a Selective ATP-Competitive Akt Inhibitor (GDC-0068) for the Treatment of Human Tumors. *J. Med. Chem.* **55**, 8110–8127 (Sept. 2012) (cit. on p. 42).
42. Davies, B. R., Greenwood, H., Dudley, P., Crafter, C., Yu, D.-H., Zhang, J., Li, J., Gao, B., Ji, Q., Maynard, J., Ricketts, S.-A., Cross, D., Cosulich, S., Chresta, C. C., Page, K., Yates, J., Lane, C., Watson, R., Luke, R., Ogilvie, D. & Pass, M. Preclinical pharmacology of AZD5363, an inhibitor of AKT: pharmacodynamics, antitumor activity, and correlation of monotherapy activity with genetic background. en. *Mol. Cancer Ther.* **11**, 873–887 (Apr. 2012) (cit. on p. 42).
43. Yap, T. A., Yan, L., Patnaik, A., Fearen, I., Olmos, D., Papadopoulos, K., Baird, R. D., Delgado, L., Taylor, A., Lupinacci, L., Riisnaes, R., Pope, L. L., Heaton, S. P., Thomas, G., Garrett, M. D., Sullivan, D. M., de Bono, J. S. & Tolcher, A. W. First-in-man clinical

- trial of the oral pan-AKT inhibitor MK-2206 in patients with advanced solid tumors. en. *J. Clin. Oncol.* **29**, 4688–4695 (Dec. 2011) (cit. on p. 42).
44. Ericson, K., Gan, C., Cheong, I., Rago, C., Samuels, Y., Velculescu, V. E., Kinzler, K. W., Huso, D. L., Vogelstein, B. & Papadopoulos, N. Genetic inactivation of *AKT1*, *AKT2*, and *PDPK1* in human colorectal cancer cells clarifies their roles in tumor growth regulation. *Proceedings of the National Academy of Sciences* **107**, 2598–2603 (2010) (cit. on pp. 43, 69).
45. Salguero, A. L., Chen, M., Balana, A. T., Chu, N., Jiang, H., Palanski, B. A., Bae, H., Wright, K. M., Nathan, S., Zhu, H., Gabelli, S. B., Pratt, M. R. & Cole, P. A. Multifaceted Regulation of Akt by Diverse C-Terminal Post-translational Modifications. *ACS Chem. Biol.* **17**, 68–76 (Jan. 2022) (cit. on pp. 43, 44).
46. Liu, P., Begley, M., Michowski, W., Inuzuka, H., Ginzberg, M., Gao, D., Tsou, P., Gan, W., Papa, A., Kim, B. M., Wan, L., Singh, A., Zhai, B., Yuan, M., Wang, Z., Gygi, S. P., Lee, T. H., Lu, K.-P., Toker, A., Pandolfi, P. P., Asara, J. M., Kirschner, M. W., Sicinski, P., Cantley, L. & Wei, W. Cell-cycle-regulated activation of Akt kinase by phosphorylation at its carboxyl terminus. en. *Nature* **508**, 541–545 (Apr. 2014) (cit. on pp. 44, 66).
47. Parang, K., Till, J. H., Ablooglu, A. J., Kohanski, R. A., Hubbard, S. R. & Cole, P. A. Mechanism-based design of a protein kinase inhibitor. en. *Nat. Struct. Biol.* **8**, 37–41 (Jan. 2001) (cit. on p. 44).
48. Cheng, K.-Y., Noble, M. E. M., Skamnaki, V., Brown, N. R., Lowe, E. D., Kontogiannis, L., Shen, K., Cole, P. A., Siligardi, G. & Johnson, L. N. The role of the phospho-CDK2/cyclin A recruitment site in substrate recognition. en. *J. Biol. Chem.* **281**, 23167–23179 (Aug. 2006) (cit. on p. 44).
49. Mildvan, A. S. Mechanisms of signaling and related enzymes. en. *Proteins* **29**, 401–416 (Dec. 1997) (cit. on p. 44).

50. Shi, Y., Fernandez-Martinez, J., Tjioe, E., Pellarin, R., Kim, S. J., Williams, R., Schneidman-Duhovny, D., Sali, A., Rout, M. P. & Chait, B. T. Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. en. *Mol. Cell. Proteomics* **13**, 2927–2943 (Nov. 2014) (cit. on pp. 47, 85).
51. Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B. & Sali, A. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. en. *PLoS Biol.* **10**, e1001244 (Jan. 2012) (cit. on pp. 48, 57, 86).
52. Liu, Q., Chang, J. W., Wang, J., Kang, S. A., Thoreen, C. C., Markhard, A., Hur, W., Zhang, J., Sim, T., Sabatini, D. M. & Gray, N. S. Discovery of 1-(4-(4-propionylpiperazin-1-yl)-3-(trifluoromethyl)phenyl)-9-(quinolin-3-yl)benzo[h][1,6]naphthyridin-2(1H)-one as a highly potent, selective mammalian target of rapamycin (mTOR) inhibitor for the treatment of cancer. en. *J. Med. Chem.* **53**, 7146–7155 (Oct. 2010) (cit. on p. 48).
53. Yang, H., Rudge, D. G., Koos, J. D., Vaidialingam, B., Yang, H. J. & Pavletich, N. P. mTOR kinase structure, mechanism and regulation. en. *Nature* **497**, 217–223 (May 2013) (cit. on pp. 48, 49).
54. Liu, Q., Wang, J., Kang, S. A., Thoreen, C. C., Hur, W., Ahmed, T., Sabatini, D. M. & Gray, N. S. Discovery of 9-(6-Aminopyridin-3-yl)-1-(3-(trifluoromethyl)phenyl)benzo[h][1,6]naphthyridin-2(1H)-one (Torin2) as a Potent, Selective, and Orally Available Mammalian Target of Rapamycin (mTOR) Inhibitor for Treatment of Cancer. *J. Med. Chem.* **54**, 1473–1480 (Mar. 2011) (cit. on p. 48).
55. Wu, W.-I., Voegtli, W. C., Sturgis, H. L., Dizon, F. P., Vigers, G. P. A. & Brandhuber, B. J. Crystal structure of human AKT1 with an allosteric inhibitor reveals a new mode of kinase inhibition. en. *PLoS One* **5**, e12913 (Sept. 2010) (cit. on pp. 49, 59).

56. Takimura, T., Kamata, K., Fukasawa, K., Ohsawa, H., Komatani, H., Yoshizumi, T., Takahashi, I., Kotani, H. & Iwasawa, Y. Structures of the PKC-iota kinase domain in its ATP-bound and apo forms reveal defined structures of residues 533-551 in the C-terminal tail and their roles in ATP binding. en. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 577–583 (May 2010) (cit. on p. 53).
57. Hu, G., Katuwawala, A., Wang, K., Wu, Z., Ghadermarzi, S., Gao, J. & Kurgan, L. fIDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. en. *Nat. Commun.* **12**, 1–8 (July 2021) (cit. on p. 59).
58. Kovalski, J. R., Bhaduri, A., Zehnder, A. M., Neela, P. H., Che, Y., Wozniak, G. G. & Khavari, P. A. The Functional Proximal Proteome of Oncogenic Ras Includes mTORC2. en. *Mol. Cell* **73**, 830–844.e12 (Feb. 2019) (cit. on p. 65).
59. Johnson, J. L., Yaron, T. M., Huntsman, E. M., Kerelsky, A., Song, J., Regev, A., Lin, T.-Y., Liberatore, K., Cizin, D. M., Cohen, B. M., Vasan, N., Ma, Y., Krismer, K., Robles, J. T., van de Kooij, B., van Vlimmeren, A. E., Andrée-Busch, N., Käufer, N. F., Dorovkov, M. V., Ryazanov, A. G., Takagi, Y., Kasthuber, E. R., Goncalves, M. D., Hopkins, B. D., Elemento, O., Taatjes, D. J., Maucuer, A., Yamashita, A., Degterev, A., Uduman, M., Lu, J., Landry, S. D., Zhang, B., Cossentino, I., Linding, R., Blenis, J., Hornbeck, P. V., Turk, B. E., Yaffe, M. B. & Cantley, L. C. An atlas of substrate specificities for the human serine/threonine kinome. en. *Nature* **613**, 759–766 (Jan. 2023) (cit. on p. 66).
60. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. en. *Nucleic Acids Res.* **32**, 1792–1797 (Mar. 2004) (cit. on p. 68).
61. Cristea, I. M. & Chait, B. T. Conjugation of magnetic beads for immunopurification of protein complexes. en. *Cold Spring Harb. Protoc.* **2011**, db.prot5610 (May 2011) (cit. on p. 71).

62. Chen, B., Gilbert, L. A., Cimini, B. A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E. H., Weissman, J. S., Qi, L. S. & Huang, B. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. en. *Cell* **155**, 1479–1491 (Dec. 2013) (cit. on p. 72).
63. Wang, W. & Malcolm, B. A. Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange Site-Directed Mutagenesis. en. *Biotechniques* **26**, 680–682 (Apr. 1999) (cit. on p. 72).
64. Mitchell, L. A., Cai, Y., Taylor, M., Noronha, A. M., Chuang, J., Dai, L. & Boeke, J. D. Multichange isothermal mutagenesis: a new strategy for multiple site-directed mutations in plasmid DNA. en. *ACS Synth. Biol.* **2**, 473–477 (Aug. 2013) (cit. on p. 72).
65. Rogala, K. B., Gu, X., Kedir, J. F., Abu-Remaileh, M., Bianchi, L. F., Bottino, A. M. S., Dueholm, R., Niehaus, A., Overwijn, D., Fils, A.-C. P., Zhou, S. X., Leary, D., Laqtom, N. N., Brignole, E. J. & Sabatini, D. M. Structural basis for the docking of mTORC1 on the lysosomal surface. en. *Science* **366**, 468–475 (Oct. 2019) (cit. on p. 72).
66. Fitzgerald, D. J., Berger, P., Schaffitzel, C., Yamada, K., Richmond, T. J. & Berger, I. Protein complex expression by using multigene baculoviral vectors. en. *Nat. Methods* **3**, 1021–1032 (Nov. 2006) (cit. on p. 72).
67. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. en. *PLoS One* **3**, e3647 (Nov. 2008) (cit. on p. 72).
68. Weissmann, F., Petzold, G., VanderLinden, R., Huis In 't Veld, P. J., Brown, N. G., Lampert, F., Westermann, S., Stark, H., Schulman, B. A. & Peters, J.-M. biGBac enables rapid gene assembly for the expression of large multisubunit protein complexes. en. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E2564–9 (May 2016) (cit. on p. 73).
69. Andersen, K. R., Leksa, N. C. & Schwartz, T. U. Optimized E. coli expression strain LOBSTR eliminates common contaminants from His-tag purification. en. *Proteins* **81**,

