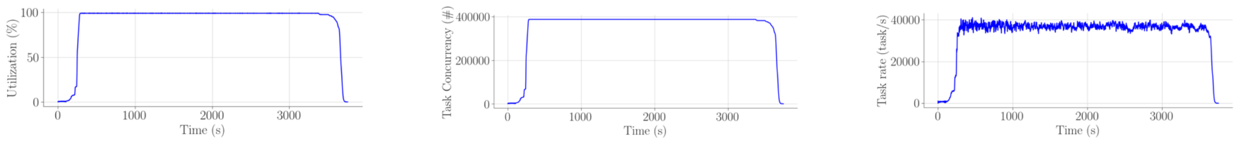
Supporting Information for: High Throughput Virtual Screening and Validation of a SARS-CoV-2 Main Protease Non-Covalent Inhibitor

Austin Clyde[1,2,14*], Stephanie Galanie[3,14], Daniel W. Kneller[4,14], Heng Ma[1,14], Yadu Babuji[2,14], Ben Blaiszik[1,14], Alexander Brace[1,2,14], Thomas Brettin[5,14], Kyle Chard[2,14], Ryan Chard[1,2,14], Leighton Coates[4,14], Ian Foster[1,2,14], Darin Hauner[6,14], Vilmos Kertesz[3,14], Neeraj Kumar[6,14], Hyungro Lee[7,14], Zhuozhao Li[1,2,14], Andre Merzky[7,14], Jurgen G. Schmidt[8,14], Li Tan[7,14], Mikhail Titov[7,14], Anda Trifan[9,14], Matteo Turilli[7,10,14], Hubertus Van Dam[10,14], Srinivas C. Chennubhotla[11,14], Shantenu Jha[7,10,14*], Andrey Kovalevsky[12,14*], Arvind Ramanathan[1,14*], Martha S. Head[13,14*], Rick Stevens[2,5,14*] *

[1]*Data Science and Learning Division,* [5]*Computing Environment and Life Sciences Directorate, Argonne National Laboratory, Lemont, IL 60439, USA.* [2] *Department of Computer Science, University of Chicago, Chicago, IL 60615, USA.* [3]*Biosciences Division,* [4]*Neutron Scattering Division,* [12]*Second Target Station,* [13]*Joint Institute for Biological Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA.* [6]*Computational Biology Group, Biological Science Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA.* [7]*Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ 08854, USA.* [8]*Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.* [9]*University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.* [10]*Computational Science Initiative, Brookhaven National Laboratory, Upton, NY 11973, USA.* [11]*Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, 15260 USA* [14]*National Virtual Biotechnology Laboratory.*

E-mail: aclyde@anl.gov;shantenu.jha@rutgers.edu;kovalevskyay@ornl.gov; ramanathana@anl.gov;headms@ornl.gov;stevens@anl.gov

High Throughput Virtual Screening Workflow



(a) (b) (c)

Figure S1: (a) resource utilization (RU), (b) execution concurrency (EC) and (c) task execution rate (TR) with RAPTOR when executing 126,471,524 OpenEye Python function calls on 7000 compute nodes/392,000 cores of Frontera with 70 master and 99 workers per master. RU = 90$\%$; EC = 4×10^5 steady state; TR = $144 \times 10^6/hour$ peak.

Table S1: Correlation matrix for docking scores across different receptors used.

	7C7P	7JU7	7BQY	6W63	6LU7	Consensus
7C7P	1.000	0.041	0.395	0.269	0.400	0.647
7JU7	0.041	1.000	0.260	0.188	0.268	0.527
7BQY	0.395	0.260	1.000	0.402	0.843	0.799
6W63	0.269	0.188	0.402	1.000	0.408	0.666
6LU7	0.400	0.268	0.843	0.408	1.000	0.806
Consensus	0.647	0.527	0.799	0.666	0.806	1.000

To minimize the overhead caused by repeated loading of the receptor data from memory, the data were loaded once per node and then reused for all docking runs assigned to that specific node. The individual cores hosting the docking computations received cloned copies of the receptor data so as to isolate the individual docking computations. To reduce the overhead of loading compound data from disk, the storage offsets in the dataset were precomputed at startup, and shared with all the nodes, which reduced the I/O operations on the compound data. Intermediate data were stored on the local storage of each node, further reducing the load on the shared file system. For the same reason, we also stored the Python virtualenv with the OpenEye docking modules on the local storage of the nodes.

