AWARD NUMBER: W81XWH-17-1-0412

TITLE: Histone Lysine Methyltransferases-Conformational Dynamics and Selective Inhibitor Design for Chromatin-Modifying Enzymes in Lymphomas and Melanomas

PRINCIPAL INVESTIGATOR: Rafal Wiewiora

CONTRACTING ORGANIZATION:
Weill Medical College of Cornell University
New York, NY 10065

REPORT DATE:
September 2018

TYPE OF REPORT:
Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command

Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
September 2018	Annual	15 Aug 2017 - 14 Aug 2018

4. TITLE AND SUBTITLE

Histone Lysine Methyltransferases-Conformational Dynamics and Selective Inhibitor Design for Chromatin-Modifying Enzymes in Lymphomas and Melanomas

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-17-1-0412

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Rafal Wiewiora

E-Mail: rafal.wiewiora@choderalab.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

WEILL MEDICAL COLLEGE OF CORNELL
UNIVERS 407 E 61ST ST RM 106 1ST FL NEW
YORK NY 10065-4805

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The long-term objective of this project is to advance the understanding of the biology, therapeutic potential and availability of small molecule drugs for cancer-implicated histone lysine methyltransferases (HKMTs) EZH2 (lymphoma and melanoma target) and SETDB1 (melanoma target). The immediate objective of this project is to present to the scientific community a number of novel small molecule binding modes and pockets in those protein targets, as well as novel small molecule chemical probe scaffolds to hit them. In this report, I show the preparation and data collection in tens of thousands of molecular dynamics trajectories on the distributed computing system Folding@home, and generation of dynamic models of the conformational ensembles of EZH2 and EED, both in apo, and in-complex form (PRC2 complex). I successfully deployed a semi-automatic Markov state model building pipeline on a pilot model system SETD8, hence creating a well sampled, prototypical Markov state model, from which to seed conformations for this project and extract reaction coordinates for adaptive and enhanced sampling simulation runs. I am continuing to refine the models, while building a small molecule ensemble pocket detection – docking – free energy calculations pipeline to select candidates for screening.

15. SUBJECT TERMS
protein methyltransferase; histone methyltransferase; EZH2; EED; PRC2; SETD8; conformational dynamics; protein dynamics; Folding@home; distributed computing; Markov state models; ensemble docking

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	54	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

INTRODUCTION

The long-term objective of this project is to advance the understanding of the biology, therapeutic potential and availability of small molecule drugs for cancer-implicated histone lysine methyltransferases (HKMTs) EZH2 (lymphoma and melanoma target) and SETDB1 (melanoma target). The immediate objective of this project is to present to the scientific community a number of novel small molecule binding modes and pockets in those protein targets, as well as novel small molecule chemical probe scaffolds to hit them. We postulate that knowledge of the conformational landscapes of these targets will: a) expose novel small molecule binding modes within known binding sites; b) expose novel allosteric binding sites, where binding will affect either the activity of the protein, disrupt protein-protein interactions within their active complexes, or result in conformational trapping and disruption of the formation of such complexes; c) allow for faster and cheaper development of more selective inhibitors.

KEYWORDS

protein methyltransferase; histone methyltransferase; EZH2; EED; PRC2; SETD8; conformational dynamics; protein dynamics; Folding@home; distributed computing; Markov state models; ensemble docking

ACCOMPLISHMENTS

What were the major goals of the project?

Major Task 1: Prepare homology models of EZH2 and SETDB1 from all-transferase templates using Ensembler and prepare them for molecular dynamics simulations. Prepare simulations of the PRC2 complex and the complex SETDB1 participates in.

a) Software engineering of Ensembler to add support for zinc ion clusters and ligand modeling.

b) Run Ensembler using all-transferase templates.

c) Prepare simulations of the multi-protein complexes.

Major Task 2: Run the molecular dynamics simulations on Folding@home to multiple-millisecond aggregate timescale.

a) Perform internal and beta testing of the simulations.

b) Run the simulations on Folding@home. Curate while running.

Major Task 3: Build Markov State Models of the molecular dynamics simulation data.

a) Manually build preliminary Markov state models.

b) Perform *Osprey* runs to find the best models.

Major Task 4: Characterize the amount of conformational flexibility seen in known binding sites.

a) Coarse-grain the Markov State Models to result in metastable states appropriate for drug design. Choose the distinct conformational states to proceed.

b) Perform bioinformatics structural analyses of the known binding pockets in the context of all distinct conformational states seen. Identify opportunities for novel modes of small molecule binding.

Major Task 5: Identify novel cryptic allosteric sites. Identify allosteric hotspots to disrupt the protein-protein interactions in the complexes.

a) Analyze the Markov State Models with the LIGSITE algorithm and rolling-probe algorithms.

b) Perform Mutual Information analyses of rotameric state coupling between the proposed allosteric sites and the active sites.

What was accomplished under these goals?

Major Task 1:

a) Completed.

b) Partially completed, so far with a smaller (not all-transferase yet) dataset.

c) Completed, except for SETDB1 (see PROBLEMS).

Major Task 2:

a) Completed.

b) Completed, continue to collect more data.

Major Task 3:

a) Completed.

b) In progress.

Major Task 4:

a) Partially completed with a preliminary model.

b) In progress with a preliminary model.

Major Task 5:

a) Preparing software, testing with a preliminary model.

b) Preparing software, testing with a preliminary model.

Development and testing of the molecular modeling pipeline on the model system SETD8 (Figures 1 and 2)

Previously and partially in this project as a model system, we studied the conformational landscape of another protein lysine methytransferase – SETD8 (*The Dynamic Conformational Landscapes of the Protein Methyltransferase SETD8*: https://www.biorxiv.org/content/10.1101/438994v1 – now accepted, in revision at eLife). SETD8 is evolutionarily closest to the common ancestor of all PKMTs, hence providing the perfect model system. For that study, our collaborators collected crystal structures of multiple novel conformations of the protein trapped with or without various ligands, providing diverse molecular dynamics simulation seeds. This has provided me with a well sampled, prototypical Markov state model, from which to seed conformations for this project and extract reaction coordinates for adaptive and enhanced sampling simulation runs. I present here two figures from the SETD8 paper as an example of the expected output of this project. Figure 1 shows 24-state and 10-state Markov state models constructed for apo and cofactor-bound SETD8 respectively, highlighting the different populations of known conformations, new ‘hidden’ conformations, and pathways of connectivity. Figure 2 shows a functional annotation of the models, highlighting the

connection between the pre-existing conformations in the apo conformational landscape and their reweighing as the protein progresses in its catalytic cycle.

Large scale homology modeling of EZH2 from all available lysine methyltransferase structures (Figure 3)

In order to classify and use in simulation all available structural information in the PDB database about protein lysine methytransferases (PKMTs), all crystal structures available were downloaded with *Ensembler.* The sequence of the methyltransferase (SET) domain of EZH2 was homology modeled using all PKMT structures, resulting in over 300 models. Those were sorted by RMSD to the real EZH2 crystal structure and a cutoff was chosen to discard models of low quality due to low sequence homology to the target, resulting in 186 final models. Figure 3 shows the resulting ensemble of EZH2 conformations. These will be used as seeds in parallel Folding@home simulations. It is expected that this conformationally wide seeding will act to boost sampling in a similar way as starting from crystal structures of SETD8 collected for this purpose did (those are also used as templates for EZH2).

Markov state model analysis of apo-EZH2 5 millisecond simulation dataset (Figure 4)

Initial simulations of the EZH2 SET domain starting from the PRC2 conformation and an inactive apo conformation resulted in ~ 5 milliseconds (10 million conformations, >200 GPU-years) of molecular dynamics data. Following the previously established protocol, protein was featurized with backbone and sidechain dihedral angles, and after kinetic mapping the data were clustered into 100 microstates. For intepretability, a 10 macrostate Hidden Markov Model (HMM) was built, and a millisecond long trajectory was simulated from it, highlighting the power of short parallel simulations to reconstruct a long-timescale landscape. Figure 4 shows conformations of EZH2 from the simulated trajectory, separated by 200 microseconds. The three slowest relaxation modes of the model are also shown, highlighting the dynamics of the SET-I loop and the post-SET. I am currently analyzing the model for the presence of changed and new binding pockets, and preparing a large scale docking experiment to the Enamine REAL library.

Comparative modeling of EZH2 and EED in apo, and in the PRC2 complex (Figure 5)

Analogically to the construction of the apo-EZH2 Hidden Markov Model, a model of the EED reader protein from over 2 ms of simulation data was constructed. Over 50 microseconds of simulation were run for the whole PRC2 complex (composed of EZH2, EED, and SUZ12) on the local cluster, as the necessary developments to Folding@home to enable the simulation of systems this large are still in

progress (see ‘Maintenance and development of Folding@home’). Dynamics of EZH2 and EED within the PRC2 complex were described using the same state space as for the apo proteins, highlighting the conformational restrictions within the complex, which can be used to design probes stabilizing conformations incompatible with complex formation (Figure 5).

Development and testing of a small molecule ensemble docking pipeline (Figure 6)

After construction of an apo Markov state model for a protein of interest, we would like to predict the changes to the conformational landscape due to ligand binding, and hence discover chemical probe candidates. We have developed a pipeline to achieve this in two complimentary ways: a) ‘brute force’: if the binding site and pose of the ligand in some protein conformation are known, the ligand is copied into all other conformations into the exact same position, and relaxed alchemically (the interactions of the ligand with the protein are gradually turned on) into a putative binding pose; b) ‘standard’ docking into the vicinity of the active site or a newly identified pocket, no previous knowledge about the binding mode is used (Figure 6). This pipeline has been tested on the states of the SETD8 model, for binding to the SAM cofactor. I will next expand the pipeline with alchemical free energy calculations to calculate the free energy of binding of a ligand to each state and validate against experimental data with known ligands. Further, I will begin adapting code for pocket and allosteric coupling detection published by another research group, and we are exploring a collaboration to use deep learning to achieve the small molecule screening goals.

Maintenance and development of Folding@home

The extensive simulation sampling required for this project, particularly for very large systems, such as the PRC2 complex, made it necessary to maximize the power of Folding@home. The simulation engine we use, OpenMM, has steadily increased its utilization of the newest GPUs in every new version. Folding@home requires compilation of new ‘cores’ for every new version of OpenMM, a time-intensive task, first from the programmatic standpoint, second due to the need to test the stability of every new core on the thousands of distributed GPUs. I have been working on building an OpenMM 7.2 core – which will give us to 50% speedup in computation speed (equivalent of ~$3M in new hardware), when combined with hydrogen mass repartitioning. This is necessary to achieve reasonable sampling for the PRC2 complex, as well as all the systems containing zinc clusters (e.g. full length EZH2) require features of the newest OpenMM to model those clusters. This work has been taking much longer than expected due to instability of the new core seen in testing so far and necessity of further debugging. Further, I have also spent considerable amount of time refining the system of processing data (tens of TBs) from the Folding@home servers.

Testing of new simulation methods

In the previous work with model system SETD8, it became apparent that even though we are able to collect unprecedented amount of molecular dynamics simulation data on Folding@home, the sampling problem is still significant for the task at hand. We would like to maximize the insight into, in particular, alternative binding pockets whose openings might be very rare. Three advancements to the simulation pipeline are being tested: a) better seeding – how to pre-generate more useful starting conformations as simulation seeds, using e.g. accelerated / steered molecular dynamics; b) adaptive sampling – how to automatically analyze existing data continuously, and propose to stop and restart trajectories from optimal new seeds - I collected ~10 ms of data for a short peptide (trp-zip) to validate different strategies against; c) enhanced sampling – once we have learnt the slow relaxation modes of a system from unbiased simulation, we can perform biased (enhanced) sampling along those coordinates, e.g. metadynamics, to rapidly calculate the free energies, and refine existing models and obtain free energy profiles for other PKMTs, mutants or ligand-bound states. I am testing different metadynamics strategies on a library of designed mini-proteins with experimental hydrogen exchange data available.

Others

I have collected multiple milliseconds of simulation data for the Tudor domain of SETDB1, which is awaiting analysis.

Another point of interest in this project is the influence of cancer mutations on the conformational landscapes of the proteins and hence chemical probe binding. I am developing a pipeline using Rosetta to reweigh Markov state model states by mutational free energy changes. Mutations could also lead to the emergence of new conformations – in order to assess the balance of the two effects on a model system, I collected nearly 30 ms of simulation on 26 mutants of SETD8, which is awaiting analysis.

The pipeline of this project depends on the ability to automate the construction and cross-validation / scoring of Markov state models. The most efficient approach to exploring the model space is still an unresolved question in the field. I am exploring these strategies on long protein folding simulation data from DE Shaw Research as model systems.

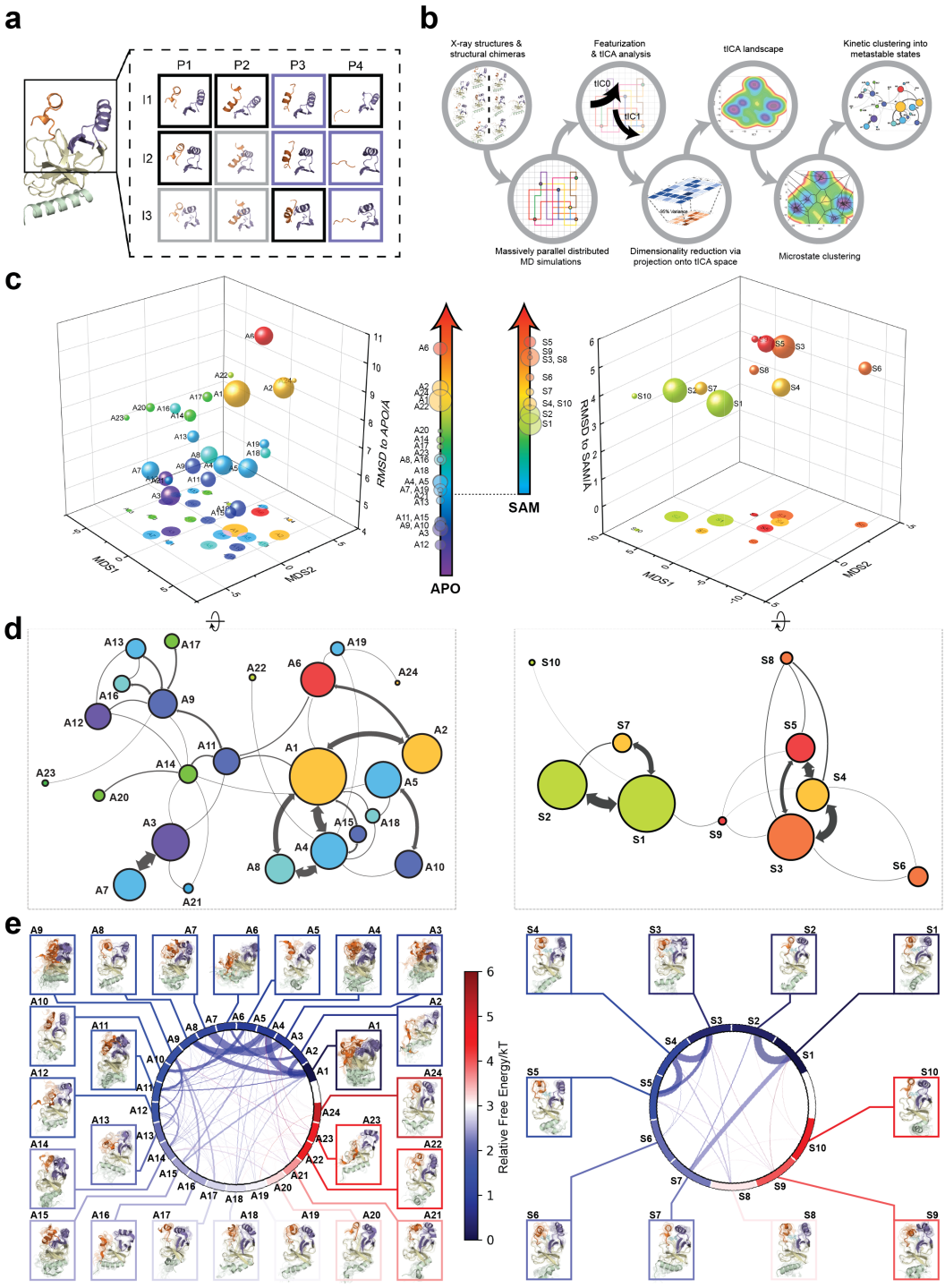


Figure 1. Successful demonstration of the modeling pipeline used for this work, on a model system - protein lysine methyltransferase SETD8. Markov state models and conformational landscapes of apo- and SAM-bound SETD8 constructed through diversely seeded, massively parallel molecular dynamics simulations. **a**, Combinatorial construction of structural domain chimeras using crystallographically-derived post-SET and SET-I conformations. Each conformer is boxed and color-coded with black for five X-ray-derived structures, blue for four putative structural chimeras included

as seed structures for MD simulations, and grey for three structural chimeras excluded from MD simulations because of obvious steric clashes. **b**, Schematic workflow to construct dynamic conformational landscapes via MSM. The five X-ray structures and the four structural chimeras were used to seed massively parallel MD simulations on Folding@home (see Method). Markov state models were constructed from these MD simulation results to reveal the conformational landscape. **c–e**, Kinetically metastable conformations (macrostates) obtained from kinetically coupled microstates via Hidden Markov Model (HMM) analysis. The revealed dynamic conformational landscapes consist of 24 macrostates for apo-SETD8 (left panel) and 10 macrostates for SAM-bound SETD8 (right panel). **c**, Kinetic and structural separation of macrostates in a 3D scatterplot. The X, Y axes represent kinetic separation of macrostates with a log-inverse flux kinetic embedding method (see Methods). The Z axis reports RMSDs of each macrostate to APO (left) or BC-SAM (right). The relative population of each macrostate of apo- or SAM-bound SETD8 ensembles is proportional to the volume of each representative sphere. **d**, Cartoon depiction of macrostates in a 2D scatterplot. The relative positions of metastable conformations were derived via the log-inverse flux kinetic embedding (see Methods). The diameter of the corresponding circle in the 2D scatterplot is proportional to the diameter of the respective sphere in the 3D scatterplot above. Equilibrium kinetic fluxes larger than 7.14×10^{2} s^{-1} for apo- and 1.39×10^{3} s^{-1} for SAM-bound SETD8 are shown for interconversion kinetics with thickness of the connections proportional to fluxes between two macrostates. **e**, Chord diagrams and representative conformers of macrostates. The colors are encoded for the free energy of each macrostate relative to the lowest free energy of the macrostates. The equilibrium flux between two macrostates is proportional to thickness of respective arcs.