- 1857–1861 (Nov. 2013) (cit. on p. 73).
70. Lau, Y.-T. K., Baytshtok, V., Howard, T. A., Fiala, B. M., Johnson, J. M., Carter, L. P., Baker, D., Lima, C. D. & Bahl, C. D. Discovery and engineering of enhanced SUMO protease enzymes. en. *J. Biol. Chem.* **293**, 13224–13233 (Aug. 2018) (cit. on p. 73).
 71. Backliwal, G., Hildinger, M., Kuettel, I., Delegrange, F., Hacker, D. L. & Wurm, F. M. Valproic acid: a viable alternative to sodium butyrate for enhancing protein expression in mammalian cell cultures. en. *Biotechnol. Bioeng.* **101**, 182–189 (Sept. 2008) (cit. on p. 74).
 72. Huang, B. X., Akbar, M., Kevala, K. & Kim, H.-Y. Phosphatidylserine is a critical modulator for Akt activation. en. *J. Cell Biol.* **192**, 979–992 (Mar. 2011) (cit. on p. 75).
 73. Ficarro, S. B., Zhang, Y., Lu, Y., Moghimi, A. R., Askenazi, M., Hyatt, E., Smith, E. D., Boyer, L., Schlaeger, T. M., Luckey, C. J. & Marto, J. A. Improved Electrospray Ionization Efficiency Compensates for Diminished Chromatographic Resolution and Enables Proteomics Analysis of Tyrosine Signaling in Embryonic Stem Cells. *Anal. Chem.* **81**, 3440–3447 (May 2009) (cit. on p. 79).
 74. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. en. *Nat. Methods* **14**, 513–520 (Apr. 2017) (cit. on p. 79).
 75. Ficarro, S. B., Alexander, W. M. & Marto, J. A. mzStudio: A Dynamic Digital Canvas for User-Driven Interrogation of Mass Spectrometry Data. en. *Proteomes* **5** (Aug. 2017) (cit. on p. 79).
 76. Shen, K. & Cole, P. A. Conversion of a Tyrosine Kinase Protein Substrate to a High Affinity Ligand by ATP Linkage. *J. Am. Chem. Soc.* **125**, 16172–16173 (Dec. 2003) (cit. on p. 81).

77. Chen, Z.-L., Meng, J.-M., Cao, Y., Yin, J.-L., Fang, R.-Q., Fan, S.-B., Liu, C., Zeng, W.-F., Ding, Y.-H., Tan, D., Wu, L., Zhou, W.-J., Chi, H., Sun, R.-X., Dong, M.-Q. & He, S.-M. A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. en. *Nat. Commun.* **10**, 1–12 (July 2019) (cit. on p. 85).
78. Xiang, Y., Sang, Z., Bitton, L., Xu, J., Liu, Y., Schneidman-Duhovny, D. & Shi, Y. Integrative proteomics identifies thousands of distinct, multi-epitope, and high-affinity nanobodies. en. *Cell Syst* **12**, 220–234.e9 (Mar. 2021) (cit. on p. 85).
79. Xiang, Y., Shen, Z. & Shi, Y. Chemical Cross-Linking and Mass Spectrometric Analysis of the Endogenous Yeast Exosome Complexes. en. *Methods Mol. Biol.* **2062**, 383–400 (2020) (cit. on pp. 85, 86).
80. Shi, Y., Pellarin, R., Fridy, P. C., Fernandez-Martinez, J., Thompson, M. K., Li, Y., Wang, Q. J., Sali, A., Rout, M. P. & Chait, B. T. A strategy for dissecting the architectures of native macromolecular assemblies. en. *Nat. Methods* **12**, 1135–1138 (Oct. 2015) (cit. on pp. 85, 86).
81. Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. en. *Cell* **177**, 1384–1403 (May 2019) (cit. on p. 86).
82. Valenstein, M. L., Rogala, K. B., Lalgudi, P. V., Brignole, E. J., Gu, X., Saxton, R. A., Chantranupong, L., Kolibius, J., Quast, J.-P. & Sabatini, D. M. Structure of the nutrient-sensing hub GATOR2. en. *Nature* **607**, 610–616 (July 2022) (cit. on pp. 86, 93).
83. Zheng, S. Q., Palovcak, E., Armache, J.-P., Verba, K. A., Cheng, Y. & Agard, D. A. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. en. *Nat. Methods* **14**, 331–332 (Apr. 2017) (cit. on p. 93).
84. Zivanov, J., Nakane, T. & Scheres, S. H. W. A Bayesian approach to beam-induced motion correction in cryo-EM single-particle analysis. en. *IUCrJ* **6**, 5–17 (Jan. 2019)

- (cit. on p. 93).
85. Zhang, K. Gctf: Real-time CTF determination and correction. en. *J. Struct. Biol.* **193**, 1–12 (Jan. 2016) (cit. on p. 93).
 86. Wagner, T., Merino, F., Stabrin, M., Moriya, T., Antoni, C., Apelbaum, A., Hagel, P., Sitsel, O., Raisch, T., Prumbaum, D., Quentin, D., Roderer, D., Tacke, S., Siebolds, B., Schubert, E., Shaikh, T. R., Lill, P., Gatsogiannis, C. & Raunser, S. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. en. *Communications Biology* **2**, 1–13 (June 2019) (cit. on p. 93).
 87. Zivanov, J., Nakane, T., Forsberg, B. O., Kimanius, D., Hagen, W. J. H., Lindahl, E. & Scheres, S. H. W. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7**, e42166 (Nov. 2018) (cit. on p. 93).
 88. Zivanov, J., Nakane, T. & Scheres, S. H. W. Estimation of high-order aberrations and anisotropic magnification from cryo-EM data sets in RELION-3.1. en. *IUCrJ* **7**, 253–267 (Mar. 2020) (cit. on p. 94).
 89. Scheres, S. H. Beam-induced motion correction for sub-megadalton cryo-EM particles. en. *Elife* **3**, e03665 (Aug. 2014) (cit. on p. 94).
 90. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. en. *Nat. Methods* **14**, 290–296 (Mar. 2017) (cit. on p. 94).
 91. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. en. *J. Comput. Chem.* **25**, 1605–1612 (Oct. 2004) (cit. on p. 94).
 92. Casañal, A., Lohkamp, B. & Emsley, P. Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. en. *Protein Sci.* **29**, 1069–1078 (Apr. 2020) (cit. on p. 94).

93. Afonine, P. V., Poon, B. K., Read, R. J., Sobolev, O. V., Terwilliger, T. C., Urzhuntsev, A. & Adams, P. D. Real-space refinement in PHENIX for cryo-EM and crystallography. en. *Acta Crystallogr D Struct Biol* **74**, 531–544 (June 2018) (cit. on p. 94).
94. Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H. & Ferrin, T. E. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. en. *Protein Sci.* **27**, 14–25 (Jan. 2018) (cit. on p. 95).

Chapter 4

Bayesian multi-state multi-condition modeling of a protein structure from X-ray crystallography

4.1 Abstract

A model of a protein structure at atomic resolution is key to rationalizing and predicting its biological function. Many such models are computed from a diffraction pattern from X-ray crystallography. Despite the protein crystal containing billions of protein molecules that independently sample the energy landscape during data collection, most models computed from X-ray data depict a single set of atomic coordinates. A model with multiple sets of atomic coordinates (multi-state) may improve the satisfaction of the X-ray data and is a more accurate, precise, and informative depiction of the protein. However, computing a multi-state model is challenging on account of a low data-to-parameter ratio. X-ray datasets collected for the same system under distinct experimental conditions (*eg*, temperature) may provide additional observations, thereby improving the data-to-parameter ratio. Here, we develop, benchmark, and illustrate MultiXray: Bayesian multi-state multi-condition modeling for

X-ray crystallography. The input information is J X-ray datasets collected under distinct conditions and a molecular mechanics force field. The model representation is N independent atomic models of the protein structure (states) and the weight of each state under each condition. A Bayesian posterior model density quantifies the match of the model with all X-ray datasets and molecular mechanics. A sample of models is drawn from the posterior model density using molecular dynamics simulations. We benchmark MultiXray on simulated X-ray data and analyze the impact of additional states and conditions on the scoring function and model search. We illustrate MultiXray on temperature-dependent X-ray datasets collected for SARS-CoV-2 M^{pro} and compute multi-state multi-condition models that improve the R^{free} relative to the PDB model by up to 0.05. MultiXray is implemented in our open-source *Integrative Modeling Platform*, relying on integration with *Phenix*, thus making it easily applicable to other systems.

4.2 Introduction

4.2.1 A protein crystal is a heterogeneous mix of structural states

X-ray crystallography is an important experimental technique for obtaining structural models of proteins at atomic resolution. In X-ray crystallography, the protein crystal contains between $10^6 - 10^{15}$ copies of the protein¹⁻³. Due to the high solvent content of the crystal, protein molecules within the crystal can fluctuate nearly independently throughout data collection and adopt distinct structural states based on the energy landscape⁴⁻⁶. Despite the resulting structural heterogeneity, a majority of models computed from X-ray datasets describe a single structural state with Gaussian isotropic B-factors⁷. Fitting a B-factor to an X-ray dataset convolutes all sources of experimental uncertainty, including the heterogeneous mix of structural states, random thermal motion, and long-range crystallographic disorder^{8,9}. The inability of a B-factor to describe the heterogeneous mix of structural states found within the protein crystal contributes to the inability of protein models to satisfy an X-ray dataset

within its theoretical uncertainty^{10,11}. A model that depicts anharmonic conformational substates found in a protein crystal will improve the satisfaction of X-ray data and reveal the structural basis of important biological properties at atomic resolution, such as allosteric networks or hidden cavities for small molecule binding (cryptic pockets)^{12,13}.

4.2.2 Approaches to modeling a heterogenous mix of structural states

There are 2 approaches for computing models that depict multiple structural states from X-ray data. First, conformational substates can be captured by computing single-state models that independently satisfy the X-ray dataset, as is seen in the modeling of Nuclear Magnetic Resonance spectra¹⁴. For example, `phenix.ensemble_refinement` computes an ensemble from snapshots of a molecular simulation restrained by a time-averaged X-ray target function¹⁵. Such approaches may not find weakly occupied states in a reasonable amount of computation time, as they are dependent on overcoming potentially large barriers in the scoring function. Alternatively, a model can depict conformational substates by introducing additional structural variables³. All variables are then collectively fit against the X-ray data. The extent and detail of the additional degrees of freedom depend on the available computational power and data quality¹⁶. For example, `qFit-3` avoids introducing excessive structural parameters by representing side chains as ensembles of one or more rotameric states¹⁷. Methods that refine multiple fully parameterized atomic models have been limited to systems that diffract to ultra-high resolution^{6,18-21}.