Fig. S1 shows core utilization (Fig. S1(a)), docking call concurrency (Fig. S1(b)) and docking call rate over time (Fig. S1(c)). The max docking time dominates the inefficiencies when the workload tapers off, as all idling nodes have to wait for a few remaining long running docking calls to terminate. The plots also show inefficiencies at start up, caused by: distributing input data, Python modules, and docking calls across the compute nodes; and preparing data structures in memory. Those inefficiencies lead to an average resource utilization of 91% and an average docking call throughput of 139M * 0.91 = 126M docks/hr, where a dock encapsulates the entire unit of scoring the conformational ensemble of a ligand. The achieved peak rate of 144M docks/hr indicates an uneven distribution of docking call times throughout the data set, but also shows that resource utilization during steady state is near perfect, as confirmed by Fig. S1(a). The min, max, and mean components of the performance data characterize the distribution of docking time per ligand. Given each ligand has a different conformational landscape and overall size, this impacts the overall time it takes to dock a single compound. Therefore, we highlight the average time to dock a single ligand, the min and the maximum.

Performance of Computational Models

Several recent papers have reported impressive performance and scaling results. However, given the diverse computing platforms and docking programs employed, as well as different measures of performance, it is difficult to provide a head-to-head comparison. We discuss two “state-of-the-art” publications, that represent the broad spectrum of performance considerations and design points.

VirtualFlow[?] submits multiple different “jobs” to different “clusters” and report a peak utilization of 160k core. However they do not report how effectively the resources are utilized and focus on application performance measures (docks/time). In contrast to our solution which submits one single job to the supercomputer and manages the entire workload within the boundaries of that single job. Our workflow on Frontera uses more than 400k cores at peak, while reporting a resource utilization consistently above 95programs are different, a direct throughput comparison is not meaningful.

In a recent submission[?] a team from ORNL demonstrated a performance of 50M docks/hour for up to 20 poses per dock (i.e., 20k docks/second) on the Summit supercomputer. This was primarily achieved through adaptation of AutoDock-GPU[?] — GPU offloading feature calculations in rescoring, and the use of GPU-accelerated database query software. Additionally, a further 10x performance improvement was achieved by using parallel database methods. In contrast, our work reports a general-purpose execution tool that is not constrained to a specific docking program, nor is it limited optimizations on a specific computing platform.

Assay details

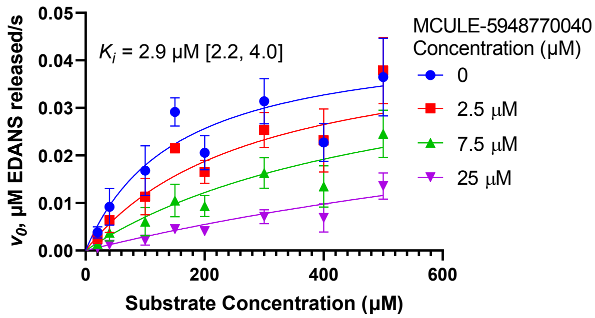


Figure S2: Quantitative high-throughput mass spectrometry-based endpoint assay depicting the K_i determined for MCULE-5948770040 as it binds to its substrate (M^{pro}).

Table S2: Plate-based M^{pro} activity inhibition screen results.

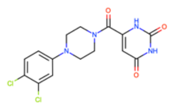
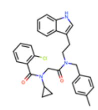
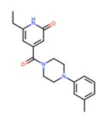
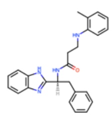
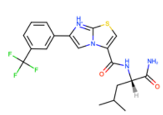
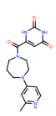
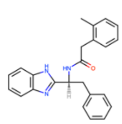
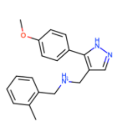
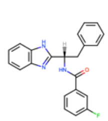
Structure	Compound	Z-Score	% Residual activity at 20 μM
	MCULE-5948770040	13.7	12%
	MCULE-9466928660	6	62%
	MCULE-8841098920	5.5	64%
	MCULE-9437816791	4.2	73%
	MCULE-9635688656	3.3	78%
	MCULE-5271818360	2.9	81%
	MCULE-2322094178	2.6	83%
	MCULE-5715268546	2.6	83%
	MCULE-6499664919	2.6	83%

Table S3: Extended Plate-based M^{pro} activity inhibition screen results.