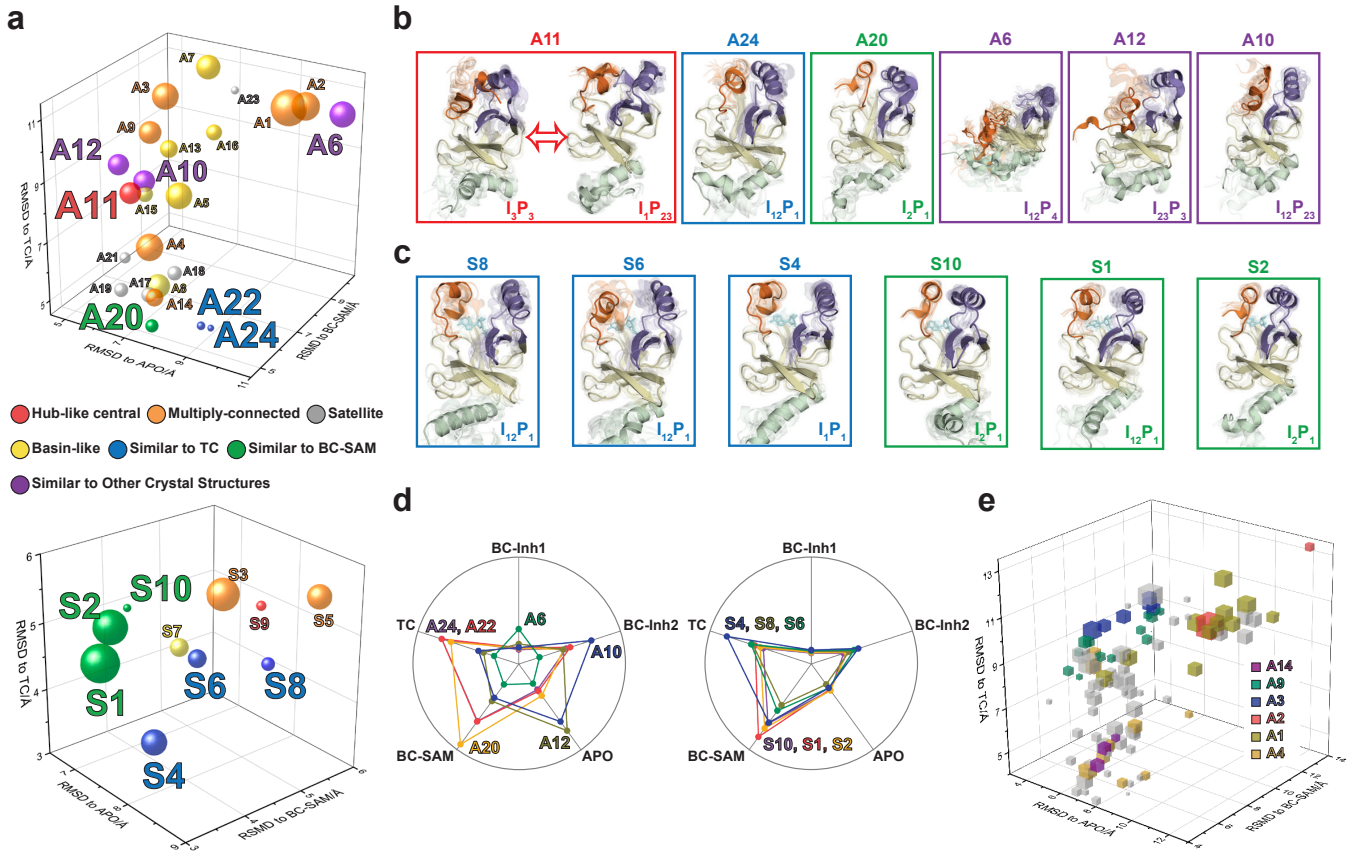


Figure 2. Functional annotation of the dynamic conformational landscapes of the model system SETD8. **a**, 3D scatterplots of the 24 macrostates of apo-SETD8 landscape and 10 macrostates of SAM-bound SETD8 landscape in the coordinates of RMSDs relative to APO, BC-SAM, and TC. Volume of each sphere is proportional to the relative population of the corresponding macrostate in the context of the 24 macrostates for apo-SETD8 or the 10 macrostates for SAM-bound SETD8. The RMSD of each macrostate is the average of its microstates weighted with their intra-macrostate population. The RMSD of each microstate is the average of the top 10 frames most closely related to the clustering center of the microstate. The feature of each macrostate is annotated in color. **b**, **c** Cartoons of representative conformations of key macrostates in the apo-SETD8 landscape and the SAM-bound SETD8 landscape, respectively. Structural annotations are shown in bottom right of each conformation. **d**, Radar chart of representative macrostates of apo (left) and SAM-bound (right) landscapes in reference to the five crystal structures. Distances between dots and cycle centers are proportional to the reciprocal values of RMSDs of macrostates relative to the crystal structures. **e**, 3D scattering plot of 100 microstates of the apo landscape in the coordinates of RMSDs to APO, BC-SAM, and TC. Volume of each cube is proportional to the relative population of the

corresponding microstate in the context of the 100 microstates. Microstates clustered in intermediate-like macrostates are highlighted in colors. Structural diversity of microstates within individual macrostates indicates that each intermediate-like state contains multiple structurally distinct but readily interconvertible microstates.

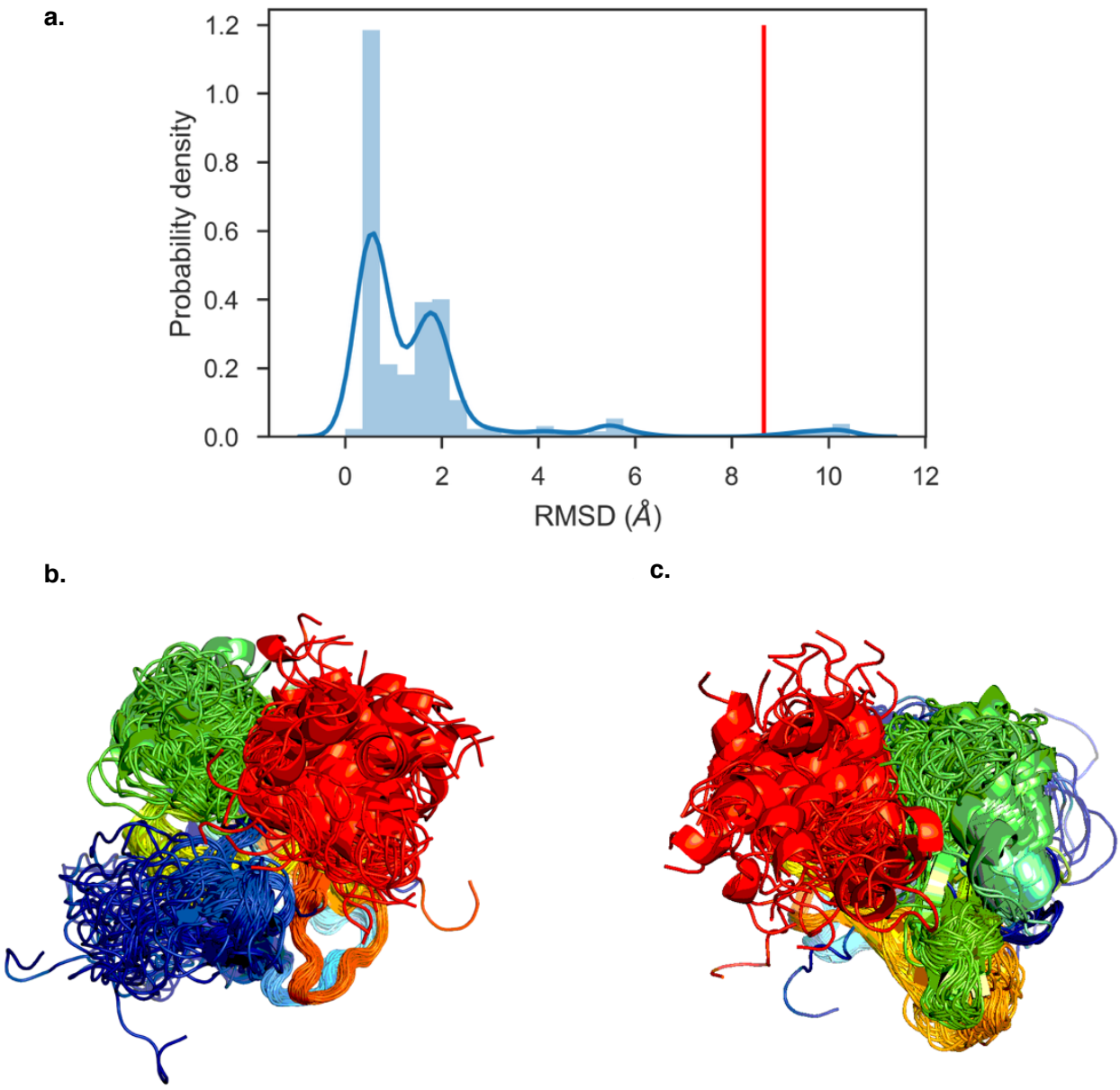


Figure 3. *Ensembler* homology modeling of the EZH2 SET domain sequence onto the structures of all SET domains deposited in the PDB database. The sequence of the methyltransferase (SET) domain of EZH2 was homology modeled using all PKMT structures, resulting in over 300 models. Those were sorted by RMSD to the real EZH2 crystal structure and a cutoff was chosen to discard models of low quality due to low sequence homology to the target, resulting in 186 final models. **a.** All-heavy-atom RMSD of over 300 initial models to the real EZH2 crystal structure, red line marks the RMSD cutoff used to discard bad models, **b., c.** two views of all 186 final homology models of the EZH2 SET domain, highlighting the diversity of conformations for seeding simulations.

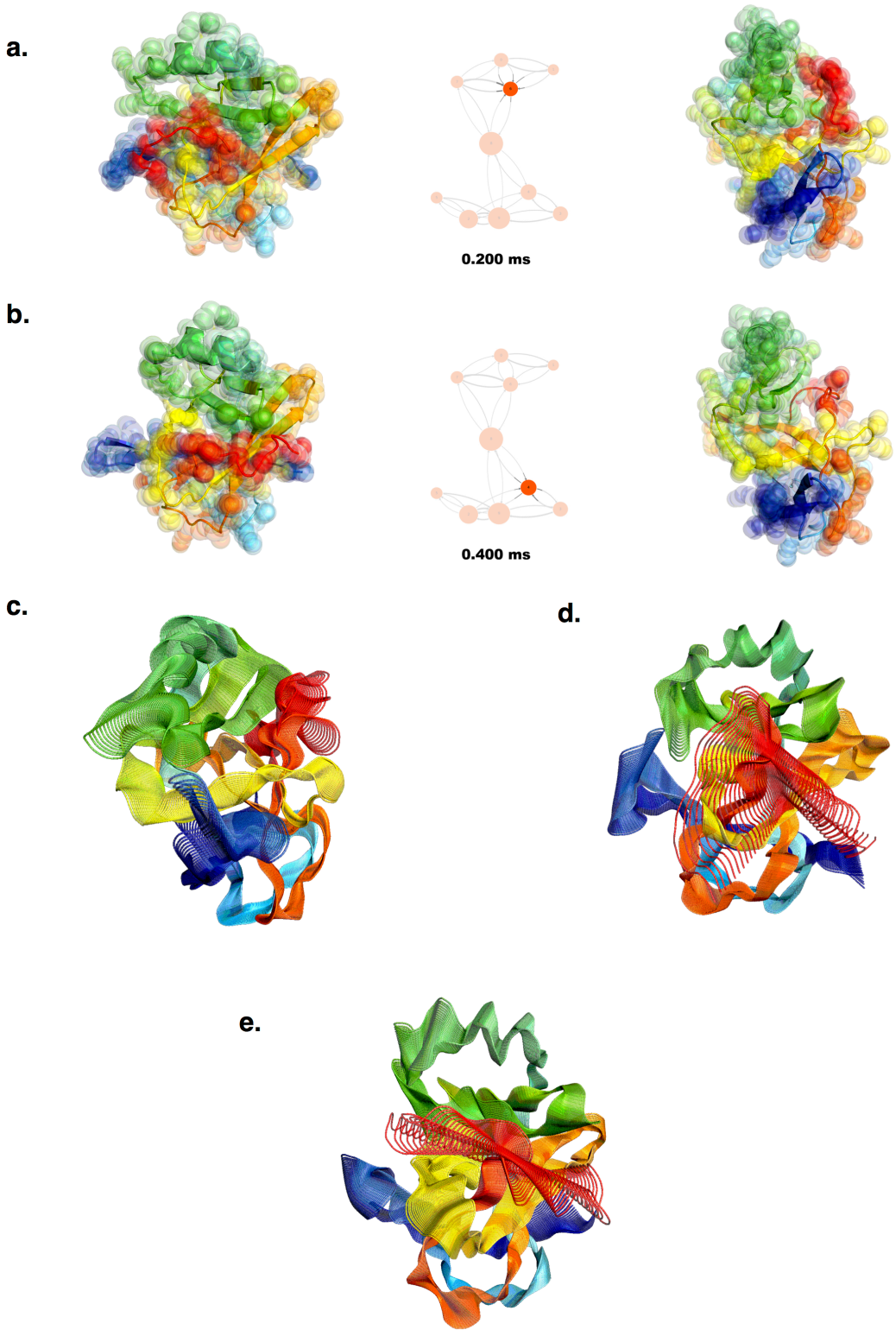


Figure 4. Hidden Markov Model of the EZH2 SET domain from 5 ms of molecular dynamics simulation. Initial simulations of the EZH2 SET domain starting from the PRC2 conformation and an inactive apo conformation resulted in ~ 5 milliseconds (10 million conformations, >200 GPU-years) of

molecular dynamics data. **a., b.** Conformations of EZH2 from a millisecond-long simulated trajectory, separated by 200 microseconds. The two conformations reside in macrostates placed in two different lobes of the 10-state Hidden Markov Model, shown in the middle. **c.**, **d.**, **e.** The three slowest relaxation modes of the model, highlighting the dynamics of the SET-I loop and the post-SET.