4.2.3 Multi-condition crystallography

Often, multiple X-ray datasets are collected for the same system under distinct experimental conditions. One example is multi-temperature crystallography, where data collection is performed at temperatures from cryogenic to near-physiological²². Data collection at higher

temperatures has been shown to dramatically modify protein dynamics within the protein crystal^{23–25}. More recently, models computed from higher temperature datasets have revealed a fuller set of structural states within the protein’s energy landscape at atomic resolution^{26–28}. Comparisons of models of the same system computed at distinct temperatures show similar structural states but shifted thermodynamic equilibrium^{26,29,30}. Therefore, multiple X-ray datasets under distinct conditions containing mutual structural information could increase the data-to-parameter ratio and inform a more accurate, precise, and complete multi-state model.

4.2.4 Computing a multi-state multi-condition model

Here, we seek to compute a multi-state model from multiple X-ray diffraction patterns collected under distinct conditions (**Fig. 4.1a**). We can fit a multi-state model using multiple X-ray datasets without needing ultra-high-resolution X-ray data. We accommodate for the thermodynamic shift of distinct experimental conditions by computing each state’s weight under each condition. In other words, we assume the sets of structural states are consistent across all experimental conditions with varied weights, allowing a single multi-state model to be informed by all X-ray datasets and massively boosting the data-to-parameter ratio. The assumption of a consistent set of structural states under all conditions is not limiting because the structural states may include sparsely populated or completely unpopulated states. We formulate a Bayesian posterior model density for the multi-state multi-condition model. We draw a sample from the Bayesian posterior model density using a molecular dynamics algorithm where all structural states are jointly restrained by the satisfaction of all X-ray datasets in addition to molecular mechanics. We benchmarked the method using synthetic X-ray datasets simulated from multi-state SARS-CoV-2 M^{pro} structural models and showed that multiple structural states are needed to satisfy the data and that by including multiple X-ray datasets, all datasets are better satisfied individually. We illustrate our method to compute a multi-state model of the SARS-CoV-2 viral target M^{pro} from multi-temperature

X-ray under 6 conditions and show that all datasets are satisfied better by increasing the number of states and X-ray datasets.

4.3 Methods

4.3.1 Overview of modeling method

A model is a depiction of our knowledge about a system or process. We wish to inform the model based on the input information, which generally includes experimental data and prior models (*eg*, physical theories and statistical preferences). A model can then rationalize past observations and predict future ones. Modeling is the search for a set of models consistent with the input information. We aim to find all models that satisfy the input information, reflecting the uncertainty of the input information and the modeling process. It is convenient to divide modeling into three steps: (i) specifying all model variables (representation), (ii) ranking alternative models by their agreement with the input information (scoring), and (iii) generating a sample of good-scoring models (sampling). A model should be validated before being interpreted, often by examining the satisfaction of input information withheld from the modeling process. Multiple iterations of gathering input information, modeling, and validation are often necessary to compute a sufficiently precise model^{31,32}.

4.3.2 Input information

The input information may be used to inform any step of modeling: representation, scoring, and sampling. Here, we use the following input information: 1) X-ray datasets D_1, \dots, D_J collected under distinct experimental conditions (*eg*, at distinct temperatures) and corresponding PDB models. Each X-ray dataset is a set of observed structure factor amplitudes $\{|F_{\vec{s}}^O|\}_{\vec{s} \in S}$ indexed by a scattering vector \vec{s} . The X-ray datasets are used in scoring. The PDB model is used in scoring and sampling. 2) The CHARMM19 force field parameters³³. The force field parameters are used in scoring to evaluate the prior.

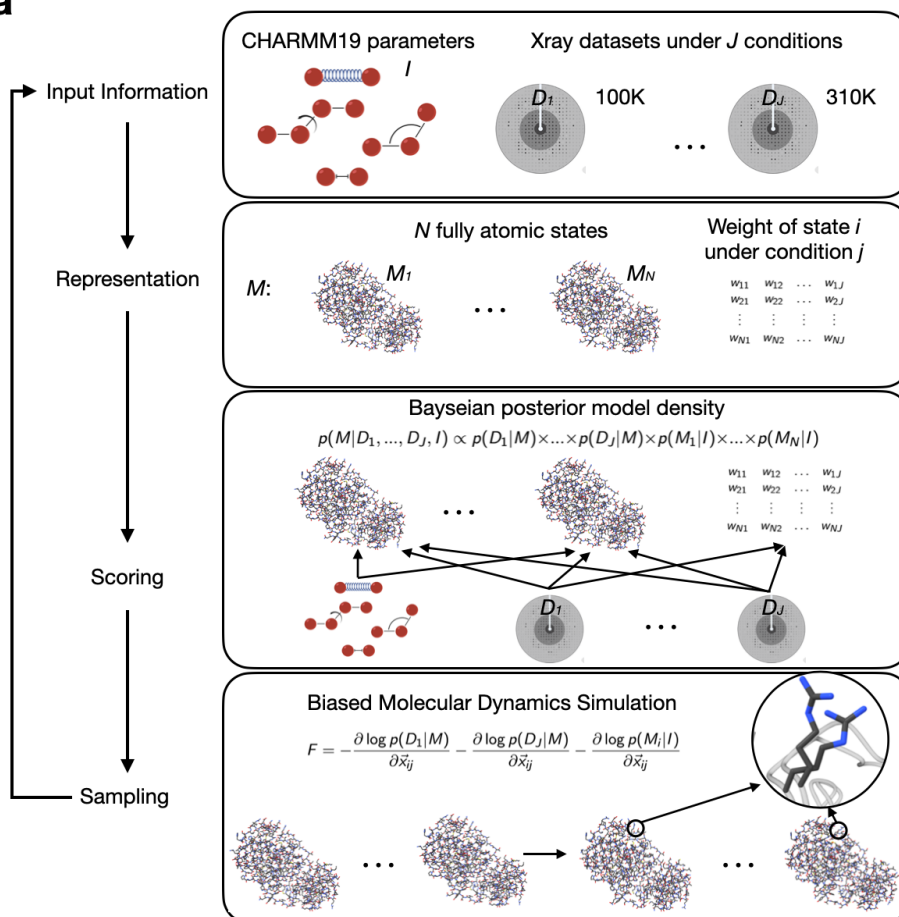
Figure 1**a**

Figure 4.1: a, modeling can be framed as a model search given some input information. Here, the input information is the CHARMM19 force field parameters and J X-ray datasets collected for the same system under distinct experimental conditions (*eg*, temperature). The representation is the N atomic states containing all the heavy atoms of the system along with the weight matrix that parameterizes the weight of each state under each condition. All states and the weight matrix are scored collectively against each X-ray dataset. Each state is individually scored against the molecular mechanics force field. A sample is drawn from the posterior model density using molecular dynamics. All states are initialized by a starting structure and the force on the atoms is computed from the satisfaction of all X-ray datasets along with the molecular mechanics, and the weights are stochastically sampled.

4.3.3 Representation

The representation defines the degrees of freedom whose values are to be determined by modeling. The multi-state multi-condition model M includes the following variables: (i) N states M_i (N is selected before modeling). Each state is an independently parameterized model of a protein structure. The number and type of atoms (composition) of each state are based on the PDB models (Input Information). Only non-hydrogen atoms from the protein are included, however, the representation may be generalized to include additional atoms/molecules (eg, hydrogen, solvent, ion, and ligand). As we are computing a model containing a set of discrete structural states, all atomic B-factors are set to 15 and atomic occupancies to 1. (ii) Weight matrix $W_{N \times J}$ containing the weight of M_i under condition j , w_{ij} . The weight matrix is subject to the constraint that the sum of the weights of all states under each condition (*ie*, columns) is 1. (iii) Nuisance variables, taken from established modeling methods, to improve the fit of the X-ray datasets by the model^{34,35}. For clarity, we do not include nuisance parameters in the notation.

4.3.4 Scoring

Bayesian posterior model density

A scoring function assesses the match of the model to the input information. Bayesian inference is one approach to assess this fit. In Bayesian inference, the posterior model density $p(M|D, I)$ describes the relationship between model M , prior information I , and data D . The posterior model density may be factored into a prior $p(M|I)$ that quantifies the state of knowledge prior to the observation of data and a likelihood $p(D|M, I)$ that quantifies what is learned from the observation of data³⁶:

$$p(M|D, I) = p(D|M, I) \times p(M|I) \quad (4.1)$$

It is often helpful to decompose the likelihood into a forward model $f(M)$ that simulates

a noiseless data observation and a noise model $N(f(M); D, \sigma)$ that quantifies the difference between the observed and simulated data based on a model of the experiment's noise parameterized by σ ³¹.

Joint-likelihood

The likelihood of observing all X-ray diffraction patterns from a multi-state multi-condition model is:

$$p(D_1, \dots, D_J | M_1, \dots, M_N, W_{N \times J}) \quad (4.2)$$

The likelihood may be factored into the likelihood of observing each condition's dataset:

$$p(D_1, \dots, D_J | M_1, \dots, M_N, W_{N \times J}) = \prod_{j=1}^J p(D_j | M_1, \dots, M_N, W_{N \times J}) \quad (4.3)$$

The likelihood for each condition's dataset depends only on the set of states and the weights representing the states under the corresponding condition (matrix column):

$$p(D_j | M_1, \dots, M_N, W_{N \times J}) = p(D_j | M_1, \dots, M_N, w_{1j}, \dots, w_{Nj}) \quad (4.4)$$

The likelihood of an X-ray dataset is factored into the likelihood of observing each structure factor amplitude $F_{\vec{s}}^O$ ³⁵:

$$p(D_j | M_1, \dots, M_N, w_{1j}, \dots, W_{Nj}) = \prod_{\vec{s} \in S} p(F_{\vec{s}}^O | M_1, \dots, M_N, w_{1j}, \dots, W_{Nj}) \quad (4.5)$$

The likelihood of a structure factor amplitude given a multi-state multi-conditional model is a noise model that quantifies the difference between the observed structure factor amplitude $F_{\vec{s}}^O$ and the model structure factor amplitude $F_{\vec{s}}^M$ simulated by the forward model from the multi-condition multi-state model³⁵:

$$p(F_{\vec{s}}^O | M_1, \dots, M_N, w_{1j}, \dots, W_{Nj}, I) = p(F_{\vec{s}}^O | F_{\vec{s}}^M) \quad (4.6)$$