Compound	Z-score	% Residual activity at 20 μM
PC	-0.1	101%
PC	-1.7	111%
PC	0.8	95%
PC	-1.0	107%
PC	1.3	91%
PC	0.8	95%
PC	-0.1	101%
PC	0.0	100%
NC	16.8	-8%
NC	14.9	4%
NC	15.4	1%
NC	15.1	3%
MCULE-5948770040	13.7	12%
MCULE-9466928660	6.0	62%
MCULE-8841098920	5.5	64%
S-613265359	4.9	69%
S-613265357	4.6	70%
S-613265358	4.6	71%
MCULE-9437816791	4.2	73%
MCULE-9635688656	3.3	78%
S-613265352	3.2	79%
MCULE-5271818360	2.9	81%
S-613265354	2.9	81%
MCULE-2322094178	2.6	83%
MCULE-5715268546	2.6	83%

MCULE-6499664919	2.6	83%
MCULE-5087943183	2.4	84%
MCULE-5517230918	2.4	84%
MCULE-6069061169	2.4	85%
MCULE-9796670140	2.2	86%
MCULE-3840831569	2.2	86%
MCULE-1316421825	2.2	86%
MCULE-1691783948	2.1	86%
MCULE-9073095150	2.1	87%
MCULE-6576711707	2.0	87%
S-613265355	2.0	87%
MCULE-9274314730	1.9	88%
MCULE-6419793505	1.9	88%
MCULE-7344200039	1.9	88%
MCULE-2015327629	1.8	88%
MCULE-2233915378	1.7	89%
MCULE-5637537904	1.5	91%
MCULE-9408357356	1.3	92%
MCULE-7367467944	1.2	92%
MCULE-3860138249	1.1	93%
MCULE-4793474277	1.1	93%
MCULE-3980688822	1.0	93%
MCULE-4125658730	1.0	94%
S-613265351	1.0	94%
MCULE-8462272476	0.9	94%
MCULE-2988500623	0.9	94%
MCULE-8798249417	0.8	95%

MCULE-7522626224	0.7	96%
MCULE-2310276978	0.6	96%
MCULE-6998567671	0.5	96%
MCULE-1618528086	0.5	97%
MCULE-1541843003	0.4	97%
MCULE-5145994619	0.4	98%
MCULE-6929749068	0.2	99%
S-613265356	0.0	100%
MCULE-7152686229	0.0	100%
MCULE-8568450548	0.0	100%
MCULE-4811796101	-0.1	101%
MCULE-2017284421	-0.2	101%
MCULE-7679599807	-0.3	102%
MCULE-5985443175	-0.3	102%
MCULE-3102049132	-0.4	102%
S-613265350	-0.4	103%
MCULE-9172580823	-0.7	105%
MCULE-9221479095	-0.8	105%
MCULE-4489145056	-1.1	107%
MCULE-7943829824	-1.4	109%
MCULE-6144665861	-1.4	109%
MCULE-1569445415	-1.6	110%
MCULE-1224951145	-1.8	111%
MCULE-5592055088	-1.8	111%
MCULE-1189442017	-2.1	113%
S-613265353	-2.1	114%
MCULE-7291272788	-2.1	114%

MCULE-6426577239	-2.3	115%
MCULE-4563504276	-2.4	116%
MCULE-6537132306	-2.7	118%
MCULE-3593629191	-3.0	119%
MCULE-3076850808	-3.3	121%