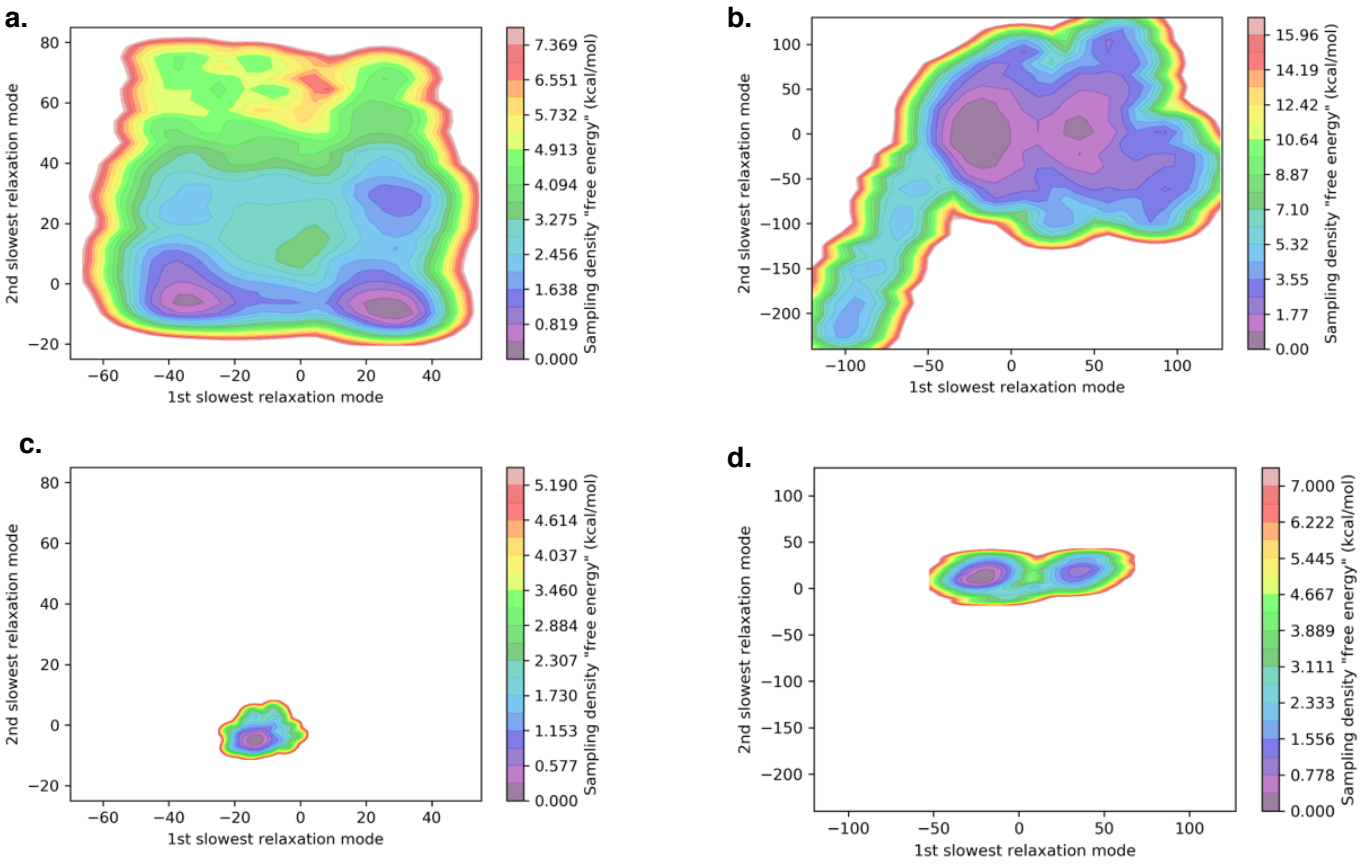
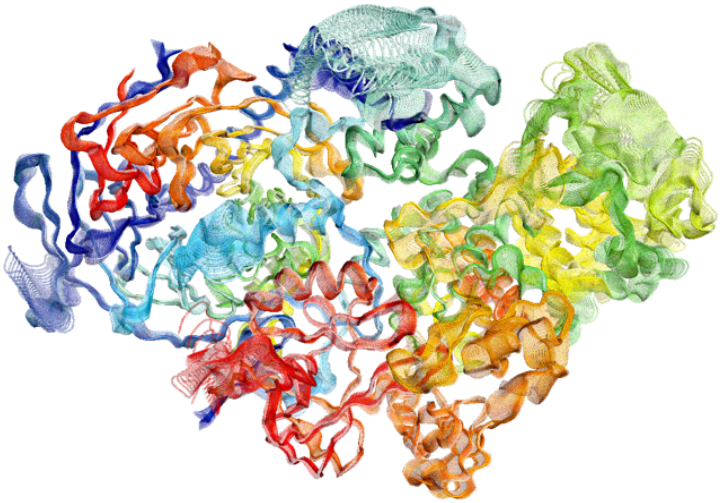


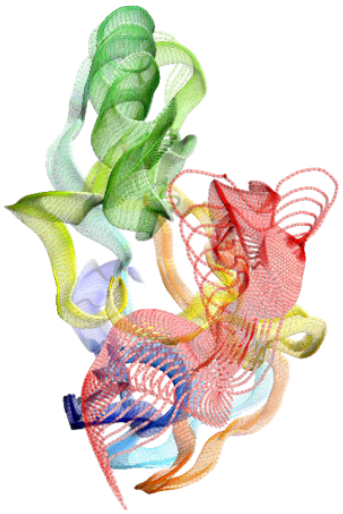
Figure 5. Comparison of the apo and in-complex slowest relaxation modes of the EZH2 and EED components of the PRC2 complex from μs-to-ms timescale molecular dynamics simulations. Dynamics of EZH2 and EED within the PRC2 complex were described using the same state space as for the apo proteins, highlighting the conformational restrictions within the complex, which can be used to design probes stabilizing conformations incompatible with complex formation. **a, b.** Apo data projected onto the two slowest relaxation modes of the apo EZH2 (**a**.) and EED (**b**.), **c,d**. In-complex data projected onto the same apo relaxation modes. **e.** Diversity of conformations along the slowest relaxation mode of the whole PRC2 complex (composed of EZH2, EED, and SUZ12). **f.,h**., (next page) Diversity of conformations along the slowest relaxation mode of the apo proteins (EZH2 and EED respectively). **g,i**. Diversity of conformations along the slowest relaxation mode of the in-complex proteins (EZH2 and EED respectively).

e.



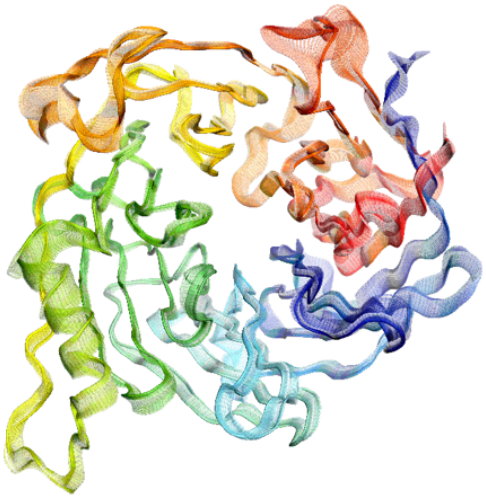
f.

g.



h.

i.



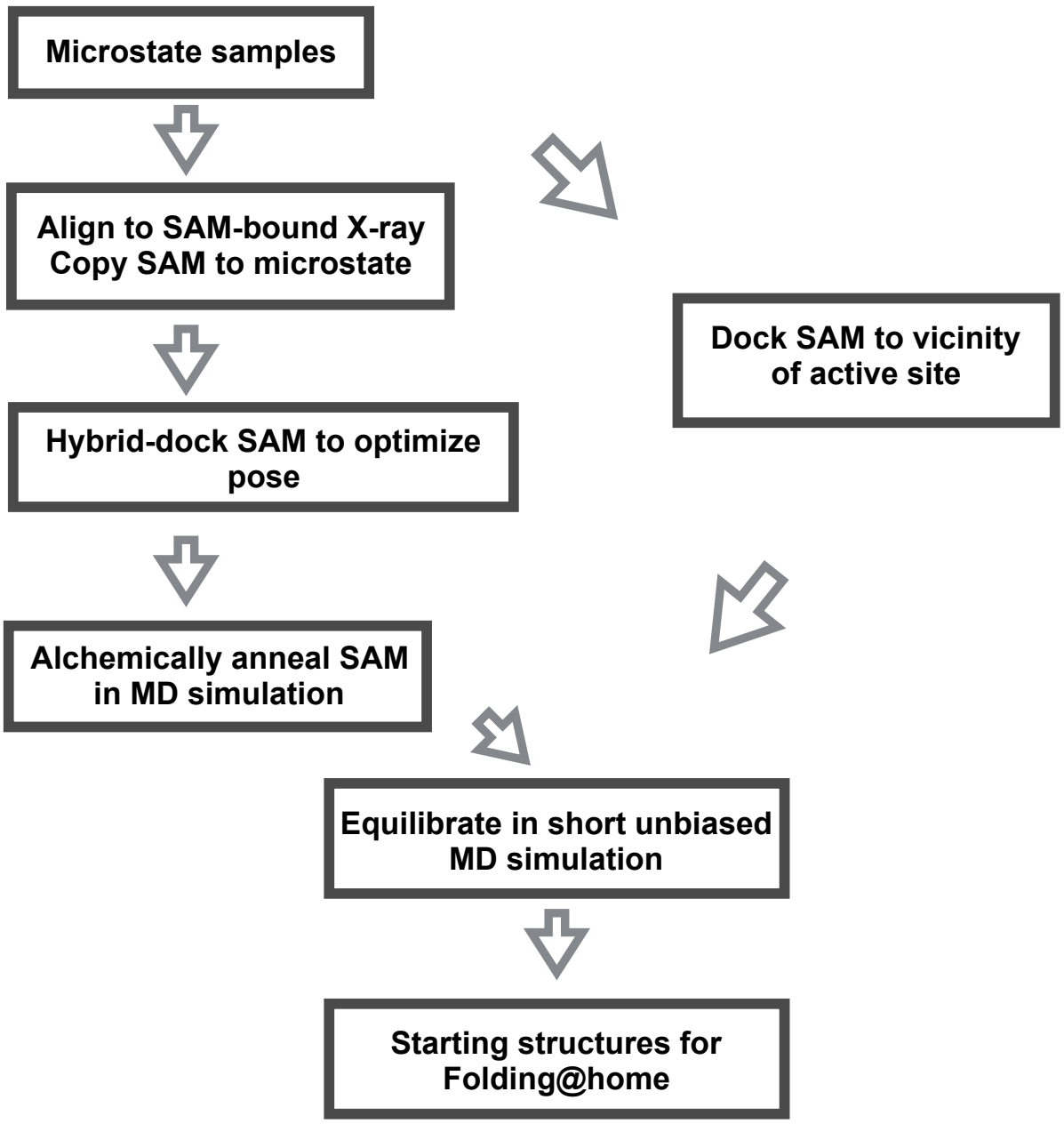


Figure 6. Development and testing of a pipeline to model ligands into states of an apo Markov state model. After construction of an apo Markov state model for a protein of interest, we would like to predict the changes to the conformational landscape due to ligand binding, and hence discover chemical probe candidates. We have developed a pipeline to achieve this in two complimentary ways: a) 'brute force': the ligand is copied into all other conformations into the exact same position, and relaxed alchemically (the interactions of the ligand with the protein are gradually turned on) into a putative binding pose; b) 'standard' docking into the vicinity of the active site or a newly identified pocket, no previous knowledge about the binding mode is used.

What opportunities for training and professional development has the project provided?

Conference attendance

1) Wiewiora, R.P., Chen, S., Luo, M., Chodera, J.D. Conformational dynamics of histone methyltransferase SET8 probed by millisecond-timescale molecular dynamics, Markov state modeling and biochemical experiments. Platform talk in Protein Dynamics & Allostery, Biophysical Society Meeting, San Francisco, February 2018. (Biophysical Journal 116 (3), 183a)

Mentorship by the PI

I mentored an undergraduate summer student and a PhD rotation student, who helped with this project (see PARTICIPANTS).

How were the results disseminated to communities of interest?

Via Twitter by John Chodera and Folding@home with a powerful visualization, where it received quite a significant attention – see e.g. https://twitter.com/jchodera/status/1051128881036632064

What do you plan to do during the next reporting period to accomplish the goals?

I will refine the Markov state models of the target proteins with newly collected data and advanced sampling schemes, and finalize the ensemble pocket detection – docking – free energy calculations pipeline. After the protein models are finalized, I will apply the pipeline to generate top candidate picks to purchase, and protein constructs to express, to finally test the predictions experimentally.

IMPACT

What was the impact on the development of the principal discipline(s) of the project?

In my work, I try to illustrate the potential that computer modeling, when done at the appropriate, expensive, scale holds for future molecular design when computational power of this magnitude becomes routinely available. We have illustrated the potential of the approach taken in this project on a model system, and continue to collect unprecedented amounts of simulation data for analysis.

What was the impact on other disciplines?

Nothing to report.

What was the impact on technology transfer?

Nothing to report.

What was the impact on society beyond science and technology?

Folding@home, as a citizen science project, puts significant visibility on enormous advances anyone can help achieve as part of a community. Many users join the project to make a contribution to progress in studying the diseases that affected their families. All of our science is open, we provide descriptions of our projects for a lay audience in the Folding@home client, and are lucky to have a dedicated community of 'donors' focused at the Folding@home forum.

CHANGES/PROBLEMS

Changes in approach and reasons for change

Additional work: While simulations were collecting data, and before attempting its analysis, additional time was available to extend our previous work on a model system methyltransferase SETD8 to pilot the exact approach used in this project. I determined that the complexity of the automated Markov state modeling pipeline proposed here necessitated tests on a better understood system first. SETD8 is evolutionarily closest to the common ancestor of all PKMTs, hence providing the perfect model system. For that study, our collaborators collected crystal structures of multiple novel conformations of the protein trapped with or without various ligands, providing diverse molecular dynamics simulation seeds. This has provided me with a well sampled, prototypical Markov state model, from which to seed conformations for this project and extract reaction coordinates for adaptive and enhanced sampling simulation runs. This will speed up and expand the possible scope of the work in this project.

Actual or anticipated problems or delays and actions or plans to resolve them

I have not been successful in creating homology models of the SET domain of SETDB1 in acceptable quality. This is due to the fact that it is the only protein lysine methyltransferase to contain an insert of unknown function in the SET domain. I was interested in studying the influence of that insert on the dynamics of the domain, however there is not enough structural data available from similar structures to model it. The Tudor domain of SETDB1 is also of interest for allosteric inhibition, simulations of it have been collecting data, but it is yet unclear how much potential there is in that type of domain to follow the experimental plan of this project. I would ask to consider a change of the second system of interest.

Folding@home maintenance and development (became necessary, but was not anticipated in the SOW): I have been working on updating the simulation 'core' of Folding@home to a new version of the OpenMM engine – which will give us to 50% speedup in computation speed (equivalent of ~$3M

in new hardware), when combined with hydrogen mass repartitioning. This is necessary to achieve reasonable sampling for the PRC2 complex, as well as all the systems containing zinc clusters (e.g. full length EZH2) require features of the new OpenMM to model those clusters. This work has been taking much longer than expected due to instability of the new core seen in testing so far and necessity of further debugging. Further, I have also spent considerable amount of time refining the system of processing data (tens of TBs) from the Folding@home servers. These became necessary due to the limitations of the existing system that were not anticipated in the SOW, and are causing delays to the computational modeling and hence the ability to make decisions on wet laboratory experiments to perform and small molecules to purchase. I would ask to consider a no-cost extension to the project to utilize these advancements fully.

Changes that had a significant impact on expenditures

Nothing to report. (note all wet lab funds will be used in year 2 of the project after completion of the computational analysis of year 1, as per the SOW).