The noise model is based on the assumption that the real and imaginary parts of a complex structure factor \mathbf{F} are sampled from a two-dimensional Gaussian with variance $\epsilon\beta$ and scale parameter $\alpha_{\vec{s}}$ ³⁵:

$$p(\text{Re}\mathbf{F}, \text{Im}\mathbf{F}) = \frac{1}{\pi\epsilon\beta_{\vec{s}}} \exp\left(-(\text{Re}\mathbf{F} - \alpha_{\vec{s}}F^M \cos \phi^M)^2 + (\text{Im}\mathbf{F} - \alpha_{\vec{s}}F^M \sin \phi^M)^2\right) / \epsilon\beta_{\vec{s}} \quad (4.7)$$

The likelihood for a single structure factor amplitude is³⁵:

$$p(F_{\vec{s}}^O | F_{\vec{s}}^M, \alpha_{\vec{s}}, \beta_{\vec{s}}) = \begin{cases} \frac{2F_{\vec{s}}^O}{\epsilon\beta_{\vec{s}}} \exp\left(-\frac{F_{\vec{s}}^{O2} + \alpha_{\vec{s}}^2 F_{\vec{s}}^{M2}}{\epsilon\beta_{\vec{s}}}\right) I_0\left(\frac{2\alpha_{\vec{s}}F_{\vec{s}}^O F_{\vec{s}}^M}{2\epsilon\beta_{\vec{s}}}\right) \\ \dots \text{if } \vec{s} \text{ acentric} \\ \left(\frac{2}{\epsilon\pi\beta_{\vec{s}}}\right)^{1/2} \exp\left(-\frac{F_{\vec{s}}^{O2} + \alpha_{\vec{s}}^2 F_{\vec{s}}^{M2}}{2\epsilon\beta_{\vec{s}}}\right) \cosh\left(\frac{2\alpha_{\vec{s}}F_{\vec{s}}^O F_{\vec{s}}^M}{2\epsilon\beta_{\vec{s}}}\right) \\ \dots \text{if } \vec{s} \text{ centric} \end{cases} \quad (4.8)$$

where I_0 is the hyperbolic Bessel function of the first kind ($\alpha=0$) and \cosh is the hyperbolic cosine function.

Forward model

The forward model $F_{\vec{s}}^M$ is the magnitude of the complex model structure factor $\mathbf{F}_{\vec{s}}^M$:

$$F_{\vec{s}}^M = |\mathbf{F}_{\vec{s}}^M| \quad (4.9)$$

The model structure factor is the sum of the protein $\mathbf{F}_{\vec{s}}^C$ and bulk solvent structure factor $\mathbf{F}_{\vec{s}}^B$ with an overall scale factor k_{total} and bulk solvent scaling factor k_{mask} ³⁷.

$$\mathbf{F}_{\vec{s}}^M = k_{\text{total}}(\mathbf{F}_{\vec{s}}^C + k_{\text{mask}}\mathbf{F}_{\vec{s}}^B) \quad (4.10)$$

It is well-established how to compute the scattering from a model of a protein structure assuming a perfect crystal³⁸. Here, we compute the scattering of a model of a protein structure with multiple weighted structural states. It is important to keep in mind that different structures and positions of molecules in a crystal can result in the same X-ray diffraction data, even when the data are noiseless, complete, and collected instantaneously. We show that the scattering in the 5 cases is equivalent (**Fig. 4.2ab**).

The degeneracy between (i) and (ii) relates the scattering of a multi-state crystal to a set of single-state crystals. The structure factor for a multi-state model may be computed from the structure factor $\mathbf{F}_{i\vec{s}}$:

$$\mathbf{F}_{\vec{s}} = \sum_{i=1}^{\text{states}} \mathbf{F}_{i\vec{s}} \quad (4.11)$$

The degeneracy of (i), (iv), and (v) complicates the interpretation of a match between a multi-state model and an X-ray dataset. For example, imagine we compute an N -state model with N rotamers for residues A and B each. Without exhaustive sampling of all possible multi-state models, we do not know if the rotamer pair always occurs in the same molecule or is one of many possible combinations of the rotameric states in multiple molecules. In other words, we cannot rule out any combination of the rotameric states of A and B in one molecule. As an aside, the degeneracy of the X-ray data might be broken by using additional information for modeling; for example, a potential energy function may rule out combinations of rotameric states that result in overlapping atoms.

Prior

The prior for the multi-state multi-condition model is the product of a prior for each state:

$$p(M|I) = \prod_i^N p(M_i|I) \quad (4.12)$$

The prior for a state is the Boltzmann distribution corresponding to the potential energy

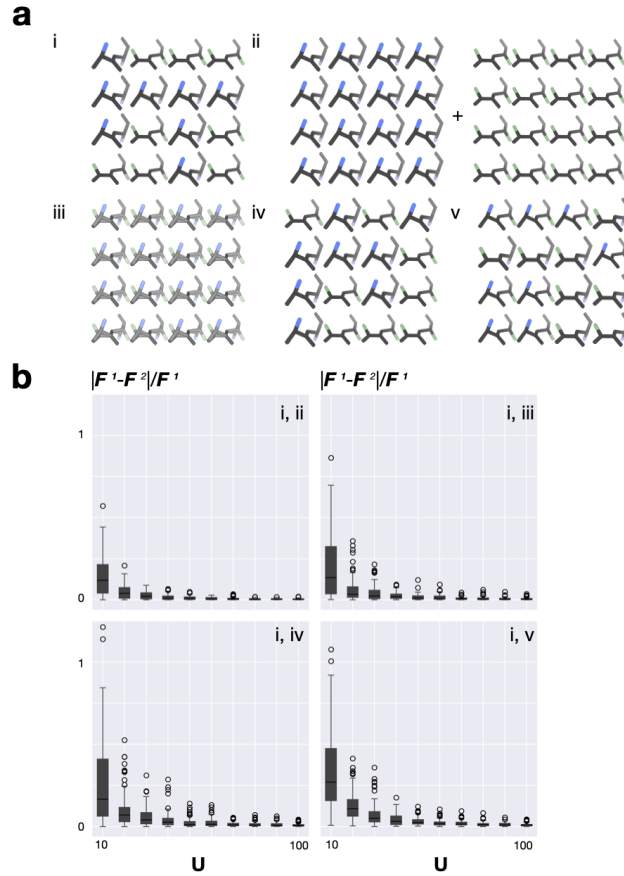
Figure 2

Figure 4.2: **a**, five cases where the scattering is degenerate based on a multi-state model containing 2 states M_1 and M_2 with w_1 and w_2 . (i) Scattering from a heterogeneous crystal containing unit cells occupied by either M_1 or M_2 with probability w_1 and w_2 , respectively. (ii) A weighted combination of the scattering from homogeneous crystals containing identical unit cells with a single copy of M_1 and M_2 respectively. (iii) Scattering from a homogeneous crystal that contains the occupancy-weighted average of M_1 and M_2 (*ie*, the atomic occupancy of an atom from M_1 and M_2 is w_1 and w_2 , respectively). (iv) Scattering from a heterogeneous crystal the same as I where the states are re-indexed (*ie*, M_1 becomes M_2 and w_1 becomes w_2 and vice versa). (v) If $w_1 = w_2$, the scattering from a heterogeneous crystal except the states contain an intermixing of the atomic positions of M_1 and M_2 . **b**, as the number of unit cells goes to infinity, the reciprocal lattice points simulated from a crystal in ii, iii, iv, and v converge to those for the crystal in i. We construct 100 random reference crystals with U^3 unit cells where each unit cell is randomly populated by M_1 or M_2 with probability w_1 , w_2 respectively. M_1 , M_2 each contain 2 scatterers with random atomic position and scattering type. We compute the reciprocal lattice points up to a maximum miller index of $hkl=(3,3,3)$ via a discrete Fourier Transform (DFT). For (ii), (iii), (iv), and (v), we construct 100 random crystals as described and compute the same reciprocal lattice points. For each reciprocal lattice point, we compute the Euclidean distance for each structure factor to the reference structure factor U from 10 to 100. As U grows larger, the average Euclidean distance between the structure factors and reference structure factors converges to 0.

of the state, including terms for bond lengths b , bond angles θ , dihedral angles ϕ , improper dihedral angles w , and non-bonded interactions r_{ij} :

$$p(M_i|I) = -\frac{1}{Q} \left(\exp \left\{ \sum_{\text{bnd}} k_b (b - b_0)^2 + \sum_{\text{ang}} k_\theta (\theta - \theta_0)^2 + \sum_{\text{dih}} k_\phi [1 + \cos(n\phi - \delta)] \right\} \right. \quad (4.13)$$

$$\left. + \exp \left\{ \sum_{\text{imp}} k_\omega (\omega - \omega_0)^2 + \sum_{\text{nb}} \left(\epsilon \left[\left(\frac{R_{\text{min}_{ij}}}{r_{ij}} \right)^{12} - \left(\frac{R_{\text{min}_{ij}}}{r_{ij}} \right)^6 \right] \right) \right\} \right)$$

where the parameters k_b , b_0 , k_θ , θ_0 , k_ϕ , δ , k_ω , w_0 , ϵ , and $R_{\text{min}_{ij}}$ are obtained from the CHARMM19 molecular mechanics force field³³ and Q is the partition function, which is ignored in practice.

Scoring function

A Bayesian scoring function for ranking alternative solutions based on the available information is:

$$S = -\log p(D|M, I) - \log p(M|I) \quad (4.14)$$

4.3.5 Sampling

The purpose of sampling is to find all models consistent with the input information. In Bayesian modeling, this is achieved by computing a sufficiently converged estimate of $p(D|M, I)$.

Atomic coordinates

The atomic positions for each state are sampled via molecular dynamics (MD). The atomic positions of a state are initialized to one of the input PDB models. The velocity of each atomic coordinate is sampled from a Boltzmann distribution with $T = 300\text{K}$. Because of the

likelihood, the force on the atoms is non-conservative, and a thermostat is used to maintain a simulation temperature of 300K¹⁵. We do not include explicit solvent molecules in the simulations, but the likelihood restrains each state to well-folded conformations. To account for discrepancies in the X-ray datasets, we reset the center of the mass to the center of mass of the corresponding PDB model before scoring against an X-ray dataset. The force computed on each atom is the partial derivative of the scoring function with respect to the atomic position of atom \vec{x} in state i , \vec{x}_i :

$$\begin{aligned} \frac{\partial S}{\partial \vec{x}_i} = & - \left(\frac{\partial \log p(D_1|M_1, \dots, M_N, w_{11}, \dots, w_{N1}, I)}{\partial \vec{x}_i} + \dots \right. \\ & + \frac{\partial \log p(D_J|M_1, \dots, M_N, w_{1J}, \dots, w_{NJ}, I)}{\partial \vec{x}_i} \\ & \left. - \frac{\partial \log p(M_i|I)}{\partial \vec{x}_i} \right) \end{aligned} \quad (4.15)$$

To calculate $\frac{\partial p(D_j|M_1, \dots, M_N, w_{1j}, \dots, w_{Nj}, I)}{\partial \vec{x}_i}$ more efficiently, a quadratic approximation of $p(F_{\vec{s}}^O|F_{\vec{s}}^M)$ is used³⁵. By computing the partial derivative of all atomic positions for a multi-state model, the atoms are restrained to satisfy a heterogeneous X-ray dataset collectively. Furthermore, all atoms of all states are restrained to satisfy all J X-ray datasets collectively.