Crystal

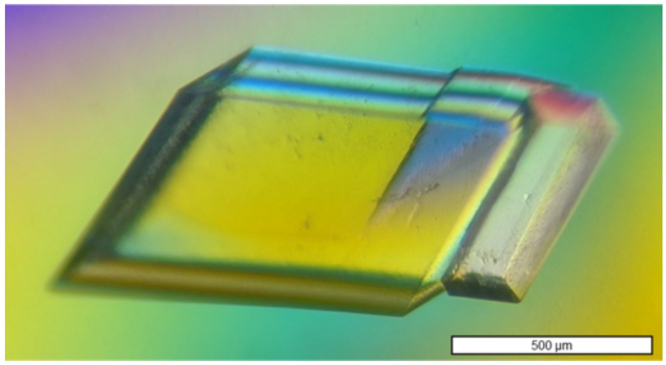


Figure S3: Protein crystal of SARS-CoV-2 3CL M^{pro} in complex with MCULE-5948770040 used for room-temperature X-ray data collection. Crystal measured ∼1x0.5x0.3 mm or ∼0.15 mm^3 and diffracted to 1.80 Å using a home source X-ray diffractometer.

Table S4: Crystallography data reduction and refinement statistics for room temperature structure of SARS-CoV-2 3CL Mpro in complex with MCULE-5948770040. Values in parenthesis indicate highest resolution shell.

	3CL Mpro- MCULE-5948770040 PDB ID 7LTJ
Data collection:	X-ray (in-house)
Diffractometer	Rigaku HighFlux Eiger R 4M
Space group	I2
Wavelength (Å)	1.5406
Cell dimensions:	
a, b, c (Å)	55.23, 81.47, 88.81
α, β, γ (°)	90, 90.56, 90
Resolution (Å))	59.85-1.80 (1.87-1.80)
No. reflections unique	35069 (3250)
R_{merge}	0.042 (0.585)
R_{pim}	0.022 (0.388)
$CC_{\frac{1}{2}}$	0.995 (0.524)
$I/\sigma I$	26.88 (1.23)
Completeness (%)	96.8 (89.6)
Redundancy	4.4 (3.1)
Refinement	
R_{work}/R_{free}	0.162/0.192
Ramachandran statistics	
Favored (%)	97.04
Allowed (%)	2.96
Outliers (%)	0
R.M.S deviations	
Bond lengths (Å)	0.012
Bond angles (°)	1.094
All atoms clashscore	1.91