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

N/A

Significant changes in use or care of human subjects

N/A

Significant changes in use or care of vertebrate animals

N/A

Significant changes in use of biohazards and/or select agents

N/A

PRODUCTS

Publications, conference papers, and presentations

Conference talks

1) Wiewiora, R.P., Chen, S., Luo, M., Chodera, J.D. Conformational dynamics of histone methyltransferase SET8 probed by millisecond-timescale molecular dynamics, Markov state modeling

and biochemical experiments. Platform talk in Protein Dynamics & Allostery, Biophysical Society Meeting, San Francisco, February 2018. (Biophysical Journal 116 (3), 183a)

Publications

1) Rafal P. Wiewiora*, Shi Chen*, Fanwang Meng, Nicolas Babault, Anqi Ma, Wenyu Yu, Kun Qian, Hao Hu, Hua Zou, Junyi Wang, Shijie Fan, Gil Blum, Fabio Pittella-Silva, Kyle A. Beauchamp, Wolfman Tempel, Hualiang Jiang, Kaixian Chen, Robert Skene, Y. George Zheng, Peter J. Brown, Jian Jin, Cheng Luo, John D. Chodera, Minkui Luo (*co-first).

The Dynamic Conformational Landscapes of the Protein Methyltransferase SETD8. bioRxiv 438994; doi: https://doi.org/10.1101/438994 (accepted, in revision at eLife; federal funding acknowledged).

Website(s) or other Internet site(s)

Code for the SETD8 modeling, Github: https://github.com/choderalab/SETD8-materials

SETD8 simulation data (6 ms), Open Science Framework: https://osf.io/2h6p4/

Technologies or techniques

I have shown an efficient automatic way to construct Markov state models and illustrated a rigorous analysis – these techniques will be immediately useful to the field.

Inventions, patent applications, and/or licenses

Nothing to report.

Other Products

SETD8 simulation data (6 ms), Open Science Framework: https://osf.io/2h6p4/

PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

1.

Name: Rafal Wiewiora

Project role: PI

ORCID: 0000-0002-8961-7183

Nearest person month worked: 12

Contribution to project: Performed all work, except work done by persons 2 and 3, which I supervised.

2.

Name: Hersh Gupta

Project role: undergraduate summer student

ORCID: -

Nearest person month worked: 2

Contribution to project: Wrote software for the small molecule hybrid docking pipeline.

Funding support: Memorial Sloan Kettering Cancer Center

3.

Name: Chloe Burnside

Project role: PhD rotation student

ORCID: -

Nearest person month worked: 1

Contribution to project: Tested the small molecule docking pipeline.

Funding support: Tri-Institutional PhD Program in Chemical Biology, Weill Cornell Medicine

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Nothing to report.

What other organizations were involved as partners?

Nothing to report.

SPECIAL REPORTING REQUIREMENTS

N/A

APPENDICES

The Dynamic Conformational Landscapes of the Protein Methyltransferase SETD8

Shi Chen[1,2,#], Rafal P. Wiewiora[1,3,#], Fanwang Meng[4,¶], Nicolas Babault[5], Anqi Ma[5], Wenyu Yu[6], Kun Qian[7], Hao Hu[7], Hua Zou[8], Junyi Wang[2], Shijie Fan[4,9], Gil Blum[2], Fabio Pittella-Silva[2], Kyle A. Beauchamp[3], Wolfram Tempel[6], Hualiang Jiang[4,9], Kaixian Chen[4,9], Robert Skene[8], Y. George Zheng[7], Peter J. Brown[6], Jian Jin[5], Cheng Luo[4,9,*], John D. Chodera[3,*], and Minkui Luo[2,10,*]

1. Tri-Institutional PhD Program in Chemical Biology, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA
2. Chemical Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA
3. Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA
4. Drug Discovery and Design Center, CAS Key Laboratory of Receptor Research, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China
5. Mount Sinai Center for Therapeutics Discovery, Departments of Pharmacological Sciences and Oncological Sciences, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
6. Structural Genomics Consortium, University of Toronto, Toronto, Ontario, M5G 1L7, Canada
7. Department of Pharmaceutical and Biomedical Sciences, University of Georgia, Athens, Georgia, 30602, USA
8. Takeda California, 10410 Science Center Drive, San Diego, CA 92121, USA
9. University of Chinese Academy of Sciences, 19 Yuquan Road, Beijing 100049, China
10. Program of Pharmacology, Weill Cornell Medical College of Cornell University, New York, New York 10021, USA

¶ Current address: Department of Chemistry and Chemical Biology, McMaster University, Ontario, L8S 4L8, Canada

# These authors made equal contribution

* Corresponding authors: luom@mskcc.org; john.chodera@choderalab.org; cluo@simm.ac.cn.

Abstract: Elucidating conformational heterogeneity of proteins is essential for understanding protein functions and developing exogenous ligands for chemical perturbation. While structural biology methods can provide atomic details of static protein structures, these approaches cannot in general resolve less populated, functionally relevant conformations and uncover conformational kinetics. Here we demonstrate a new paradigm for illuminating dynamic conformational landscapes of target proteins. SETD8 (Pr-SET7/SET8/KMT5A) is a biologically relevant protein lysine methyltransferase for *in vivo* monomethylation of histone H4 lysine 20 and nonhistone targets. Utilizing covalent chemical inhibitors and depleting native ligands to trap hidden high-energy conformational states, we obtained diverse novel X-ray structures of SETD8. These structures were used to seed massively distributed molecular simulations that generated

six milliseconds of trajectory data of SETD8 in the presence or absence of its cofactor. We used an automated machine learning approach to reveal slow conformational motions and thus distinct conformational states of SETD8, and validated the resulting dynamic conformational landscapes with multiple biophysical methods. The resulting models provide unprecedented mechanistic insight into how protein dynamics plays a role in SAM binding and thus catalysis, and how this function can be modulated by diverse cancer-associated mutants. These findings set up the foundation for revealing enzymatic mechanisms and developing inhibitors in the context of conformational landscapes of target proteins.

Introduction

Proteins are not static, but exist as an ensemble of conformations in dynamic equilibrium[1]. Characterization of conformational heterogeneity can be an essential step towards interpreting function, understanding pathogenicity, and exploiting pharmacological perturbation of target proteins[2-4]. Conventional efforts to map functionally relevant conformations rely on biophysical techniques such as X-ray crystallography[5], nuclear magnetic resonance (NMR)[6], and cryo-electron microscopy[7], which provide static snapshots of highly-populated conformational states. While complementary techniques such as relaxation-dispersion NMR can resolve a limited number of low-population states, they are incapable of providing detailed structural information[8]. By contrast, molecular simulations provide atomistic detail---a prerequisite to structure-guided rational ligand design---and insight into relevant conformational transitions[1]. The emergence of Markov state models (MSMs) has shown the power of massively distributed molecular simulations in resolving complex kinetic landscapes of proteins[9,10]. By integrating simulation datasets with MSMs, functionally relevant conformational dynamics as well as atomistic details can be extracted[10]. Recently, MSMs have been used to identify key intermediates for enzyme activation[11,12] and allosteric modulation[13]. However, these approaches are limited by the number of seed structures and timescales accessible by molecular simulations (generally microseconds for one structure) relative to the reality of complicated conformational transitions (up to milliseconds for multiple structures)[14]. To overcome the limitations of individual techniques, we envisioned an integrated approach that combines simulation with experiment to characterize

conformational landscapes of enzymes and elucidate their functions with the consideration of dynamic conformations.

Protein lysine methyltransferases (PKMTs) comprise a subfamily of posttranslational modifying enzymes that transfer a methyl group from the cofactor *S*-adenosyl-L-methionine (SAM)[15]. PKMTs play epigenetic roles in gene transcription, cellular pluripotency, and organ development[16,17]. Their dysregulation has been implicated in neurological disorders and cancers[18,19]. SETD8 (SET8/Pr-SET7/KMT5A) is the sole PKMT annotated for monomethylation of histone H4 lysine 20 (H4K20me)[20,21] and many non-histone targets such as the tumor suppressor p53 and the p53-stabilizing factor Numb[22,23]. Disruption of endogenous SETD8 leads to cell cycle arrest and chromatin decondensation, consistent with essential roles for SETD8 in transcriptional regulation and DNA damage response[24-26]. SETD8 has also been implicated in cancer invasiveness and metastasis[27]. High expression of SETD8 is associated with pediatric leukemia and its overall low survival rate[28]. As a result, there is enormous interest in elucidating functional roles of SETD8 in disease and developing pharmacological agents to perturb this target[29-31].

Given the essential roles of conformational dynamics in enzymatic catalysis[1,32] and our current limited knowledge of conformational landscapes of PKMTs, we envisioned leveraging an integrated experimental-computational approach to characterize dynamic conformational landscapes of SETD8 and its cancer-associated mutants with atomic resolution. To access previously-unseen, less-populated conformational states of SETD8 to seed massively parallel distributed molecular dynamics (MD) simulations, we envisioned trapping these conformations with small-molecule ligands. Here we solved four distinct crystal structures of SETD8 in alternative ligand-binding states with covalent SETD8 inhibitors and native ligands. With the aid of these new structures, we generated an aggregate of six milliseconds of explicit solvent MD simulation data for apo- and SAM-bound SETD8. Using a machine learning approach to select features and hyperparameters for MSMs via extensive cross-validation, we identified 24 kinetically distinct metastable conformational states of apo-SETD8 and determined how the conformational landscape is remodeled upon SAM binding. We then validated these conformational landscapes with stopped-flow kinetics and isothermal titration calorimetry by examining SAM binding, characterizing rationally-designed SETD8 variants with increased

catalytic efficiency, and resolving multiple timescales associated with transitions among these conformers. The resulting model furnishes unprecedented key insights on how these dynamic conformations play a role in catalysis and how cancer-associated SETD8 mutants alter this process.

Results

Crystal structures of SETD8 associated with hidden conformations. To identify hidden high-energy conformational states of SETD8, we envisioned a strategy of trapping the associated conformers with small-molecule ligands. The development of high-affinity SETD8 inhibitors with canonical target-engagement modes is challenging[29], and led us to exploit covalent inhibitors[31,33]. These compounds can overcome the high energy penalties associated with hidden high-energy conformers through the irreversible formation of energetically-favored inhibitor-SETD8 adducts. Our prior efforts led to the development of covalent inhibitors containing 2,4-diaminoquinazoline arylamide and multi-substituted quinone scaffolds by targeting Cys311[31,33]. Upon further optimization of these scaffolds, we identified MS4138 (**Inh1**) and SGSS05NS (**Inh2**)[34], two structurally distinct covalent inhibitors with the desired potency against SETD8 (**Figures 1a**, **S9**). X-ray crystal structures of SETD8 were then solved in complex with **Inh1** and **Inh2**, respectively (**Figures 1b**,**c**, **S10**, **S11**). Notably, despite the overall structural similarity of the pre-SET, SET, and SET-I motifs, the **Inh1**- and **Inh2**-SETD8 binary complexes (**BC-Inh1** and **BC-Inh2**) differ from the SETD8-SAH-H4 ternary complex (**TC**)[35-37] by the distinct conformations of their post-SET motifs. The post-SET motif of **TC** was characterized by its U-shaped topology with a double-kinked loop-helix-helix architecture, which appears to be optimally oriented for binding both SAM and a peptide substrate (**Figure 1c**,**d**)[35-37]. In comparison, **BC-Inh1** and **BC-Inh2** rotate their post-SET motifs by 140 ° and 60 °, respectively (**Figure 1d**). Moreover, the post-SET motifs of **BC-Inh1** and **BC-Inh2** adapt more extended configurations with a less structured loop and a singly-kinked helix, respectively (**Figure 1c**,**d**). Whereas multiple factors may influence the overall conformations, the formation of Cys311 adducts likely made the key contribution to the discovery of these hidden post-SET motif conformers.

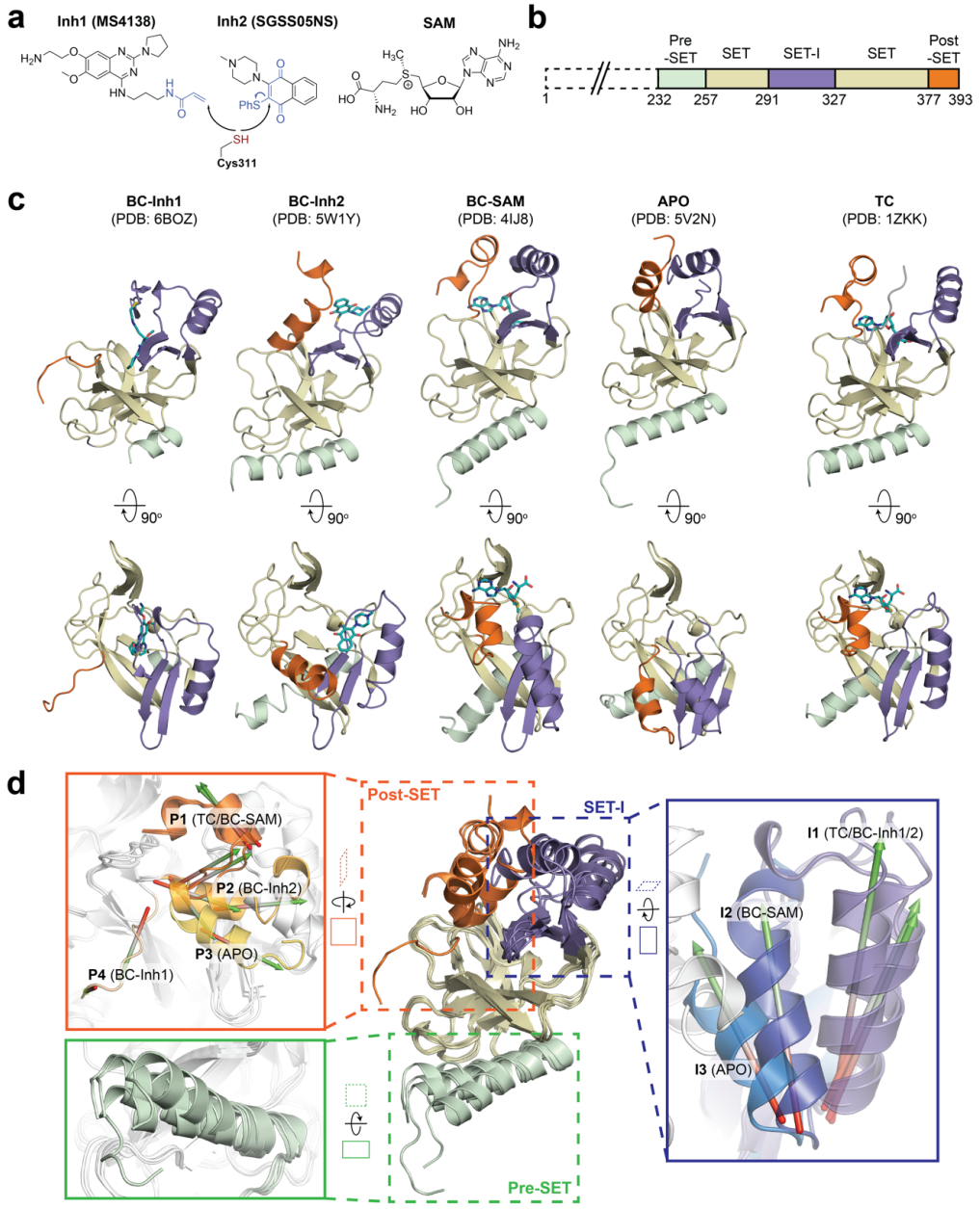


Figure 1. Diverse SETD8 conformations captured in altered ligand-binding states. a, Structures of SETD8 ligands involved in this work. Two covalent inhibitors targeting Cys311 (MS4138 as **Inh1** and SGSS05NS as **Inh2**) and the cofactor SAM were used as ligands to trap neo-conformations of SETD8. **b**, Domain topology of SETD8. Four functional motifs at SETD8's catalytic domain are colored: pre-SET (light green), SET (dark yellow), SET-I (purple), and post-SET (orange). **c**, Cartoon representations of

four neo-structures of SETD8 (**BC-Inh1**, **BC-Inh2**, **BC-SAM**, and **APO**) and a structure of a SETD8-SAH-H4 ternary complex (**TC**). These structures are shown in two orthogonal views with ligands, pre-SET, SET, SET-I, and post-SET colored in cyan, light green, dark yellow, purple, and orange, respectively. **d,** Superposition of five crystal structures highlighted with detailed views of post-SET, SET-I, and pre-SET motifs. The five X-ray structures reveal four distinct conformational states of the post-SET motif (**P1-4**) and three distinct conformational states of the SET-I motif (**I1-3**).