Because the magnitudes of the derivative of the likelihood and prior can vary significantly, we introduce 2 scaling parameters w_{xray} and w_{auto} :

$$\begin{aligned} \frac{\partial S}{\partial \vec{x}_i} = & -w_{\text{xray}} \left(w_{\text{auto}}^0 \frac{\partial \log p(D_1|M_1, \dots, M_N, w_{11}, \dots, w_{N1}, I)}{\partial \vec{x}_i} + \dots \right. \\ & \left. + w_{\text{auto}}^J \frac{\partial \log p(D_J|M_1, \dots, M_N, w_{1J}, \dots, w_{NJ}, I)}{\partial \vec{x}_i} \right) \\ & - \frac{\partial \log p(M_i|I)}{\partial \vec{x}_i} \end{aligned} \quad (4.16)$$

w_{auto}^j is computed automatically to ensure the average force on the atoms from the likelihood of D_j is approximately equal to that from the prior²¹. To account for the presence

of multiple X-ray datasets, w_{xray} is selected before modeling. We select $w_{\text{xray}} \in \{2^{-i}; -5 \leq i \leq 5\}$ by computing a trial sample using each w_{xray} and identifying the w_{xray} that results in a sample that best satisfies the data (R^{free}) after filtering out w_{xray} that produce non-physiological conformations (**Fig. 4.3c**).

We perform a short refinement minimization of $p(M_i|I)$ for each state of each model.

Weight matrix

For every 100 steps of sampling of the atomic positions, we propose 10 sets of weights per condition. A proposal weight is accepted for a condition if it improves the likelihood for the corresponding X-ray dataset. Each proposal is sampled from a Normal distribution with a mean equal to the current weight and a standard deviation of 0.05 and normalized.

Nuisance parameters

The nuisance variables are not stochastically sampled to enhance computational efficiency but optimized by minimizing the least squares residual between the model and the observed X-ray diffraction data³⁴.

4.3.6 Software availability

The software, input files, and output files are freely available at as part of our open-source *Integrative Modeling Platform* (<https://integrativemodeling.org/2.20.0/doc/manual/>)³⁹. The software relies on integration with *Phenix* (<https://phenix-online.org/>)³⁷, as described previously⁴⁰.

4.4 Results

4.4.1 Synthetic Benchmark

Framework for analyzing a modeling method

If the ground truth (native) is known, a modeling method may be evaluated by its ability to find the native as the best-scoring model. Assuming all models are enumerated with sufficient precision, the only relevant feature of the scoring function is the native being the global minimum of the scoring function (global minimum accuracy) (**Fig. 4.3a1, a2**). However, in practice, all models cannot be enumerated with sufficient precision due to the high dimensionality of the search space. Therefore, we use stochastic search methods to find the global minimum. The efficacy of most search methods depends strongly on the scoring function having a funnel shape around the global minimum. The radius of convergence of the funnel is the area where meaningful gradients can be computed to assist convergence to the global minimum (the width of the funnel) (**Fig. 4.3a3**). The smoothness of the funnel is defined as the correlation between the score and model coordinates (**Fig. 4.3a4**). Most search methods benefit from a smoother funnel with a larger radius of convergence.

We benchmark our modeling method by comparing the 1-state 1-condition, 1-state 2-condition, 2-state 1-condition, and 2-state 2-condition scoring functions. Because the prior has been evaluated previously, we focus on the negative log-likelihood and refer to the negative log-likelihood as the scoring function. First, we estimate the global minimum of the scoring function. Second, we quantify the thoroughness of the search process by plotting the best score found as a function of the amount of sampling (convergence curve). Third, we interpret the differences between these plots for different scoring functions by considering their radius of convergences and smoothnesses. If the differences in the convergence curve cannot be explained by the differences in the radius of convergence and smoothness of the scoring function, then they must be a result of higher-dimensional characteristics of the landscape.

Figure 3

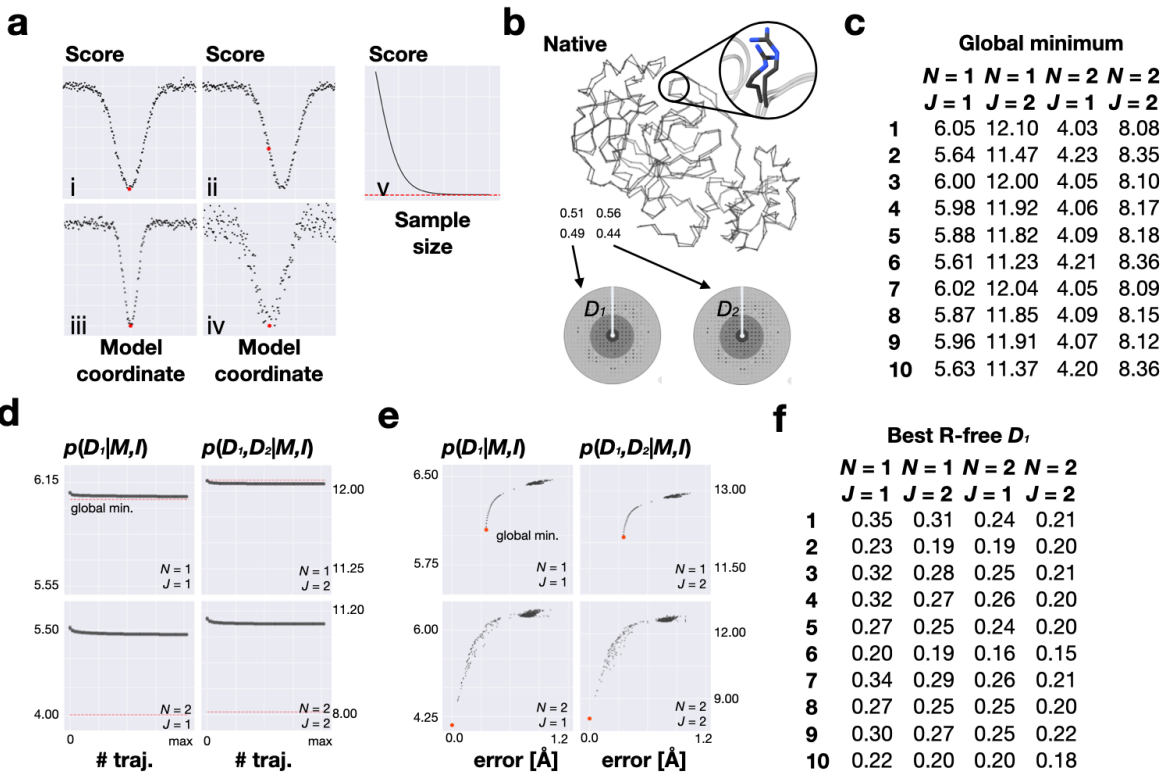


Figure 4.3: **a**, i-iv, depictions of a scoring function relating the model error to the score. The native is shown in red. ii) a scoring function where the native is not the scoring function global minimum (poor global minimum accuracy). iii) a scoring function with a small funnel (poor radius of convergence). iv) a scoring function where the correlation between the model error and score (poor smoothness). **v**) a depiction of the thoroughness of the search relating the best score found to the size of the sample. The search process finds the global minimum with a sufficiently large sample (red line). **b**, one of the native models used in the synthetic benchmark. The native model contains 2 states under 2 conditions. A native dataset is simulated from the native model using the forward model for each set of weights (D_1 and D_2). **c**, the thoroughness of the search process for an example native showing the best score found as a function of the amount of the sampling for all 4 scoring functions. The score is either the likelihood of D_1 (if $J = 1$) or the joint likelihood of D_1 and D_2 (if $J = 2$). The amount of sampling is quantified by the number of trajectories included in the sample from 1 to 250 (max). **d**, the multi-state multi-condition scoring function for N in 1, 2 and J in 1, 2 for an example native. The error is the multi-state multi-condition error in Å relative to the native. The score is either the likelihood of D_1 (if $J = 1$) or the joint likelihood of D_1 and D_2 (if $J = 2$). Each decoy structure is a point and the global minimum of the scoring function is in red. **f**, The best R^{free} found in the sample for each pair of N, J for all natives.

Natives

There is no large set of multi-state multi-condition models with corresponding experimentally measured X-ray datasets. Thus, we use a synthetic benchmark, where a reflection dataset is computed, rather than measured, from a native. The benchmark consists of 10 2-state 2-condition models constructed, in part, from a previously determined SARS-CoV-2 M^{PRO} structure²⁹ (**Fig. 4.3a**).

Native reflection datasets

For a native, X-ray datasets D_1 and D_2 , corresponding to condition 1 and condition 2, are simulated via the forward model at 2.0 Å resolution based on the experimentally determined experimental unit cell dimensions and space group²⁹ (**Fig. 4.3b**). Gaussian noise is applied to the simulated structure factor amplitudes. Reflections within each resolution shell are randomly withheld for model assessment.

Model error

We define the model error of a N_A -state J -condition model by its deviation from a N_B -state J -condition model where all states are compositionally identical as follows:

$$\text{error} = \frac{1}{J} \sum_{j=0}^J \text{RMSD}(\text{avg}(M_1^A, \dots, M_{N_A}^A, w_{1j}, \dots, w_{N_A j}), \text{avg}(M_1^B, \dots, M_{N_B}^B, w_{1j}, \dots, w_{N_B j})) \quad (4.17)$$

The inside sum is the RMSD between the weighted average of M^A and the weighted average of M^B using the j column of the respective weight matrices. In the weighted average, the atomic position vector k is the weighted sum of the respective atomic position vectors across all states. Computing the model error in this fashion is useful for the following reasons. First, we can compute the model error when $N_A \neq N_B$. Second, if $N_A = N_B = 1$, then the model error is the average of the RMSD function across all conditions. Third, the model

error equals 0, if and only if the atomic position vector of the model average of M^A and the model average of M^B are equal for all atoms for all conditions. It is possible, however, for distinct models to have the same weighted average structure and thus an error of 0. Fourth, the model error shares the symmetry properties with the multi-state X-ray forward model for a given condition. The model error and multi-state X-ray forward model are a function of the set of atomic positions for a given atom k (along with their weights) and, therefore, independent of the assignment of an atom to a specific state. For example, the model error would not change, if the states in M^A or M^B are re-indexed. More generally, the model error would not change, if the positions of atoms k in any 2 states are swapped (state mixing).