MD simulations

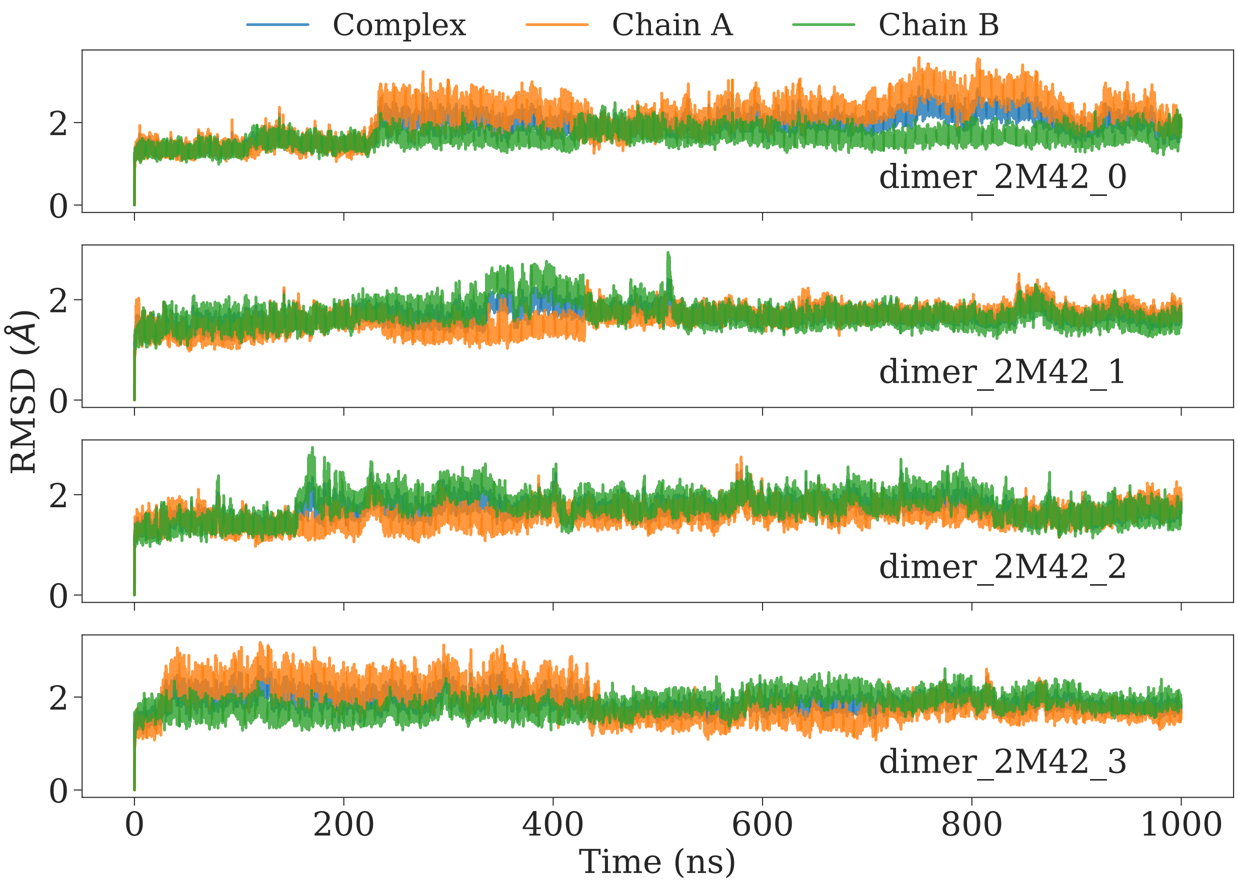


Figure S4: Root mean squared deviations from MD simulations of 2 MCULE-5948770040 molecules bound to each M^{pro} protomer. The RMSD for the complex is depicted in blue color whereas the individual protomers (chains A and B) are depicted in orange and green respectively.

ANCA-AE implementation details

In the methods section, we provided an overview of how ANCA-AE was implemented. Here we provide supporting information regarding its evaluation. In particular, ANCA-AE implements an autoencoder that takes as inputs the linear projections of the simulation datasets analyzed using ANCA. For our implementation, we noted that the optimal set of linear dimensions for ANCA with the LB and LF simulations (where the ligand was bound to both protomers) was 40, whereas including the simulations with the ligand bound to only one protomer, the optimal number of