To reveal additional hidden conformers that are structurally distinct from **TC**, we also solved crystal structures of SETD8 upon depleting native ligands and obtained structures of the SAM-SETD8 binary complex (**BC-SAM**) and apo-SETD8 (**APO**) (**Figures 1c**, **S12**, **S13**). Strikingly, **BC-SAM** and **APO** differ from **TC** by their distinct SET-I motifs in the context of the otherwise similar SET-domain (**Figure 1d**). Furthermore, the post-SET motif of **APO** structurally resembles an intermediate state between **BC-Inh1** and **BC-Inh2** but is distinct from those of **BC-SAM** and **TC** (**Figure 1d**). In contrast to the structurally diverse SET-I (**I1-3**) and post-SET motifs (**P1-4**) in these structures, their pre-SET motifs show only slightly altered configuration (**Figure 1d**). The differences between these structures highlight the conformational plasticity of the SET-I and post-SET motifs. Collectively, these observations provide strong structural rationale for the existence of a highly dynamic conformational landscape of SETD8.

Hidden conformations of apo-SETD8 revealed by structural chimeras. The **BC-SAM**, **BC-Inh1**, **BC-Inh2**, **APO**, and **TC** structures can be readily classified into three distinct SET-I configurations (**I1-3**) and four distinct post-SET configurations (**P1-4**) (**Figure 1d**). Given the relative independence between the SET-I and post-SET motifs, we expected that additional combinations of discrete motifs can represent yet-unobserved functionally relevant conformations of SETD8. We thus constructed putative "structural chimeras" of apo-SETD8 containing orthogonal **I1-3** and **P1-4** in a combinatorial (3×4) manner (**Figures 2a**, **S14**). Among the twelve structural chimeras as potential seeds for MD simulations, five were crystallographically-determined conformers (**BC-Inh1**, **BC-Inh2**, **BC-SAM**, **TC** with ligands removed, and **APO**), four were new structurally-chimeric conformers, and three were excluded because of obvious steric clashes (**Figures 2a**, **S15**).

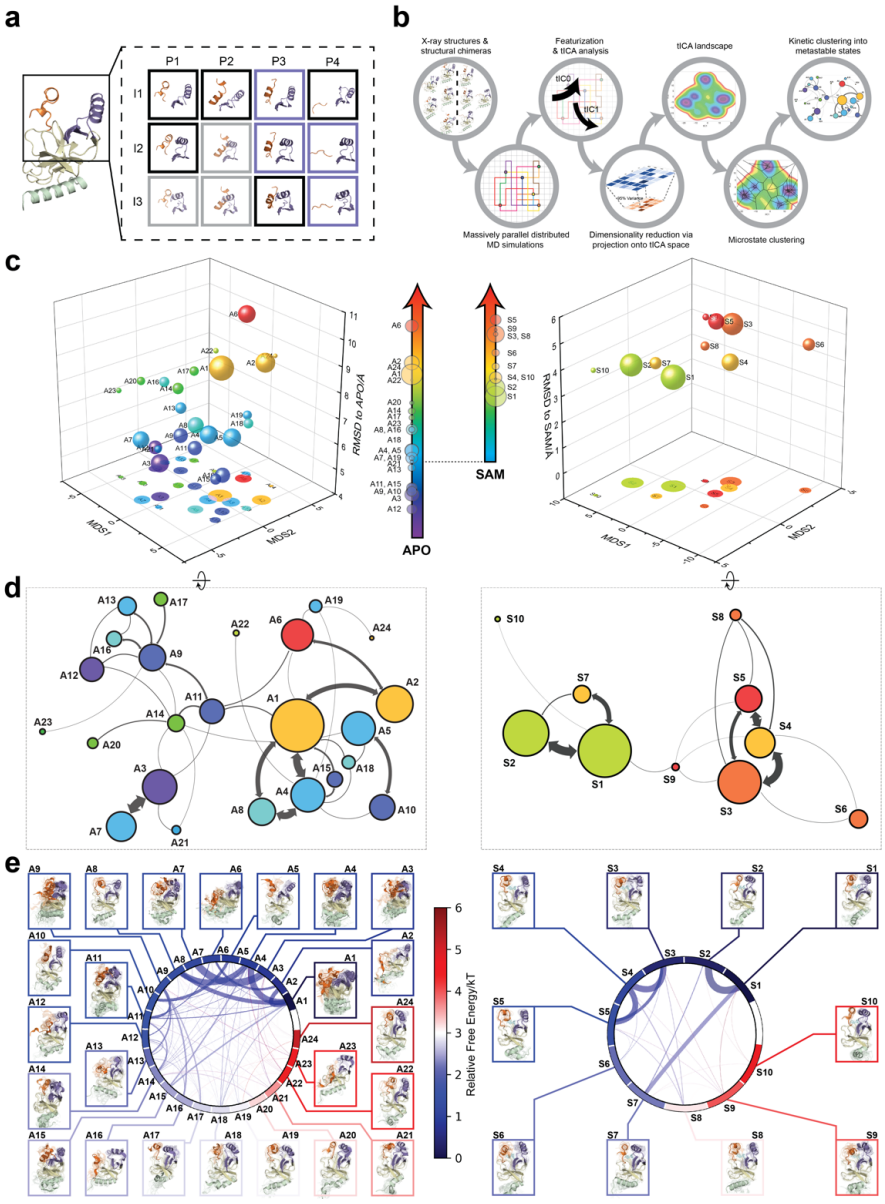


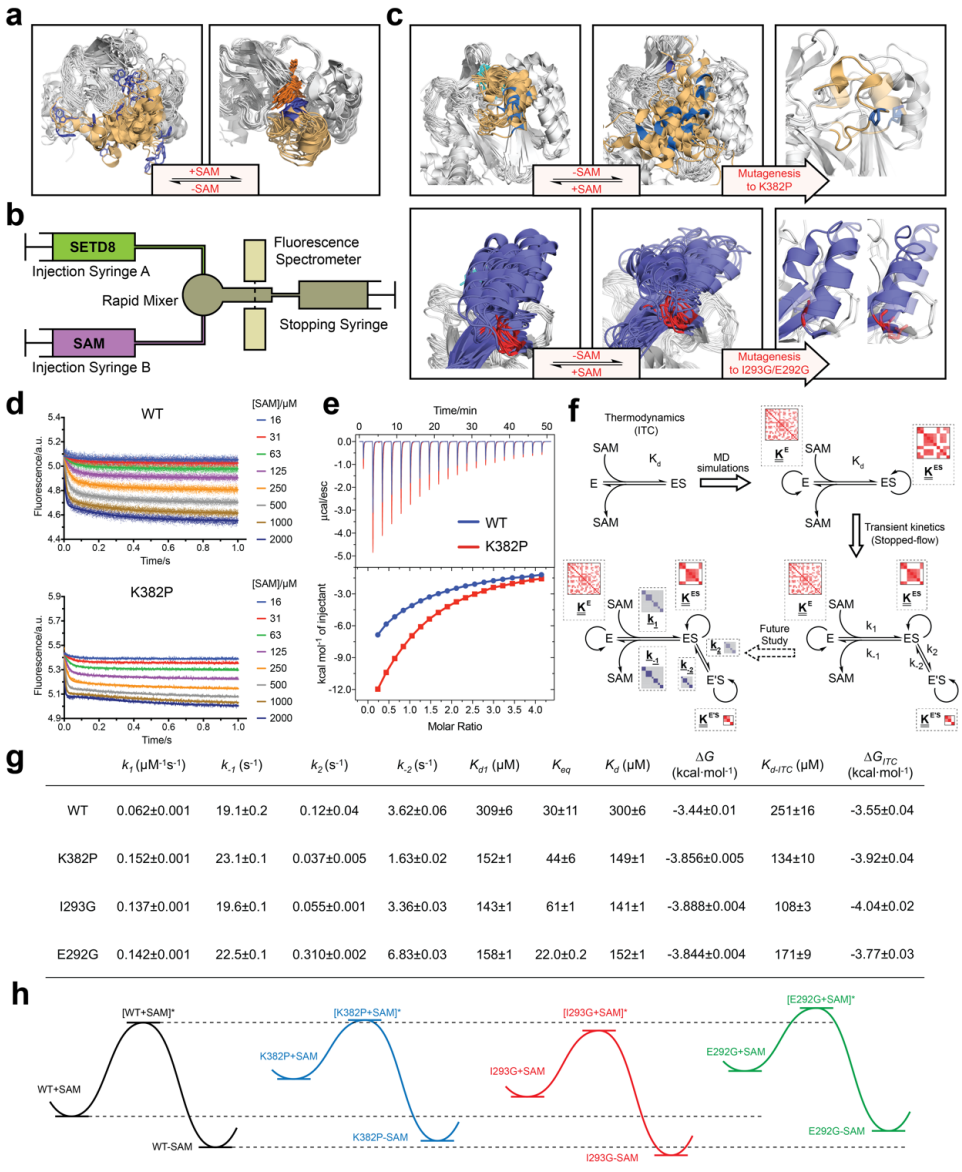
Figure 2. Markov state models and conformational landscapes of apo- and SAM-bound SETD8 constructed through diversely seeded, massively parallel molecular dynamics simulations. a, Combinatorial construction of structural domain chimeras using crystallographically-derived post-SET and SET-I conformations. Each conformer is boxed and color-coded with black for five X-ray-derived structures, blue for four putative structural chimeras included as seed structures for MD simulations, and grey for three structural chimeras excluded from MD simulations because of obvious steric clashes. **b**,

Schematic workflow to construct dynamic conformational landscapes via MSM. The five X-ray structures and the four structural chimeras were used to seed massively parallel MD simulations on Folding@home (see Method). Markov state models were constructed from these MD simulation results to reveal the conformational landscape. **c–e**, Kinetically metastable conformations (macrostates) obtained from kinetically coupled microstates via Hidden Markov Model (HMM) analysis. The revealed dynamic conformational landscapes consist of 24 macrostates for apo-SETD8 (left panel) and 10 macrostates for SAM-bound SETD8 (right panel). **c**, Kinetic and structural separation of macrostates in a 3D scatterplot. The X, Y axes represent kinetic separation of macrostates with a log-inverse flux kinetic embedding method (see Methods). The Z axis reports RMSDs of each macrostate to **APO** (left) or **BC-SAM** (right). The relative population of each macrostate of apo- or SAM-bound SETD8 ensembles is proportional to the volume of each representative sphere. **d**, Cartoon depiction of macrostates in a 2D scatterplot. The relative positions of metastable conformations were derived via the log-inverse flux kinetic embedding (see Methods). The diameter of the corresponding circle in the 2D scatterplot is proportional to the diameter of the respective sphere in the 3D scatterplot above. Equilibrium kinetic fluxes larger than 7.14×10^{2} s^{-1} for apo- and 1.39×10^{3} s^{-1} for SAM-bound SETD8 are shown for interconversion kinetics with thickness of the connections proportional to fluxes between two macrostates. **e**, Chord diagrams and representative conformers of macrostates. The colors are encoded for the free energy of each macrostate relative to the lowest free energy of the macrostates. The equilibrium flux between two macrostates is proportional to thickness of respective arcs.

Dynamic conformational landscape of apo-SETD8 via Markov state modeling from 5-ms MD simulation dataset. With seed conformations prepared as above, we envisioned illuminating the conformational landscape with massively distributed long-time MD simulations and resolving its kinetic features with Markov state models (MSMs) (**Figures 2b**, **S14**). We conducted approximately 500×1 μs explicit-solvent MD simulations from each seed and accumulated 5 milliseconds of aggregate data in 10 million conformational snapshots for apo-SETD8 (**Figures S16**, **Table S3**). To identify functionally relevant conformational states and their transitions, we built MSMs using a pipeline that employs machine learning and extensive hyperparameter optimization to identify slow degrees of freedom and structural and kinetic criteria to cluster conformational snapshots into discrete conformational states (**Figures S17-24**, **Tables S4**, **S5**)[38]. This approach identified 24 kinetically metastable conformations (macrostates) from an optimized, cross-validated set of 100 microstates (**Figures 2c**, **S25-30**, **Tables S6**, **S7**).

These macrostates are remarkably diverse, spanning up to 10.5 Å Cα RMSD from **APO**. To visualize the kinetic relationships between functionally important conformations, dimensionality reduction was used to project the landscape into 2D while preserving log inverse fluxes between states (**Figure 2d**). The relative populations of these macrostates and their interconversion kinetics were calculated on the basis of their transition fluxes, resolving rare conformational states up to 6 kT in free energy (**Figure 2d**,**e**).

The dynamic conformational landscape of SAM-bound SETD8. Given the success in constructing the dynamic conformational landscape of apo-SETD8, we applied the same strategy to SAM-bound SETD8. With the two crystal structures of SETD8 in complex with SAM (**BC-SAM** and **TC**) as the seed conformations, we conducted ~ 500×1 μs explicit-solvent MD simulations from each structure and accumulated 1 millisecond of aggregate data (2M snapshots) (**Figure S25**). The resulting MSM for SAM-bound SETD8 contained 10 kinetically metastable macrostates arising from 67 microstates (**Figure S31**, **Tables S8**, **S9**). Similar to those of apo-SETD8, the relative macrostate populations of SAM-bound SETD8 and their flux kinetics were computed and embedded into 3D/2D scatter plots and chord diagram (**Figure 2c**,**d**,**e**). The smaller number of metastable states identified for SAM-bound SETD8 is expected given that SAM binding restricts conformational accessibility.



	k_1 ($\mu M^{-1}s^{-1}$)	k_{-1} (s^{-1})	k_2 (s^{-1})	k_{-2} (s^{-1})	K_{d1} (μM)	K_{eq}	K_d (μM)	ΔG (kcal·mol^{-1})	$K_{d\text{-}ITC}$ (μM)	ΔG_{ITC} (kcal·mol^{-1})
WT	0.062±0.001	19.1±0.2	0.12±0.04	3.62±0.06	309±6	30±11	300±6	-3.44±0.01	251±16	-3.55±0.04
K382P	0.152±0.001	23.1±0.1	0.037±0.005	1.63±0.02	152±1	44±6	149±1	-3.856±0.005	134±10	-3.92±0.04
I293G	0.137±0.001	19.6±0.1	0.055±0.001	3.36±0.03	143±1	61±1	141±1	-3.888±0.004	108±3	-4.04±0.02
E292G	0.142±0.001	22.5±0.1	0.310±0.002	6.83±0.03	158±1	22.0±0.2	152±1	-3.844±0.004	171±9	-3.77±0.03

Figure 3. Biochemical characterization of gain-of-function mutations revealed by conformational landscapes of SETD8. a, Comparison of binding environments of Trp390 between apo and SAM-bound SETD8 in the context of their dynamic conformational landscapes. **b**, Illustration of rapid-quenching stopped-flow experiments. These experiments were conducted to trace fluorescence changes of Trp390 upon SAM binding. **c**, Comparison of the conformations of post-SET kink and SET-I helix between apo and SAM-bound SETD8 in the context of their dynamic conformational landscapes. Analysis of key structural motifs indicated K282P, I293G and E292G as potential gain-of-function variants. **d**,

Fluorescence changes of wild-type and K382P SETD8 traced with a rapid-quenching stopped-flow instrument within 1 s upon SAM binding. **e**, SAM-binding ITC enthalpogram of wild-type and K382P SETD8. **f**, Stepwise SAM-binding of SETD8 in the integrative context of biochemical, biophysical, structural, and simulation data. ITC determines the thermodynamic constant of SAM binding by SETD8. MD simulations and MSM uncover metastable conformations and interconversion rates of apo- and SAM-bound SETD8 ($\underline{K}^{apo}$ and $\underline{K}^{SAM}$). Stopped-flow experiments revealed that SETD8 binds SAM via biphasic kinetics. Rate constants uncovered by stopped-flow experiments (k_1, k_{-1}, k_2, k_{-2}) represent macroscopic rates of SAM binding by SETD8 with multiple metastable conformations. The microscopic behavior of individual metastable states and corresponding rates ($\underline{k}_1$, $\underline{k}_{-1}$, $\underline{k}_2$, $\underline{k}_{-2}$) have not been resolved. Transition probability matrices (red) and microscopic rate constant matrices (blue) are shown as colored grids. A rigorous mathematical model for these processes is shown in **Figure S36**. **g**, Kinetic and thermodynamic constants of wild-type SETD8 and its mutants. For k_1, k_{-1}, k_2, k_{-2}, data are best fitting values ± standard error (s.e.) from KinTek. For $K_{d\text{-}ITC}$, data are mean ± s.e. of at least 3 replicates. K_{d1}, K_{eq}, and K_d are calculated based on equations in online method. Uncertainties of K_{d1}, K_{eq}, K_d, and ΔG are s.e. calculated by the propagation of uncertainties from individual rate constants and dissociation constants, respectively. **h**, Relative energy landscapes of apo- and SAM-bound SETD8 and its gain-of-function mutants.