Benchmark: accuracy of the global minimum

We estimate the value of the global minimum of the 4 scoring functions for each native (**Fig. 4.3c**). For a 2-state scoring function, we know that the global minimum error is 0 for all natives. We estimate the value of the global minimum of the 1-state scoring functions by computing the score of a large set of decoy models. The 1-state scoring functions do not have a global minimum near the native, because the 1-state model representation does not provide a good match to the 2-state model from which the data was simulated. Using a representation that better matches the data-generating process increases the accuracy of the global minimum of the scoring function.

4.4.2 Sampling convergence

To quantify the thoroughness of the search process, we plot the sampling convergence of the 4 scoring functions for each native (**Fig. 4.3d**).

The model search with a 1-state scoring function converges to the best score near the estimated global minimum for the respective scoring functions. In other words, the model search for satisfying the 2-state datasets finds a nearly best-scoring 1-state model. The model search for the 2-state scoring functions converges to a score considerably better than

the 1-state samples. However, as expected for a stochastic search, it does not find the global minima of the scoring functions, instead becoming stuck in local minima. The convergence curves suggest that further sampling would not significantly improve the best score found.

For both the 1-state and 2-state scoring functions, the consideration of an additional X-ray dataset does not significantly improve the best score found. The mean improvement in the 1-state 1-condition likelihood of D_1 vs the 1-state 2-condition marginal likelihood of D_1 is 0.050 ± 0.070 . The mean improvement in 2-state 1-condition likelihood of D_1 vs the 2-state 2-condition marginal likelihood of D_1 is 0.054 ± 0.112 . However, we see a significant improvement in R^{free} of D_1 by including a second X-ray dataset (**Fig. 4.3f**). The mean improvement in R^{free} from 1-state 1-condition vs 1-state 2-condition is 0.054 ± 0.026 . The mean improvement in R^{free} from 2-state 1-condition vs 2-state 2-condition is 0.030 ± 0.021 .

4.4.3 Benchmark: the shape of the scoring function

The convergence of the model search to the global minima of the 1-state scoring functions suggests that the scoring functions are smooth and have a large enough radius of convergence to facilitate the model search by molecular dynamics simulations, given the initial conditions.

The score and model error are correlated with each other more weakly for the 2-state vs 1-state scoring functions (**Fig. 4.3e**). This observation is consistent with the 2-state scoring function depending on twice the number of atomic variables and state weights. The inability of the model search to find the global minimum of the 2-state scoring function can likely be attributed to this decrease in the smoothness of the scoring function. Despite the randomized initial conditions, the trajectories tend to be trapped in similarly scoring, but distinct, local minima, as evidenced by the low variance in the convergence curves.

The smoothness and radius of convergence of the scoring function do not appear to be sensitive to considering additional conditions. Thus, we attribute the improvement in R^{free} from adding an additional condition to characteristics of the scoring function beyond our definitions of its smoothness and radius of convergence.

4.4.4 Multi-state multi-condition modeling of SARS-CoV-2 M^{Pro}

We illustrate our method by computing multi-state multi-condition models of the SARS-CoV-2 main protease (M^{Pro}), a therapeutic target⁴¹. Data was collected and modeled previously at 100K, 240K, 277K, 298K, 298K* (high humidity), and 310K²⁹.

Using our method as described, we compute a sample of models for each N in 1, 2, 4, 8, and 16 based on each of the 63 combinations of datasets, containing 1 to 6 datasets (J) (**Fig. 4.4a**). We construct a supersample containing all models for all N and J . For an X-ray dataset, the best R^{free} N -state J -condition model is defined as the supersample model with N states and J conditions with the minimum R^{free} ; similarly, the best R^{free} model is defined as the supersample model with the minimum R^{free} (**Fig. 4.4b**). For each dataset, we compare the R^{free} values of the PDB model, best R^{free} model, and best R^{free} 1-state 1-condition model (**Fig. 4.4c**), as follows.

For all 6 datasets, the best R^{free} 1-state 1-condition model has a worse R^{free} than the PDB model. In the 1-state 1-condition case, the accuracy of the global minimum of our scoring function is less accurate, presumably because the PDB model representation includes hydrogen, ion, ligand, and solvent atoms, thus depicting the reality more completely; in addition, another reason may be decreased search thoroughness of our method due to the lack of manual adjustments between rounds of refinement.

For all 6 datasets, the best R^{free} model has a better R^{free} than the best R^{free} 1-state 1-condition model. The best R^{free} model contains 2-16 states (N) computed based on 2-4 datasets (J). As shown previously, the model’s fit of the free reflections likely improves for 2 reasons: (i) increased global minimum accuracy from adding additional states to the model representation and (ii) better R^{free} from adding additional datasets. For 4 of the 6 datasets ($\geq 277\text{K}$), the best-free model has $N \geq 8$ states and fits the dataset better than the PDB model. For higher temperature datasets, the improvement in R^{free} from including additional states and datasets is greater than the loss in fit on account of the representation. This is not the case for the 2 lower-temperature datasets. For all 6 datasets, the fit of the free

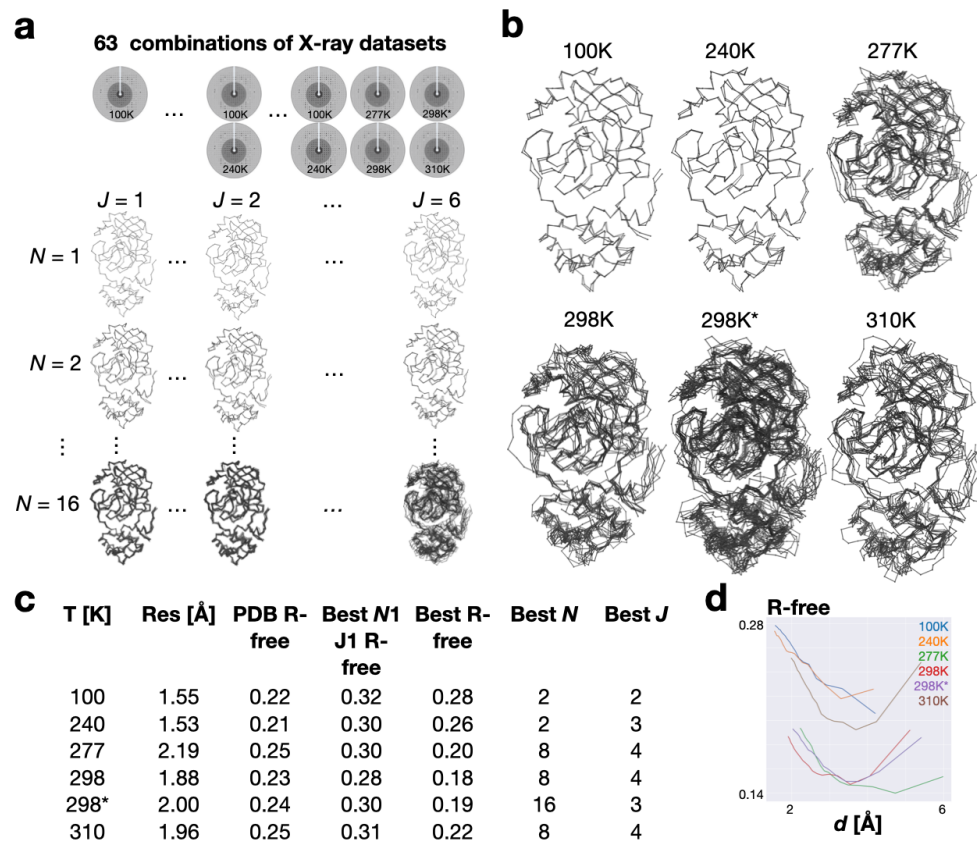
Figure 4

Figure 4.4: **a**, overview of all samples built for SARS-Cov-2 M^{pro} . A sample of good-scoring models was computed from the model posterior density for N in 1, 2, 4, 8, 16, and all combinations of X-ray datasets (J from 1 to 6). There are 63 non-empty combinations of X-ray datasets. Each sample contains 25 trajectories (7,500 models total). The super-sample contains $7,500 \times 63 \times 5 = 2,362,500$ models. **b**, The best-free model from the super-sample for each of the 6 temperature-dependent X-ray datasets. **c**, Statistics from the super-sample for each of the temperature-dependent X-ray datasets. The columns include the temperature the dataset was collected under, the resolution of the dataset, the R^{free} of the PDB model, the R^{free} of the best-free 1-state 1-condition model in the supersample for the dataset, the R^{free} of the best-free model in the supersample for the dataset, the number of states (N) in the best-free model, and the number of conditions used to inform the best-free model (J). **d**, The R^{free} of the best-free model for each dataset as a function of the resolution. The observed and model structure factors computed from the best-free model were divided into 20 resolution bins. The y-axis is the R^{free} of the resolution bucket and the X-axis is the minimum resolution (Å) of the resolution bucket.

reflections by the best-free model worsens as a function of resolution, also contributing to the relatively poor fit of the free reflections of the highest resolution datasets (100K, 240K) (**Fig. 4.4d**). The fit of the free reflections of the 277K, 298K, and 298K* is considerably better than the fit of corresponding reflections in 100K and 240K datasets, suggesting the better fit of the higher resolution datasets is at least in part from the including structural heterogeneity rather than the dataset being lower resolution and therefore more ambiguous.

We report the improvement in the R^{free} (ΔR^{free}) for each dataset by the best-free N -state J -condition model relative to the best-free 1-state 1-condition model for each N and J (**Fig. 4.3a**). Generally ΔR^{free} exhibits unimodal behavior with a maximum improvement at $J = 3.4$ and 2.4 for $\geq 277\text{K}$ and $< 277\text{K}$ datasets respectively. For 277K, 298K*, and 310K, the minimum ΔR^{free} (biggest improvement) occurs when all 4 of the $\geq 277\text{K}$ datasets were used together. For 298K, the minimum improvement occurs using 3 of the 4 $\geq 277\text{K}$ datasets. The lower temperature datasets are similarly grouped with 240K minimum ΔR^{free} based on the 100K, 240K, and 310K datasets and the 100K minimum ΔR^{free} based on the 100K and 240K datasets.