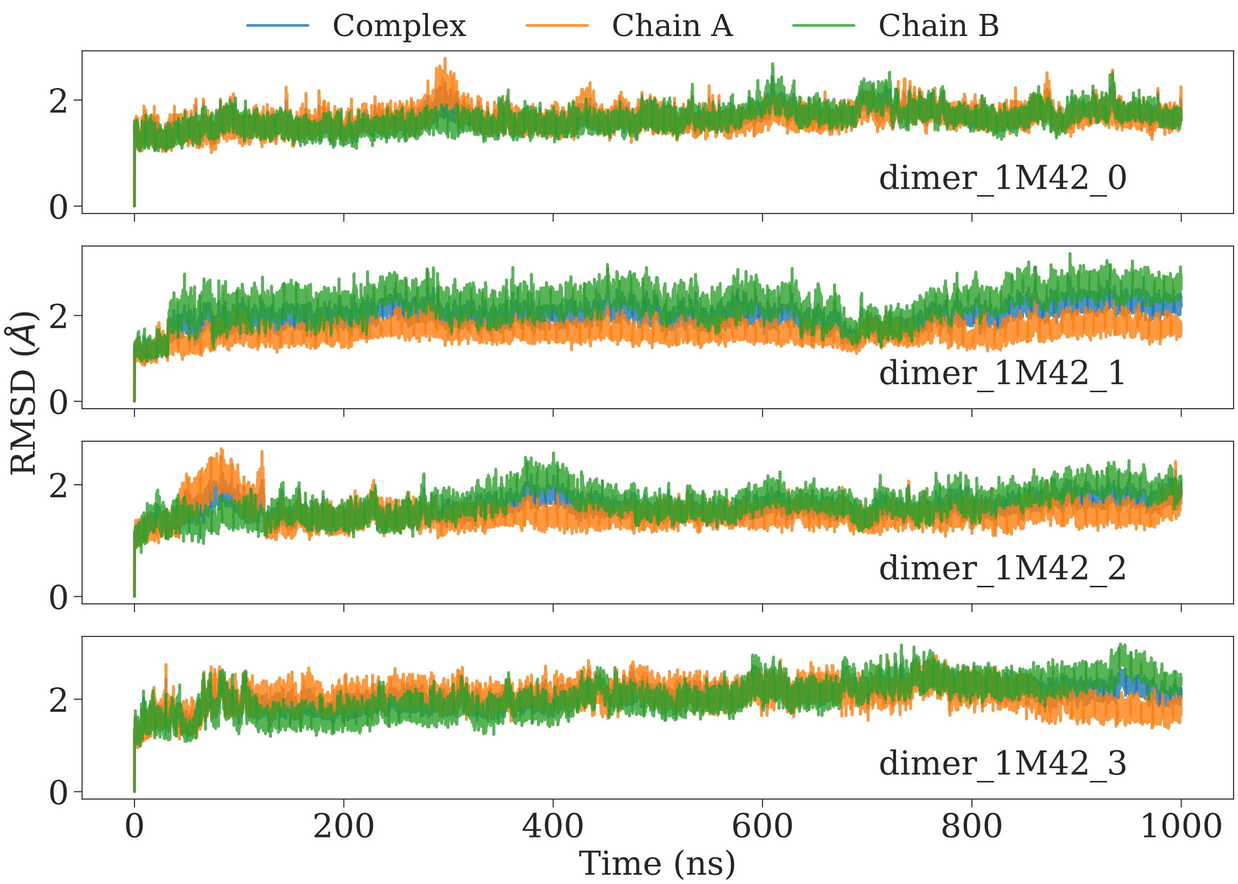


Figure S5: Root mean squared deviations from MD simulations of one MCULE-5948770040 bound to one of the M^{pro} protomer chains in the dimer. The RMSD for the complex is depicted in blue color whereas the individual protomers (chains A and B) are depicted in orange and green respectively.

dimensions for ANCA was 50. We then mapped the autoencoder loss (based on the mean squared error in reconstructing the latent space) as a function of the number of latent dimensions (Fig. S7A), which showed that as we increased the number of latent dimensions, the loss value typically reduces. The other plots (Fig. S7B-D) summarize the characteristics of training and validation loss based on the model and the hyperparameters that were used in our training runs.

Analysis of MCULE-5948770040 bound to individual protomer units

We present an analysis of MCULE-5948770040 bound to one of the protomer units, rather than both protomers to check (1) if such a simulation would be stable, and (2) if the ligand bound to