Experimental validation of the conformational landscapes of SETD8. Upon uncovering the dynamic conformational landscapes of apo- and SAM-bound SETD8, we were able to extract new structural information and designed experiments to further validate this model (**Figure 3**). Comparison of the conformational ensembles between apo- and SAM-bound SETD8 revealed that SAM binding dramatically alters the environment of Trp390 (**Figure 3a**, blue sticks), the sole tryptophan residue in the catalytic domain of SETD8. This residue is flexible and mainly solvent-exposed in apo-SETD8 conformational ensembles but restricted in a hydrophobic environment through SAM-mediated pi-pi stacking in SAM-bound SETD8 conformational ensembles (**Figure 3a**). Such environmental changes upon SAM binding are expected to quench fluorescence of Trp390[39]. To verify this prediction, we designed rapid-mixing stopped-flow kinetic experiments with 5 ms dead time and 0.1 ms resolution to track the fluorescence change of Trp390 upon SAM binding (**Figure 3b**). We observed SAM-dependent biphasic kinetics of the fluorescence decrease within 1 s with > 80% of the change occurring in the fast phase (0 – 0.1 s) (**Figure 3d**). In the context of the conformational landscape of apo-SETD8, we interpreted

the major decrease in fluorescence intensity (fast-phase kinetics) as a consequence of the collective changes of Trp390 from the solvent-exposed hydrophilic environment in apo conformations to the hydrophobic environment in SAM-bound conformations (**Figure 3a**,**c**). In contrast, the minor decrease in fluorescence intensity (slow-phase kinetics) reflects the slow conformational changes of Trp390 in the SAM-bound SETD8 conformational ensembles (**Figure 3d**). With unsupervised global fitting to this two-step model, we obtained forward and reverse rate constants for the fast- and slow-phase kinetics, which are in agreement with conventional fitting to double exponential kinetics[40] (**Figures 3d**,**f**,**g**, **S32**, **Table S10**). The k_{-1} value was also confirmed independently by rapid-mixing stopped-flow dilution of SAM-bound SETD8[41] ("ES+E'S", **Figure S33**, **Table S10**). Here the k_{-1}/k_1 ratio of 309±6 μM corresponds to the average SAM dissociation constant K_{d1} of apo-SETD8 conformers, which is consistent with independently determined ITC K_d of 251±16 μM (**Figures 3e**,**f**, **S34**). In contrast, the large k_{-2}/k_2 ratio of 30±11 suggests that the second phase corresponds to a slow equilibrium between ES and E'S with minimal contribution of E'S to the overall SAM dissociation constant K_d (**Figure 3e**). The conformational ensembles we identified for apo- and SAM-bound SETD8 demonstrate the statistical nature of its SAM-binding process. Therefore, the observed fluorescence changes and herein determined macroscopic kinetic constants represent an ensemble-weighted average of microscopic behaviors of all species that exist in the solution. A rigorous mathematical description of microscopic kinetics of SAM binding was thus obtained under the consideration of interconversion of the metastable conformational states of apo- and SAM-bound SETD8 (**Figure S36**).

We then proposed to confirm our understanding of functionally-relevant conformations and their thermodynamics by identifying SETD8 variants with increased affinity for SAM. We uncovered a collection of characteristic kink motifs around Lys382 in the post-SET motif of SAM-bound SETD8 conformational ensembles (**Figure 3c**), while this region is less structured in apo-SETD8 conformational ensembles. We hypothesized that a proline mutation (K382P) could better stabilize the conformational ensembles of SAM-bound SETD8 than apo-SETD8 (**Figure 3c,h**). We also identified a characteristic α-helix in the SET-I motif, which adapts flexible and diverse configurations in apo ensembles but constrained and structurally distorted configurations in SAM-bound ensembles (**Figure 3c**). We proposed that the replacement of I293 or E292 adjacent to the α-helix with a flexible glycine should relax this distortion to better

stabilize SAM-bound ensembles (**Figure 3c**,**h**). We therefore characterized the SAM-binding kinetics and affinities of K382P, I293G, and E292G variants of SETD8 with stopped-flow kinetics and ITC (**Figures 3c**,**d**,**e**,**f**, **S32-34**). While exhibiting biphasic kinetics similar to that of wild-type SETD8, the stopped-flow mixing experiment revealed the three variants showed a significant two-fold decrease of $K_{d,SAM}$ (**Figure 3d**,**e**). The stopped-flow data further revealed that the two-fold change of $K_{d,SAM}$ mainly arises from increased SAM-binding rates k_1 with relatively unchanged k_{-1} (**Figure 3g**). These results are consistent with independently-determined K_d and k_{-1} from ITC and stopped-flow dilution, respectively (**Figures 3e**,**f**, **S33**, **S34**, **Table S10**). Collectively, these observations confirm the robustness of our conformational landscape model for apo- and SAM-bound SETD8.

Effects of key simulation parameters on construction of conformational landscapes. We systematically investigated how the choices of seed structures and simulation time---key computational parameters---influence microstate discovery and quality of conformational landscapes of SETD8 (**Figure 4**). The simulations of apo-SETD8 initiated from any single X-ray structure (**BC-Inh1**, **BC-Inh2**, **BC-SAM**, **APO**, or **TC** in **Figure 1c**) only reveal a partial conformational landscape (28–61% microstate coverage, **Figure 4a**). To achieve >90% microstate coverage, at least two crystal structures---**BC-SAM** in combination with either **BC-Inh1** or **BC-Inh2**---must be included (**Figure 4a**). If three crystal structures are included, **BC-SAM** in combination with **TC** and **APO** can provide >90% coverage (**Figure 4a**). In terms of the structural motifs (**I1-3** or **P1-4**, **Figures 1d**, **2a**), simulations originating from the SET-I motif **I1**, **I2**, or **I3** alone led to the discovery of 69, 58, or 39 of the 100 microstates, respectively (**Figure 4b, Table S15**). The combination of **I1** and **I2** is sufficient to cover all 100 microstates, arguing for the redundant character of **I3**. For the post-SET motif, any combination of two post-SET configurations except **P2**–**P3** leads to >90 microstate coverage (**Figure 4b, Table S15**). These findings are in agreement with the key requirement of structural motif conformations **I1** (equivalent to **BC-Inh1**, **BC-Inh2**, or **TC**), **I2** (equivalent to **BC-SAM**), and any two of **P1**–**4** except **P2**–**P3** (*e.g.* **P1**–**P3** is equivalent to the combination of **APO** with **BC-SAM** or **TC**) to achieve >90% microstate coverage. For SAM-bound SETD8, the seed conformations derived from **BC-SAM** and **TC** structures contribute 31 and 38 of 67 microstates (**Figure 4c,d, Table S14**). These findings argue for the importance of using multiple structures to construct the

landscape within achievable computer time. The seed conformations prepared from ligand-trapped SETD8 structures are essential to discovering the complete conformational landscapes of SETD8.

For simulation time, we observed that the fewer seed conformations of apo-SETD8 were employed, the more computing power (the product between the number of simulation trajectories and the time length per trajectory) was required to reach a comparable level of microstate coverage (**Figure 4e**, **Tables S16**, **S17**). When computing power is fixed, comparable microstate coverages of apo- and SAM-bound SETD8 can be obtained by running either multiple short trajectories or few long trajectories (**Figures 4f**, **S38**). The current simulation time (5 ms for apo-SETD8 and 1 ms for SAM-bound SETD8) provides 2–10-fold redundant computing power to map the conformational landscapes of SETD8.

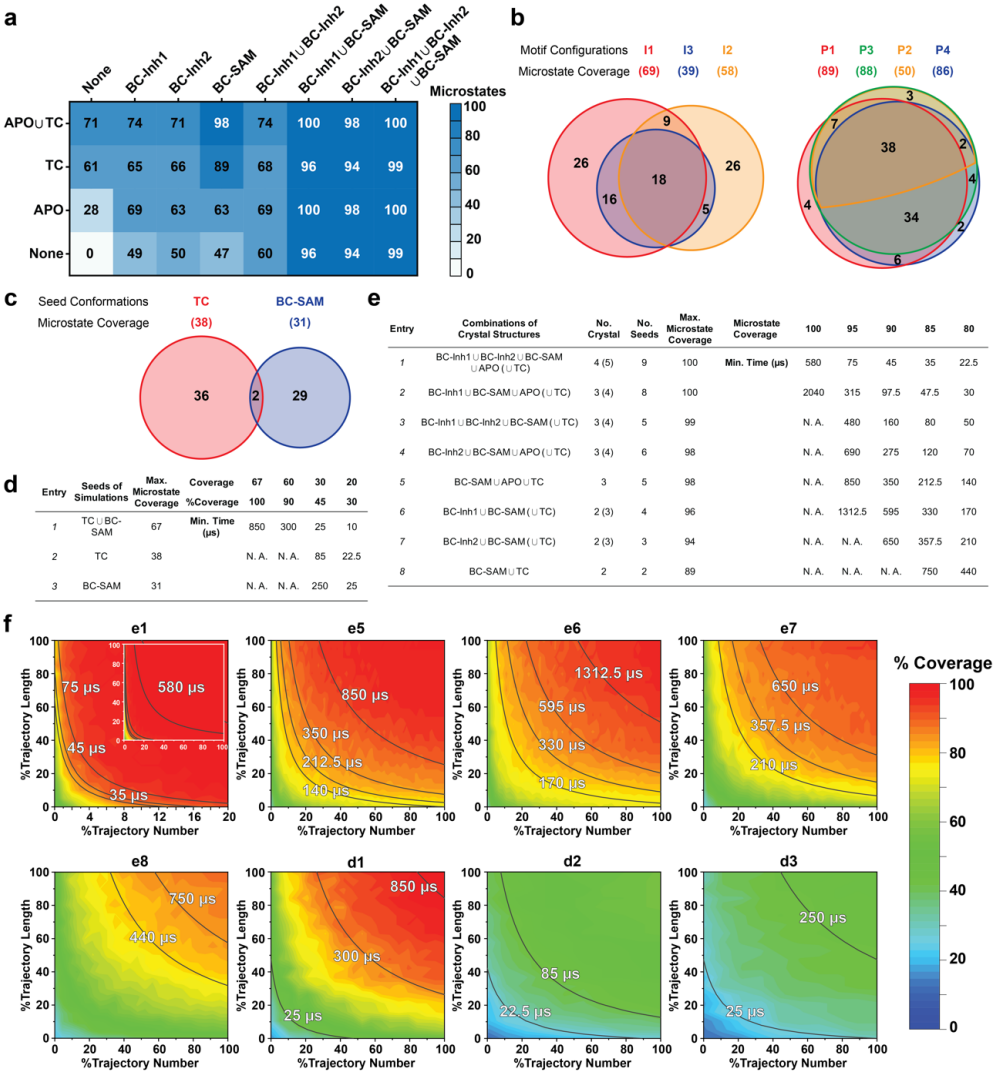


Figure 4. Evaluation of key simulation parameters of massively distributed molecular simulations. **a–b**, Robustness of simulations of apo-SETD8: **a**, Heat map for the coverage of the 100 microstates with all combinations of the crystal structures (**BC-Inh1, BC-Inh2, BC-SAM, APO, and TC**) as seed conformations; **b**, Venn diagrams of the coverage of the 100 microstates with all conformational combinations of SET-I and post-SET motifs (**I1-3** and **P1-4**) as seed structures for MD simulations. **c–d**, Robustness of simulations of SAM-bound SETD8: **c**, Venn diagram of the coverage of the 67 microstates with **TC**, **BC-SAM** or both as seed structures for MD simulation; **d**, Minimal time required by MD simulations to reach certain coverage of the 67 microstates with representative combinations of seed structures. **e**, Minimal time required by MD simulations to reach certain coverage of the 100 microstates

of apo-SETD8 with representative combinations of seed structures. **f**, Contour map of microstate coverage at various combinations of trajectory lengths and numbers as percentage of the maximal trajectory length and number of MD simulations. The seed structures of each panel are listed as the simulation entries e1, e5−8 for apo-SETD8, and d1−3 for SAM-bound SETD8. Each curve corresponds to the aggregation of specific simulation time.

Functionally relevant conformations in the dynamic landscapes of apo- and SAM-bound SETD8. After validating the conformational landscapes of apo- and SAM-bound SETD8, we explored the dynamic details of these landscapes with the focus on the connectivity and equilibrium fluxes between kinetically metastable macrostates (henceforth referred to as the "network"). When projected into two dimensions, the conformational landscape of apo-SETD8 takes the form of a dumbbell-like shape containing two lobes, each composed of about 12 macrostates primarily connected via a single hub-like central macrostate A11 (**Figures 2d**, **5**, **Table S7**). The conformational landscape also consists of other multiply-connected macrostates, including A1−A4, A9, and A14, as characterized by their rapid kinetic interconversion with multiple other macrostates (**Figure 2d**,**e**). Most low-populated macrostates (A17−A24) appear as satellite macrostates in the periphery of the network with few high-flux channels of interconversion to other macrostates (**Figure 2d**,**e**). The remaining states were classified as basin-like macrostates including (A5, A10), A7, A8, (A12, A13, A16) and A15, because these macrostates are highly populated and either relatively isolated or appear in tightly interconnected but globally isolated groups.

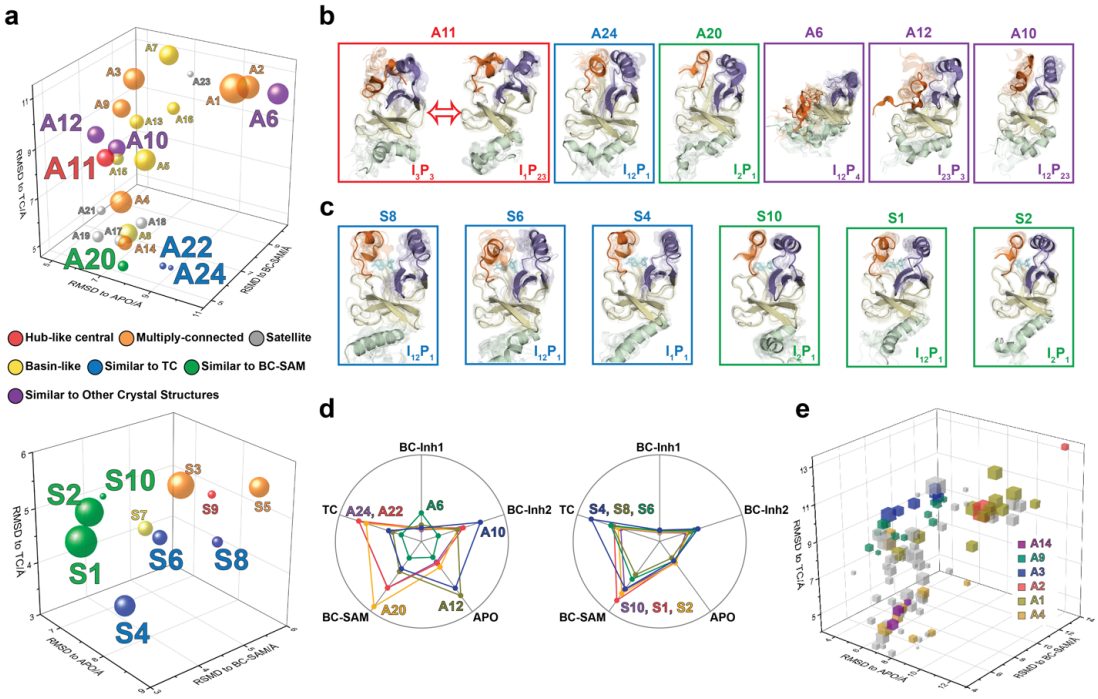


Figure 5. Functional annotation of the dynamic conformational landscapes of SETD8. a, 3D scatterplots of the 24 macrostates of apo-SETD8 landscape and 10 macrostates of SAM-bound SETD8 landscape in the coordinates of RMSDs relative to **APO**, **BC-SAM**, and **TC**. Volume of each sphere is proportional to the relative population of the corresponding macrostate in the context of the 24 macrostates for apo-SETD8 or the 10 macrostates for SAM-bound SETD8. The RMSD of each macrostate is the average of its microstates weighted with their intra-macrostate population. The RMSD of each microstate is the average of the top 10 frames most closely related to the clustering center of the microstate. The feature of each macrostate is annotated in color. **b**, **c** Cartoons of representative conformations of key macrostates in the apo-SETD8 landscape and the SAM-bound SETD8 landscape, respectively. Structural annotations are shown in bottom right of each conformation. **d**, Radar chart of representative macrostates of apo (left) and SAM-bound (right) landscapes in reference to the five crystal structures. Distances between dots and cycle centers are proportional to the reciprocal values of RMSDs of macrostates relative to the crystal structures. **e**, 3D scattering plot of 100 microstates of the apo landscape in the coordinates of RMSDs to **APO**, **BC-SAM**, and **TC**. Volume of each cube is proportional to the relative population of the corresponding microstate in the context of the 100 microstates. Microstates clustered in intermediate-like macrostates are highlighted in colors. Structural diversity of microstates within individual macrostates indicates that each intermediate-like state contains multiple structurally distinct but readily interconvertible microstates.