The partitioning of X-ray datasets by temperature condition is likely a result of the structural heterogeneity in the crystal being similar at similar temperatures. Datasets observed under temperature conditions with similar structural heterogeneity provide additional independent observations that prevent the multi-state model from overfitting to the noise of a single dataset. Computing a sample of N -state 1-condition models with w_{xray} as a multiple (1-6x) of the optimized w_{xray} value causes the best R^{free} to worsen in most cases, demonstrating that adding additional datasets does not satisfy the data better simply because of an increase in the contribution from X-ray datasets to the overall score (**Fig. 4.5b**).

We report the improvement in the fit of the force field (ΔFF) by the best-free N -state J -condition model relative to the best-free 1-state 1-condition model for each N and J for each dataset (**Fig. 4.5c**). The satisfaction of the force field decreases as a function of N and J . The dissatisfaction of the force field as a function of N is likely a result of the model

Figure 5

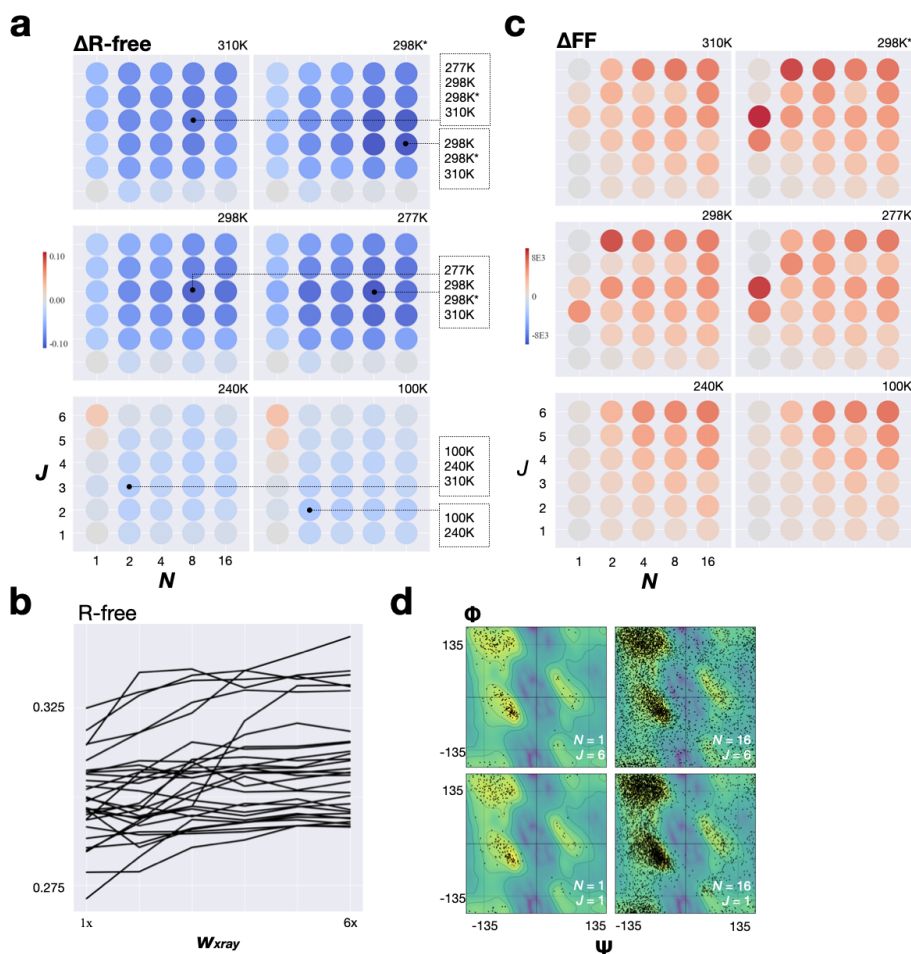


Figure 4.5: **a**, The difference between the R^{free} of the 1-state 1-condition best-free model and the best-free N -state J -condition model (ΔR^{free}) for all N and all J for all 6 datasets. A negative ΔR^{free} indicates greater improvement in satisfaction of the X-ray dataset. For each dataset, the ΔR^{free} corresponding to the best-free model is identified along with the corresponding input datasets. **b**, The difference between the force field energy (divided by the number of states) of the 1-state 1-condition best-free model and the best-free N -state J -condition model (ΔFF) for all N and all J for all 6 datasets. A negative ΔFF indicates greater improvement in satisfaction of the molecular mechanics force field. **c**, The R^{free} of the best-free N -state 1-condition model as a function of increased w_{xray} , up to 6x w_{xray} . As the w_{xray} for each sample was picked based on the w_{xray} that produced a best-free model with the minimum R^{free} out of a set of trial w_{xray} values, increasing w_{xray} generally does not improve the R^{free} of the best-free model. **d**, The Ramachandran plot for the best-free 1-state 1-condition, 1-state 6-condition, 16-state 1-condition, and 16-state 6-condition models for the 277K dataset.

being able to use a larger number of parameters to satisfy density in the Fo-Fc map that would not be able to be satisfied using a lesser number of states. For example, the model may violate the force field to satisfy the electron density that represents an ordered solvent molecule. The dissatisfaction of the prior as a function of J is likely a result of the attempt to satisfy distinct datasets with a single set of states. To further explore why the impact of the number of states and conditions on the force field, we see that the number of Ramachandran outliers is a synergistic function of both N and J (**Fig. 4.5d**). Integrative modeling aims to compute a model that best explains all input information, even when the input information represents varied experimental conditions and prior models.

4.5 Discussion

Here, we developed, benchmarked, and illustrated a Bayesian method to compute a model depicting multiple structural states from multiple X-ray diffraction patterns representing distinct experimental conditions.

First, we discuss the impact of introducing multiple states into the scoring function for a single condition (X-ray dataset). A multi-state scoring function evaluates the collective satisfaction of an X-ray dataset by all states. The scoring function can be generalized for scoring a model depicting multiple states by exploiting the degeneracy of the X-ray forward model for a weighted set of structural states and the weighted sum of the forward model for a single structural state (**Fig. 4.2b**). In contrast, the degeneracy of the X-ray forward model of assigning an atom to a particular state complicates the interpretation of the multi-state model. To resolve this degeneracy, additional information, such as a molecular mechanics force field, is required.

We show that by introducing additional states, the global minimum of the scoring function is closer to the native (**Fig. 4.3c**). For experimental data, where the true number of structural states far exceeds the number of states reasonably included in a model, adding

additional states will continue improving the accuracy of the global minimum of the scoring function. The accuracy of the global minimum will also be further improved by including additional variables to represent the atomic content of a protein crystal, such as water molecules, ions, and ligands. When the optimal representation is unknown before modeling, the number and type of model parameters may be sampled alongside the values of the remaining model parameters. In such cases, it may be beneficial to consider model selection criteria such as Bayesian or Akaike Information Criterion to explain the data with as few model parameters as possible⁴².

Adding additional states decreases the smoothness of the scoring function (**Fig. 4.3e**), which may contribute to the inability of the model search to find the global minimum of the multi-state multi-condition scoring function. The decreased smoothness of the scoring function results from the increased number of model parameters. It is conceivable that removing some nuisance parameters from the likelihood, such as the scaling factors α and β , could increase the smoothness of the scoring function.

Next, we discuss the impact of introducing multiple conditions into the scoring function. A multi-state multi-condition scoring function evaluates the collective satisfaction of all X-ray datasets by all states. The scoring function can be generalized to multiple conditions by computing a joint likelihood from the multiplicative product of the likelihood of each independent X-ray dataset. For a gradient-based sampling algorithm, such as molecular dynamics simulations, the force on an atom is the sum of the partial derivatives of each log-likelihood, thereby pushing the atom to a position that better satisfies all X-ray datasets. It was previously found advantageous to balance the first derivatives of the likelihood and the force field prior (w_{auto})²¹. Correspondingly, we also found by trial and error the best scaling factor for the partial derivatives for all log-likelihoods (w_{xray}).

Considering additional conditions significantly improves the satisfaction of the synthetic and experimental free reflections (**Fig. 4.3f**, **Fig. 4.5a**), while it does not appear to impact the properties of the scoring function nor the satisfaction of the synthetic or experimental

work reflections. Thus, consideration of additional conditions must impact more complex properties of the scoring function than its smoothness and radius of convergence. In other words, the geometry of a scoring function with multiple conditions improves the ability of the model search to find a minimum reflecting the actual native rather than the noise of the work reflections.

The model search may be further improved using advanced sampling algorithms, such as simulated annealing instead of the constant-temperature molecular dynamics simulations used here. Such methods would expand the explored model space for a given amount of CPU time⁴³. Additionally, future sampling may benefit from including additional information about stereochemistry (*eg*, sampling side chains using known rotameric states).

Several applications may benefit from Bayesian multi-state multi-condition modeling in X-ray crystallography. First, in serial crystallography, a series of small and often imperfect crystals are exposed to the X-ray beam, generating partial diffraction patterns at a series of time points⁴⁴. This trajectory could be represented as differently weighted mixtures of the same states along the sampled time points. Using our multi-state multi-condition scoring function would inform a model of each state at each time point based on the totality of the X-ray data, presumably improving the accuracy of the models as demonstrated here. Second, in a fragment screen, the protein is crystallized in the presence of one fragment at a time, corresponding to a single condition⁴⁵. Fragments often have low affinity, and therefore, only a fraction of the protein molecules in the crystal will be bound. The multi-state multi-condition representation can conceivably combine data from multiple experiments in several ways. For example, a multi-state multi-condition model may be computed where the states correspond to each fragment's apo state and the dominant holo state. As a result, X-ray diffraction patterns from each fragment will inform all states. Third, non-Bragg scattering is often ignored in modeling based on X-ray crystallography data. For example, diffuse scattering contains information on the correlated motions of atoms in the crystal⁴⁶. Our Bayesian framework may facilitate including another likelihood term based on diffuse scattering data

in computing a multi-state model.

To facilitate the application of our method by scientists to many other systems, we implemented it as a module of our open-source *Integrative Modeling Platform* (IMP)³⁹, freely available at <https://integrativemodeling.org/2.20.2/doc/manual/>.

Acknowledgments

We would like to thank Ben Webb (UCSF), Pavel Afonine (LBNL), Billy Poon (LBNL).