The hub-like macrostate A11 consists of two structurally-distinct microstates with comparable populations (**Figures 2d, 5a**). One microstate structurally resembles the conformation of **APO** (I_3P_3), while the other microstate represents a conformer with the I_1P_{23} feature for its SET-I and post-SET motifs (**Figure 5b**, **Table S6**). Rapid conformational interconversions within A11 is consistent with its hub-like character, centered between the two lobes of the dumbbell-like network. Interestingly, macrostates kinetically adjacent to A11 have structurally similar SET-I motifs within each lobe but distinct SET-I motifs between the two lobes (**I2~3** for the left and **I1~2** for the right) (**Figures 2d**, **5b**). Therefore, A11 is a transition-type state essential for the conformational fluxes of the macrostates between the two lobes, involved in a key step of conformational changes of the SET-I motif between **I1~2** and **I2~3**.

The intermediate-like macrostates A1–A4, A9, and A14 each contains multiple structurally distinct but kinetically associated microstates (**Figures 2d, 5a**,**b**). The satellite macrostates A17–A24 are less populated and more structurally homogeneous (**Figures 2d, 5a**,**b**). Conformers in the macrostates A22, A24 and A20 are structurally similar to **TC** and **BC-SAM** with slightly different but well-defined SAM-binding pockets, suggesting minimal conformational reorganization of A22, A24, and A20 is required to accommodate the cofactor (**Figure 5a**,**b**,**e**). Interestingly, A22 and A24, whose overall structures are similar to each other (**TC**-like), rarely interconvert in the apo landscape (**Figure 2d**). In contrast, the basin-like macrostates (A5, A10), A7, A8, (A12, A13, A16) and A15 do not contain a well-defined SAM-binding pocket (**Figure 5a**,**b**,**e**). Here the conformers in macrostate A12 are similar to **APO**, the conformers in the macrostate A6 are similar to **BC-Inh1**, and the conformers in the macrostates A10 are similar to **BC-Inh2** (**Figure 5d**). The structural similarity between the simulated conformers and **BC-Inh1/2** strongly argue that the two covalent inhibitors successfully trapped key hidden conformers of apo-SETD8.

Similar to that of apo-SETD8, the interconversion network of the macrostates of SAM-bound SETD8 also displays a dumbbell-like shape with S9 as the hub-like state connecting the two lobes of the network (**Figures 2d, 5a**). The macrostates S1 and S3–S5 are multi-connected states; S6, S8, and S10 are satellite-like states; S2 and S7 are basin-like states (**Figure 5a**,**b**).

Notably, the complexity of the overall conformational landscape of SAM-bound SETD8 is dramatically reduced in comparison with those of apo-SETD8 (**Figures 2d**, **5a**). The conformers in S1, S2, and S10 are structurally similar to those of A20, as well as **BC-SAM**; the conformers in S4, S6, and S8 are structurally similar to those in A22 and A24, as well as **TC** (**Figure 5c**,**d**). The structural similarities between these apo and SAM-bound macrostates suggest possible pathways for connecting the two conformational landscapes upon SAM binding.

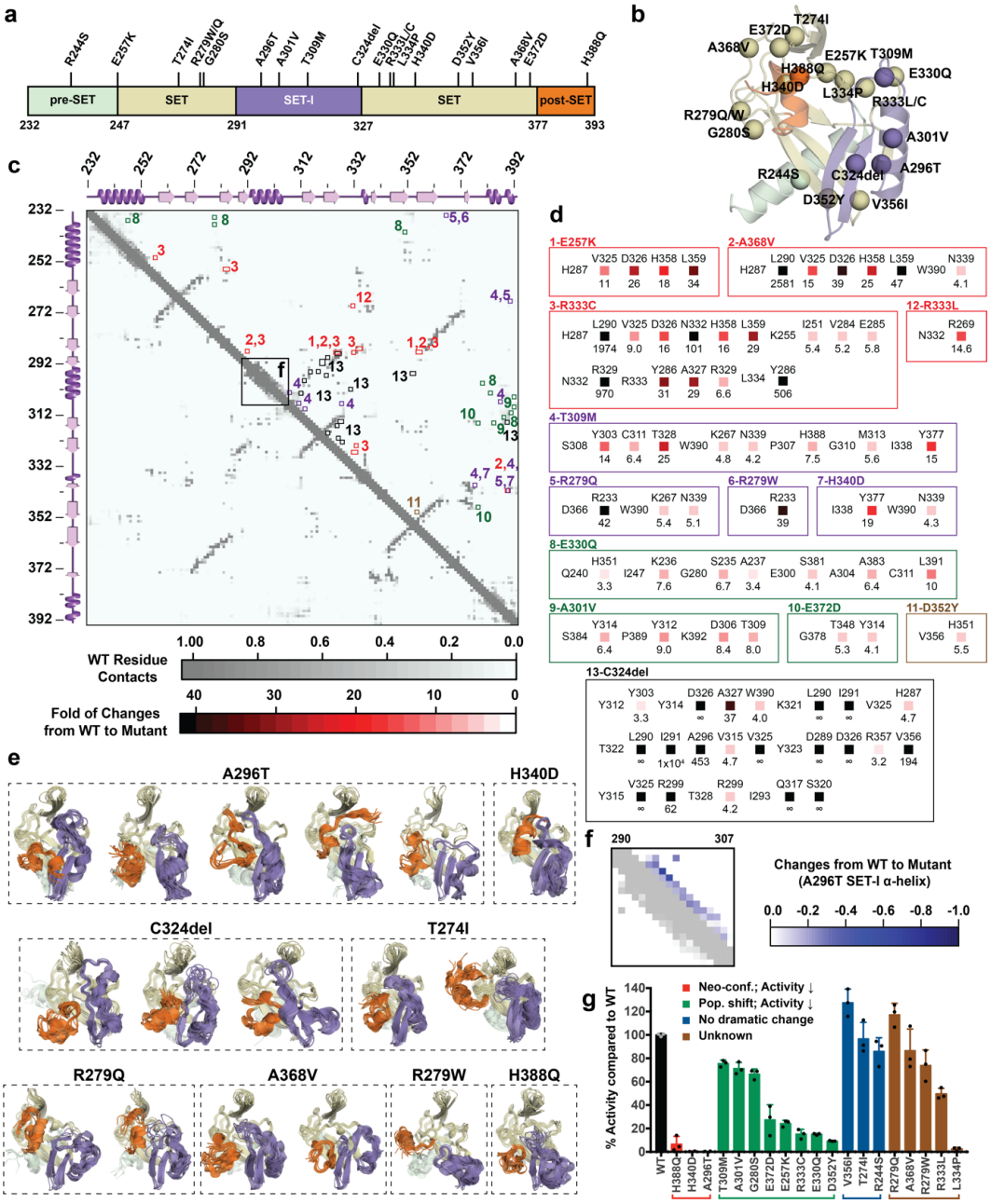


Figure 6. Computational and experimental characterization of cancer-associated SETD8 mutants. a, Cancer-associated mutations in the catalytic domain of SETD8 examined in this work. **b**, Cartoon representations of **TC** with cancer-associated SETD8 mutations highlighted. **c**, Differential residue-contact maps of cancer-associated SETD8 mutants in reference to wild type apo-SETD8 (grey). **d**,

Representative contacts in the differential residue-contact maps of cancer-associated SETD8 mutants. The contacts of SETD8 mutants with >3-fold gain of contact fraction relative to wild-type SETD8 are listed and color-coded according to the increased magnitude of the contact fraction. **e**, Cartoon representations of neo-conformations revealed by simulations of SETD8 mutants. **f**, Differential residue-contact maps of the structurally relaxed α-helix at the SET-I motif of SETD8 A296T mutant. Decrease of contact fraction of SETD8 mutants relative to wild-type SETD8 is colored in blue. **g**, Enzymatic activities of wild-type and mutated SETD8 determined by an *in vitro* radiometric assay with H4K20 peptide substrate. Here SETD8 mutants are categorized as the following: red, uncovered neo-conformations (Neo-conf.) with > 90% loss of methyltransferase activity; green, populated inactive conformations (Pop. shift) with partially abolished methyltransferase activity; blue, no dramatic change of differential contact maps with comparable methyltransferase activity with wild-type SETD8; brown, unknown relationship between differential contact maps and methyltransferase activities. Data are mean ± standard deviation (s.d.) of 3 replicates.

Characterization of cancer-associated SETD8 mutants. Sequences from tumor samples retrieved from cBioPortal[42-44] contain two dozen point mutations in the catalytic domain of SETD8 (**Figure 6a**,**b**, **Table S11**). We expect that some of these mutations perturb SETD8 function. Because of conformational heterogeneity, it has historically been challenging for *in silico* approaches to annotate how mutations---in particular those structurally remote from functional sites---affect a target protein on the basis of its static structure(s)[45-47]. Here, we envisioned addressing this challenge with the aid of the dynamic conformational landscapes of SETD8. To characterize mutations remote from catalytic sites (around 80% of known mutations), 40 independent microsecond-long MD simulations for each of the cancer-associated apo-SETD8 mutants were conducted with seed structures prepared from one ternary complex (**TC**) conformer. We then constructed a differential residue-contact map for each variant (**Figure 6c**,**d**) and extracted snapshots representing most dramatic conformational deviations from the wild type conformational ensembles (**Figure 6e**). Remarkably, even with modest simulation time, several cancer-associated mutants displayed neo-conformations that were not observed in the 5 ms wild-type dataset and cannot be predicted from static X-ray crystal structures. Strikingly, all of the neo-conformations display distinct reorganizations at the SET-I motif (**Figure 6e**). For instance, a single point mutation A296T, ~16 Å remote from the active site, yields five distinct neo-conformations (**Figure 6e**). In addition, relative to wild-type apo-SETD8, this mutant

populates several conformations with a structurally relaxed α-helix at the SET-I motif (**Figure 6e**). C324del, ~20 Å from the SET-I motif, is associated with three neo-conformations and displays the most dramatic changes in the differential contact map (**Figure 6d,** panel 13). The remote H340D mutation is associated with one neo-conformation as well as more populated conformations containing spatially compressed active sites (**Figure 6d,** panel 7; **6e**). Using *in vitro* radiometric assays, the A296T and H340D mutants were characterized by loss of the methyltransferase activity on H4K20 peptide substrate (**Figure 6g**). The failure to purify recombinant C324del also supports the impact of this deletion on SETD8 function. H388Q, which mutates a histidine involved in substrate binding, is also associated with neo-conformations as well as loss of the methyltransferase activity (**Figure 6e,g**). These observations provide potential molecular rationale for how remote mutations can alter the active sites and the SET-I motif---and hence catalysis---via modulating the conformational landscape. Exceptions are T274I, R279W, R279Q, and A368V, which yielded neo-conformations but showed activity comparable to wild-type SETD8 (**Figure 6e,g**).

The differential residue-contact maps further revealed that remote mutations can alter conformational landscapes by altering populations of pre-existing conformations (**Figure 6c,d**). For instance, E257K, G280S, A301V, T309M, E330Q, D352Y mutations populate conformations containing spatially compressed active sites (**Figure S37**); E372D populates conformations containing a constrained post-SET motif; R333C populates conformations with reorganized SET motifs adjacent to the peptide binding pocket. All of these mutations showed partial loss of methyltransferase activity (**Figure 6g**). Notably, these structural alterations are often remote from the corresponding mutation sites (**Figure 6b**). In contrast, R244S, T274I, and V356I showed no significant conformational change on the basis of their differential contact maps, consistent with their comparable methyltransferase activity to wild type SETD8 (**Figure 6g**). Likely due to insufficient simulation time (40×1 μs/mutant), R333L and L334P variants, characterized by partial-to-complete loss of the methyltransferase activity (**Figure 6g**), showed similar conformational landscapes to that of wild-type apo-SETD8. Exploring conformational landscapes is thus an effective strategy to reveal structural alterations associated with majority of remote-site mutations for functional annotation.

Discussion

Here we have demonstrated that tight integration of structural determination---using covalent probes and multiple ligand-binding states to trap hidden conformations (**Figure 1**)---with massively distributed molecular simulations and the powerful framework of Markov state models (**Figure 2b**) can provide unprecedented insights into the detailed conformational dynamics of an enzyme. The current work demonstrates the merit of an approach that leverages multiple X-ray structures with distinct diverse conformations for MD simulations and machine-learning-based MSM construction to elucidate complex conformational dynamics, and validates the resulting model experimentally with testable biophysical predictions (**Figure 3**). Previously, individual components of our integrative strategy have been employed to study the dynamics of transcriptional activators[48], kinases[11,12], and allosteric regulation[13]. However, it is the first time that these diverse approaches are consolidated explicitly with the goal of illuminating conformational dynamics of an enzyme in a comprehensive and feasible manner. Assessment of key computational parameters concluded that we have utilized sufficient diverse seed structures and simulation time for microstate discovery and thus robust construction of conformational landscapes (**Figure 4**). Notably, we relied on a unique computational resource---Folding@home---to collect remarkable six-millisecond simulation data (see Method). Without access to Folding@home, contemporaneous progress on developing *adaptive* Markov state model construction algorithms---where iterative model building guides the collection of additional simulation data[49,50]---will still allow research groups to achieve this feat on local GPU clusters or cloud resources in the near future. Furthermore, the concept of adaptive model construction can be extended to identify which new structural or biophysical data would be valuable in reducing uncertainty[51-53] and producing refined MSMs. The integrated platform and concept formulated via this work can be readily transformed to explore dynamic conformational landscapes of other proteins.

This work represents the first time that conformational dynamics of a protein methyltransferase has been definitively characterized with atomic details. Strikingly, SETD8 adopts extremely diverse dynamic conformations in apo and SAM-bound states (24 and 10 kinetically metastable macrostates, respectively, **Figure 2**). Interconversions between metastable conformers cover a broad spatio-temporal scale in particular associated with motions of

SETD8's SET-I and post-SET motifs (**Figures 1,5**). In the apo landscape, the general structural features of the X-ray structures of **BC-Inh1**, **BC-Inh2**, **APO**, **BC-SAM** and **TC** (**Figure 1**) are recapitulated by a subset of macrostates (*e.g.* A6 for **BC-Inh1**; A10 for **BC-Inh2**; A12 for **APO**; A20 for **BC-SAM**; A22, A24 for **TC**, 6 of 24 macrostates, **Figure 5**). Such observation indicates that these X-tray structures trapped in the different ligand-binding states are not ligand-induced artifacts but indeed relevant snapshots of hidden conformations of apo-SETD8. Similarly, a few macrostates in the SAM-bound landscape also recapitulate major structural features of the two cofactor-bound X-ray structures (*e.g.* S1, S2, S10 for BC-SAM, S4, S6, S8 for TC, 6 of 10 macrostates, **Figure 5**). Meanwhile, our results also demonstrate that X-ray crystallography alone is insufficient to capture all metastable conformations of SETD8. Remarkably, there is no correlation of overall structural similarity and interconversion rates between metastable conformers. Though the anticipated findings of fast transitions between structurally similar conformers and slow transitions between structurally distinct conformers (*e.g.* microstates within individual satellite macrostates A17–A24 of apo SETD8; S6, S8, and S10 of SAM-bound SETD8, **Figure 5**), we frequently observed fast kinetics of transitions between structurally distinct microstates (*e.g.* microstates within hub-like macrostates A11 and S8; multi-connected states A1–A4, A9, A14, S1 and S3–S5) and *vice versa* (*e.g.* macrostates A22 and A24) (**Figures 2**,**5**). It is thus interesting to examine how other factors such as specific residue contacts and cooperative long range motions of certain structural motifs play roles on interconversion kinetics.

Functional annotation of the landscapes revealed that the SET-I motif adopts diverse conformations (Figure 2), and its overall configuration is a key feature that differentiates the lobes of the dumbbell-like conformational landscape of SETD8. The conformational dynamics within the hub-like macrostate A11 mainly involves motions of the SET-I motif. Two gain-of-function I293G and E292G variants of SETD8 were designed for relaxing distorted configurations of the SET-I motif upon SAM binding (**Figure 3**). These findings argue the functional essentiality of the intrinsically dynamic motions of SET-I motif for SETD8 SAM binding and catalysis. Importance of dynamic conformational modulation of the SET-I motif has also been shown for other SET-domain PKMTs. For instance, the SET domains of MLLs and EZH1/2 alone are catalytically inert but active in the presence of binding partners WDR5-RbBP5-Ash2L-Dpy30 (referred as MLL-WRAD) and EED-Suz12 (referred as PRC2),

respectively[15]. Recent structural evidence implicated that the formation of these complexes regulates the conformational dynamics of the SET-I motif, which is essential for catalysis[54,55]. Interestingly, this region has also been exploited by cancer-associated mutants of PKMTs. For instance, NSD2's E1099 is located in its SET-I motif and its E1099K mutant was characterized as a hot-spot cancer mutation with the gain-of-activity of H3K36 methylation[56]. Additionally, many mutations of PKMTs have been mapped in their SET-I motifs, implicating their potential roles in alternation of function (**Figure S39, Table S12**).

In contrast to static X-ray structures, dynamic conformational landscapes greatly facilitated the characterization of cancer-associated SETD8 mutants (**Figure 6**). A significant portion of cancer-associated, loss-of-function SETD8 mutations, though remote from active sites, were revealed to act allosterically through perturbing the SET-I motif and thus catalysis (**Figure 6**). We also discovered significant changes in the connective networks and a dramatic decrease in conformational heterogeneity upon SAM binding (**Figure 2**). This finding highlights how enzyme-ligand interactions reshape conformation landscapes. The conformational landscapes of SETD8 thus provide an unprecedented platform for virtual screening of ligand candidates as inhibitors via exploring different modes of interaction (SAM-competitive, substrate-competitive, covalent or allosteric). Uncovering hidden conformations can thus be essential for developing potent and selective SETD8 inhibitors. The conformations of individual SETD8 microstates can be further explored to derive their thermodynamic, kinetic, and even transition-state parameters in a catalytic cycle. Similar strategies can be generally applied to native or disease-associated PKMTs for functional annotation.

References

1 Wei, G. *et al.* Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **116**, 6516-6551 (2016).
2 Ferguson, F. M. *et al.* Kinase inhibitors: the road ahead. *Nat. Rev. Drug. Discov.* **17**, 353-377 (2018).
3 Latorraca, N. R. *et al.* GPCR Dynamics: Structures in Motion. *Chem. Rev.* **117**, 139-155 (2017).
4 Lu, S. *et al.* Ras Conformational Ensembles, Allostery, and Signaling. *Chem. Rev.* **116**, 6607-6665 (2016).
5 Shi, Y. A glimpse of structural biology through X-ray crystallography. *Cell* **159**, 995-1014 (2014).
6 Huang, C. *et al.* Structures of Large Protein Complexes Determined by Nuclear Magnetic Resonance Spectroscopy. *Annu. Rev. Biophys.* **46**, 317-336 (2017).

7 Fernandez-Leiro, R. *et al.* Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **537**, 339-346 (2016).

8 van den Bedem, H. *et al.* Integrative, dynamic structural biology at atomic resolution-it's about time. *Nat. Methods* **12**, 307-318 (2015).

9 Husic, B. E. *et al.* Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **140**, 2386-2396 (2018).

10 Plattner, N. *et al.* Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **9**, 1005-1011 (2017).

11 Shukla, D. *et al.* Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nat. Commun.* **5**, 3397 (2014).

12 Sultan, M. M. *et al.* Millisecond dynamics of BTK reveal kinome-wide conformational plasticity within the apo kinase domain. *Sci. Rep.* **7**, 15604 (2017).

13 Bowman, G. R. *et al.* Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. *Proc. Natl. Acad. Sci. U S A* **112**, 2734-2739 (2015).

14 Klepeis, J. L. *et al.* Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* **19**, 120-127 (2009).

15 Luo, M. Chemical and Biochemical Perspectives of Protein Lysine Methylation. *Chem. Rev.* (2018).

16 Allis, C. D. *et al.* The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17**, 487-500 (2016).

17 Murn, J. *et al.* The winding path of protein methylation research: milestones and new frontiers. *Nat. Rev. Mol. Cell. Biol.* **18**, 517-527 (2017).

18 Dawson, M. A. The cancer epigenome: Concepts, challenges, and therapeutic opportunities. *Science* **355**, 1147-1152 (2017).

19 Flavahan, W. A. *et al.* Epigenetic plasticity and the hallmarks of cancer. *Science* **357** (2017).

20 Fang, J. *et al.* Purification and Functional Characterization of SET8, a Nucleosomal Histone H4-Lysine 20-Specific Methyltransferase. *Curr. Biol.* **12**, 1086-1099 (2002).

21 Nishioka, K. *et al.* PR-Set7 Is a Nucleosome-Specific Methyltransferase that Modifies Lysine 20 of Histone H4 and Is Associated with Silent Chromatin. *Mol. Cell* **9**, 1201-1213 (2002).

22 Shi, X. B. *et al.* Modulation of p53 function by SET8-mediated methylation at lysine 382. *Mol. Cell* **27**, 636-646 (2007).

23 Dhami, G. K. *et al.* Dynamic methylation of Numb by Set8 regulates its binding to p53 and apoptosis. *Mol. Cell* **50**, 565-576 (2013).

24 Liu, W. *et al.* PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. *Nature* **466**, 508-512 (2010).

25 Beck, D. B. *et al.* PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. *Genes Dev.* **26**, 325-337 (2012).

26 Veschi, V. *et al.* Epigenetic siRNA and Chemical Screens Identify SETD8 Inhibition as a Therapeutic Strategy for p53 Activation in High-Risk Neuroblastoma. *Cancer Cell* **31**, 50-63 (2017).

27 Yang, F. *et al.* SET8 promotes epithelial-mesenchymal transition and confers TWIST dual transcriptional activities. *EMBO J.* **31**, 110-123 (2012).

28 Hashemi, M. *et al.* Association of functional polymorphism at the miR-502-binding site in the 3' untranslated region of the SETD8 gene with risk of childhood acute lymphoblastic leukemia, a preliminary report. *Tumour Biol.* **35**, 10375-10379 (2014).

29 Milite, C. *et al.* Progress in the Development of Lysine Methyltransferase SETD8 Inhibitors. *ChemMedChem* **11**, 1680-1685 (2016).

30 Milite, C. *et al.* The emerging role of lysine methyltransferase SETD8 in human diseases. *Clin. Epigenetics* **8**, 102 (2016).

31 Blum, G. *et al.* Small-molecule inhibitors of SETD8 with cellular activity. *ACS Chem. Biol.* **9**, 2471-2478 (2014).

32 Schramm, V. L. Enzymatic transition states, transition-state analogs, dynamics, thermodynamics, and lifetimes. *Annu. Rev. Biochem.* **80**, 703-732 (2011).

33 Butler, K. V. *et al.* Structure-Based Design of a Covalent Inhibitor of the SET Domain-Containing Protein 8 (SETD8) Lysine Methyltransferase. *J. Med. Chem.* **59**, 9881-9889 (2016).

34 Luo, M. *et al.* Naphthaquinone methyltransferase inhibitors and uses thereof. (2015). (Sloan-Kettering Institute for Cancer Research, USA), Int. PCT Pub. No. WO2015172076 A1, 2015.

35 Couture, J. F. *et al.* Structural and functional analysis of SET8, a histone H4 Lys-20 methyltransferase. *Genes Dev.* **19**, 1455-1465 (2005).

36 Xiao, B. *et al.* Specificity and mechanism of the histone methyltransferase Pr-Set7. *Genes Dev.* **19**, 1444-1454 (2005).

37 Couture, J. F. *et al.* Structural origins for the product specificity of SET domain protein methyltransferases. *Proc. Natl. Acad. Sci. U S A* **105**, 20659-20664 (2008).

38 Husic, B. E. *et al.* Optimized parameter selection reveals trends in Markov state models for protein folding. *J. Chem. Phys.* **145**, 194103 (2016).

39 Royer, C. A. Probing protein folding and conformational transitions with fluorescence. *Chem. Rev.* **106**, 1769-1784 (2006).

40 Johnson, K. A. 1 Transient-State Kinetic Analysis of Enzyme Reaction Pathways. *The Enzymes* **20**, 1-61 (1992).

41 Agafonov, R. V. *et al.* Energetic dissection of Gleevec's selectivity toward human tyrosine kinases. *Nat. Struct. Mol. Biol.* **21**, 848-853 (2014).

42 Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401-404 (2012).

43 Gao, J. J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling* **6** (2013).

44 Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J. Mol. Diagn.* **17**, 251-264 (2015).

45 Stefl, S. *et al.* Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.* **425**, 3919-3936 (2013).

46 Klinman, J. P. *et al.* Evolutionary aspects of enzyme dynamics. *J. Biol. Chem.* **289**, 30205-30212 (2014).

47 Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function. *Nat. Chem. Biol.* **12**, 944-950 (2016).

48 Wang, N. *et al.* Ordering a dynamic protein via a small-molecule stabilizer. *J. Am. Chem. Soc.* **135**, 3363-3366 (2013).

49 Bowman, G. R. *et al.* Enhanced modeling via network theory: Adaptive sampling of Markov state models. *J. Chem. Theory. Comput.* **6**, 787-794 (2010).

50 Shamsi, Z. *et al.* Enhanced unbiased sampling of protein dynamics using evolutionary coupling information. *Sci. Rep.* **7**, 12700 (2017).

51 Olsson, S. *et al.* Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc. Natl. Acad. Sci. U S A* **114**, 8265-8270 (2017).

52 Dixit, P. D. *et al.* Caliber Corrected Markov Modeling (C2M2): Correcting Equilibrium Markov Models. *J. Chem. Theory. Comput.* **14**, 1111-1119 (2018).

53 Matsunaga, Y. *et al.* Linking time-series of single-molecule experiments with molecular dynamics simulations by machine learning. *Elife* **7** (2018).

54 Li, Y. *et al.* Structural basis for activity regulation of MLL family methyltransferases. *Nature* **530**, 447-452 (2016).

55 Justin, N. *et al.* Structural basis of oncogenic histone H3K27M inhibition of human polycomb repressive complex 2. *Nat. Commun.* **7**, 11316 (2016).

56 Oyer, J. A. *et al.* Point mutation E1099K in MMSET/NSD2 enhances its methyltranferase activity and leads to altered global chromatin methylation in lymphoid malignancies. *Leukemia* **28**, 198-201 (2014).

Acknowledgements

The authors thank for the National Institutes of Health of USA (ML: R01GM096056, R01GM120570; JDC: R01GM121505; JJ: R01GM122749, R01HD088626; YGZ, R01GM126154), National Cancer Institute (ML, JDC: 5P30 CA008748; JJ: R01CA218600), MSKCC Functional Genomics Initiative (ML), the Sloan Kettering Institute (ML, JDC, KAB), Mr. William H. Goodwin and Mrs. Alice Goodwin Commonwealth Foundation for Cancer Research, and the Experimental Therapeutics Center of Memorial Sloan Kettering Cancer Center (ML), and Louis V. Gerstner Young Investigator Award (JDC), K. C. Wong Education Foundation (CL), Chinese Academy of Sciences (CL: XDA12020353), National Natural Science Foundation of China (CL: 81625022 and 81430084), the Tri-Institutional PhD Program in Chemical Biology (RPW and SC), Peer Reviewed Cancer Research Program of the Department of Defense (RPW: W81XWH-17-1-0412) for research supports; the Marie-Josée and Henry R. Kravis Center for Molecular Oncology, and the Molecular Diagnostics Service in the Department of Pathology for the access of tumor mutation data via cBioPortal; Carolina Adura at High Throughput and Spectroscopy Resource Center at The Rockefeller University for the assistance of ITC experiments; Henry Zebroski III and Susan Powell at Proteomics Resource Center at The Rockefeller University for peptide synthesis; the Folding@home project for computational resources; Kanishk Kapilashrami, Josh Fass, Sonya Hanson, Frank Noé, Simon Olsson, and Martin Scherer for insightful discussions or software support. The Structural Genomics Consortium is a registered charity (no. 1097737) that receives funds from AbbVie; Bayer Pharma AG; Boehringer Ingelheim; Canada Foundation for Innovation; Eshelman Institute for Innovation; Genome Canada; Innovative Medicines Initiative (EU/EFPIA) (ULTRA-DD grant no. 115766); Janssen; Merck & Co.; Novartis Pharma AG; Ontario Ministry of Economic Development and Innovation; Pfizer; São Paulo Research Foundation-FAPESP; Takeda; and the Wellcome Trust. The X-ray structure results of **BC-Inh2** and **BC-SAM** are derived from work performed at Argonne National Laboratory, Structural Biology Center (SBC) at the Advanced Photon Source. SBC-CAT is operated by UChicago Argonne, LLC, for the U.S.

Department of Energy, Office of Biological and Environmental Research under contract DE-AC02-06CH11357. These experiments were performed using beamline 08ID-1 at the Canadian Light Source, which is supported by the Canada Foundation for Innovation, Natural Sciences and Engineering Research Council of Canada, the University of Saskatchewan, the Government of Saskatchewan, Western Economic Diversification Canada, the National Research Council Canada, and the Canadian Institutes of Health Research. The X-ray experiment of **BC-Inh1** was conducted with NE-CAT beam line 24-ID-E (GM103403) and an Eiger detector (OD021527) at the APS (DE-AC02-06CH11357).

Author Contributions

M.L., J.D.C, C.L., S.C., and F.M. initialized this project. M.L., J.D.C, and C.L. designed experiments and directed the project. N.B., A.M., J.W., G.B., F.P.-S., J.J., and M.L. developed inhibitors. N.B., A.M., Y.Y., W.T., H.Z., R.S., P.J.B., and J.J. solved X-ray structures. F.M. S.F., H.J., K.C. and C.L. performed initial simulation and analysis. R.P.W. and K.A.B. set up MD simulations. R.P.W. and S.C. conducted computational analysis. S.C., K.Q., H.H., J.W., Y.G.Z., and M.L. designed and conducted biochemical experiments. M.L., J.D.C., S.C., and R.P.W. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

PDB files: 6BOZ for **BC-Inh1**, 5W1Y for **BC-Inh2**, 4IJ8 for **BC-SAM**, and 5V2N for **APO**.

Code and Data Availability. The molecular dynamics datasets generated and analyzed in this study are available via the Open Science Framework at https://osf.io/2h6p4. The code used for the generation and analysis of the molecular dynamics data is available via a Github repository at https://github.com/choderalab/SETD8-materials.