References

1. Rejto, P. A. & Freer, S. T. Protein conformational substates from X-ray crystallography. en. *Prog. Biophys. Mol. Biol.* **66**, 167–196 (1996) (cit. on p. 137).
2. Smith, C. A., Ban, D., Pratihari, S., Giller, K., Schwiegk, C., de Groot, B. L., Becker, S., Griesinger, C. & Lee, D. Population shuffling of protein conformations. en. *Angew. Chem. Int. Ed Engl.* **54**, 207–210 (Jan. 2015) (cit. on p. 137).
3. Woldeyes, R. A., Sivak, D. A. & Fraser, J. S. E pluribus unum, no more: from one crystal, many conformations. en. *Curr. Opin. Struct. Biol.* **28**, 56–62 (Oct. 2014) (cit. on pp. 137, 138).
4. DePristo, M. A., de Bakker, P. I. W. & Blundell, T. L. Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. en. *Structure* **12**, 831–838 (May 2004) (cit. on p. 137).
5. Jensen, L. H. Refinement and reliability of macromolecular models based on X-ray diffraction data. en. *Methods Enzymol.* **277**, 353–366 (1997) (cit. on p. 137).
6. Ringe, D. & Petsko, G. A. Study of protein dynamics by X-ray diffraction. en. *Methods Enzymol.* **131**, 389–433 (1986) (cit. on pp. 137, 138).
7. Sun, Z., Liu, Q., Qu, G., Feng, Y. & Reetz, M. T. Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. en. *Chem. Rev.* **119**, 1626–1665 (Feb. 2019) (cit. on p. 137).
8. Kuzmanic, A., Pannu, N. S. & Zagrovic, B. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. en. *Nat. Commun.* **5**, 3220 (2014) (cit. on p. 137).
9. Kuriyan, J., Petsko, G. A., Levy, R. M. & Karplus, M. Effect of anisotropy and anharmonicity on protein crystallographic refinement. An evaluation by molecular dynamics. en. *J. Mol. Biol.* **190**, 227–254 (July 1986) (cit. on p. 137).

10. Holton, J. M., Classen, S., Frankel, K. A. & Tainer, J. A. The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. en. *FEBS J.* **281**, 4046–4060 (Sept. 2014) (cit. on p. 138).
11. Vitkup, D., Ringe, D., Karplus, M. & Petsko, G. A. Why protein R-factors are so large: a self-consistent analysis. en. *Proteins* **46**, 345–354 (Mar. 2002) (cit. on p. 138).
12. Van den Bedem, H. & Fraser, J. S. Integrative, dynamic structural biology at atomic resolution—it’s about time. en. *Nat. Methods* **12**, 307–318 (Apr. 2015) (cit. on p. 138).
13. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. en. *Nature* **450**, 964–972 (Dec. 2007) (cit. on p. 138).
14. Schiffer, C. A. in *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications* (eds van Gunsteren, W. F., Weiner, P. K. & Wilkinson, A. J.) 265–269 (Springer Netherlands, Dordrecht, 1997) (cit. on p. 138).
15. Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. Modelling dynamics in protein crystal structures by ensemble refinement. en. *Elife* **1**, e00311 (Dec. 2012) (cit. on pp. 138, 148).
16. Wankowicz, S. A., Ravikumar, A., Sharma, S., Riley, B. T., Raju, A., Flowers, J., Hogan, D., van den Bedem, H., Keedy, D. A. & Fraser, J. S. Uncovering Protein Ensembles: Automated Multiconformer Model Building for X-ray Crystallography and Cryo-EM. en. *bioRxiv* (Apr. 2024) (cit. on p. 138).
17. Riley, B. T., Wankowicz, S. A., de Oliveira, S. H. P., van Zundert, G. C. P., Hogan, D. W., Fraser, J. S., Keedy, D. A. & van den Bedem, H. qFit 3: Protein and ligand multiconformer modeling for X-ray crystallographic and single-particle cryo-EM density maps. en. *Protein Sci.* **30**, 270–285 (Jan. 2021) (cit. on p. 138).
18. Wilson, M. A. & Brunger, A. T. The 1.0 Å crystal structure of Ca(2+)-bound calmodulin: an analysis of disorder and implications for functionally relevant plasticity. en. *J. Mol. Biol.* **301**, 1237–1256 (Sept. 2000) (cit. on p. 138).

19. Levin, E. J., Kondrashov, D. A., Wesenberg, G. E. & Phillips Jr, G. N. Ensemble refinement of protein crystal structures: validation and application. en. *Structure* **15**, 1040–1052 (Sept. 2007) (cit. on p. 138).
20. Kuriyan, J., Osapay, K., Burley, S. K., Brünger, A. T., Hendrickson, W. A. & Karplus, M. Exploration of disorder in protein structures by X-ray restrained molecular dynamics. en. *Proteins* **10**, 340–358 (1991) (cit. on p. 138).
21. Burling, F. T. & Brünger, A. T. Thermal motion and conformational disorder in protein crystal structures: Comparison of multi-conformer and time-averaging models. en. *Isr. J. Chem.* **34**, 165–175 (Jan. 1994) (cit. on pp. 138, 148, 160).
22. Thompson, M. C. Combining temperature perturbations with X-ray crystallography to study dynamic macromolecules: A thorough discussion of experimental methods. en. *Methods Enzymol.* **688**, 255–305 (Aug. 2023) (cit. on p. 138).
23. Frauenfelder, H., Petsko, G. A. & Tsernoglou, D. Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. en. *Nature* **280**, 558–563 (Aug. 1979) (cit. on p. 139).
24. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. en. *Science* **254**, 1598–1603 (Dec. 1991) (cit. on p. 139).
25. Tilton Jr, R. F., Dewan, J. C. & Petsko, G. A. Effects of temperature on protein structure and dynamics: X-ray crystallographic studies of the protein ribonuclease-A at nine different temperatures from 98 to 320 K. en. *Biochemistry* **31**, 2469–2481 (Mar. 1992) (cit. on p. 139).
26. Fraser, J. S., Clarkson, M. W., Degnan, S. C., Erion, R., Kern, D. & Alber, T. Hidden alternative structures of proline isomerase essential for catalysis. en. *Nature* **462**, 669–673 (Dec. 2009) (cit. on p. 139).
27. Halle, B. Biomolecular cryocrystallography: structural changes during flash-cooling. en. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4793–4798 (Apr. 2004) (cit. on p. 139).

28. Keedy, D. A., van den Bedem, H., Sivak, D. A., Petsko, G. A., Ringe, D., Wilson, M. A. & Fraser, J. S. Crystal cryocooling distorts conformational heterogeneity in a model Michaelis complex of DHFR. en. *Structure* **22**, 899–910 (June 2014) (cit. on p. 139).
29. Ebrahim, A., Riley, B. T., Kumaran, D., Andi, B., Fuchs, M. R., McSweeney, S. & Keedy, D. A. The temperature-dependent conformational ensemble of SARS-CoV-2 main protease (Mpro). en. *bioRxiv* (Nov. 2021) (cit. on pp. 139, 152, 155).
30. Du, S., Wankowicz, S. A., Yabukarski, F., Doukov, T., Herschlag, D. & Fraser, J. S. Refinement of Multiconformer Ensemble Models from Multi-temperature X-ray Diffraction Data. en. *bioRxiv* (May 2023) (cit. on p. 139).
31. Rout, M. P. & Sali, A. Principles for Integrative Structural Biology Studies. en. *Cell* **177**, 1384–1403 (May 2019) (cit. on pp. 140, 143).
32. Sali, A. From integrative structural biology to cell biology. en. *J. Biol. Chem.* **296**, 100743 (Jan. 2021) (cit. on p. 140).
33. Brooks, B. R., Brooks 3rd, C. L., Mackerell Jr, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caffisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M. & Karplus, M. CHARMM: the biomolecular simulation program. en. *J. Comput. Chem.* **30**, 1545–1614 (July 2009) (cit. on pp. 140, 147).
34. Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D. & Urzhumtsev, A. Bulk-solvent and overall scaling revisited: faster calculations, improved results. en. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 625–634 (Apr. 2013) (cit. on pp. 142, 149).
35. Lunin, V. Y., Afonine, P. V. & Urzhumtsev, A. G. Likelihood-based refinement. I. Irremovable model errors. en. *Acta Crystallogr. A* **58**, 270–282 (May 2002) (cit. on pp. 142–144, 148).

36. Mcelreath, R. Statistical rethinking: A Bayesian course with examples in R and Stan (Dec. 2015) (cit. on p. 142).
37. Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. en. *Acta Crystallogr D Struct Biol* **75**, 861–877 (Oct. 2019) (cit. on pp. 144, 149).
38. Guinier, A. X-ray Diffraction in Crystals. *Imperfect Crystals, and Amorphous Bodies*. Dorer (1963) (cit. on p. 145).
39. Russel, D., Lasker, K., Webb, B., Velázquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B. & Sali, A. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. en. *PLoS Biol.* **10**, e1001244 (Jan. 2012) (cit. on pp. 149, 162).
40. Hancock, M., Peulen, T.-O., Webb, B., Poon, B., Fraser, J. S., Adams, P. & Sali, A. Integration of software tools for integrative modeling of biomolecular systems. en. *J. Struct. Biol.* **214**, 107841 (Mar. 2022) (cit. on p. 149).
41. Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., Duan, Y., Yu, J., Wang, L., Yang, K., Liu, F., Jiang, R., Yang, X., You, T., Liu, X., Yang, X., Bai, F., Liu, H., Liu, X., Guddat, L. W., Xu, W., Xiao, G., Qin, C., Shi, Z., Jiang, H., Rao, Z. & Yang, H. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. en. *Nature* **582**, 289–293 (June 2020) (cit. on p. 155).
42. Stoica, P. & Selen, Y. Model-order selection: a review of information criterion rules. *IEEE Signal Process. Mag.* **21**, 36–47 (July 2004) (cit. on p. 160).

43. Kirkpatrick, S., Gelatt Jr, C. D. & Vecchi, M. P. Optimization by simulated annealing. en. *Science* **220**, 671–680 (May 1983) (cit. on p. 161).
44. Barends, T. R. M., Stauch, B., Cherezov, V. & Schlichting, I. Serial femtosecond crystallography. en. *Nat Rev Methods Primers* **2** (Aug. 2022) (cit. on p. 161).
45. Krojer, T., Fraser, J. S. & von Delft, F. Discovery of allosteric binding sites by crystallographic fragment screening. en. *Curr. Opin. Struct. Biol.* **65**, 209–216 (Dec. 2020) (cit. on p. 161).
46. Pei, X., Bhatt, N., Wang, H., Ando, N. & Meisburger, S. P. Introduction to diffuse scattering and data collection. en. *Methods Enzymol.* **688**, 1–42 (Aug. 2023) (cit. on p. 161).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

Matthew Hancock

F5E8F28F6DAA457...

Author Signature

5/30/2024

Date