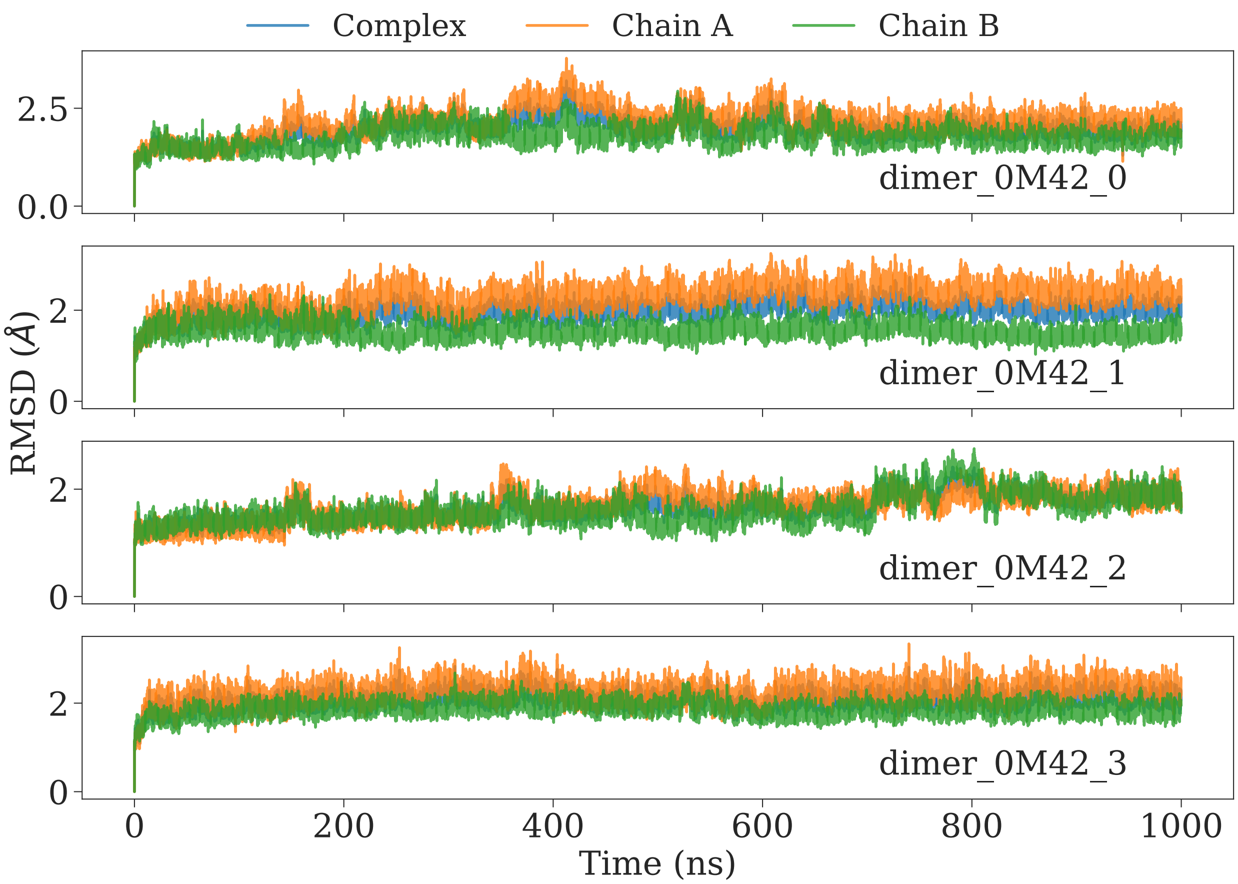
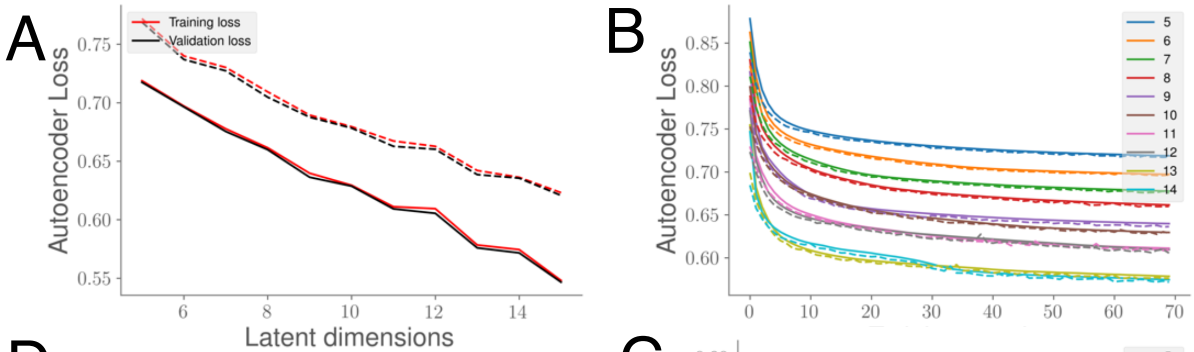


Figure S6: Root mean squared deviations from MD simulations of M^{pro} ligand-free state.

a particular protomer would induce the same long-range stability of the region R5 from our data. Our analysis here based on the RMSF profiles in Fig. S9A shows that compared to the ligand-free (LF) simulations, the RMSF of the ligand-bound (LB) states (either one or two ligands, indicated in the brackets) are significantly smaller. Thus, the presence of the ligand in either protomer can stabilize R5, apart from the loops in the binding region (identified as S2, S1, S4 and S3).

Further, the projections of these simulations using the ANCA-AE methods also illustrates that there are significant differences in the collective motions characterized by the LB and LF simulations (Fig. S9B and Fig. S11B). Notably, while there is some overlap between the LF and LB states in their collective motions, we observe that once the ligand binds to a particular chain, the fluctuations in LB-chain B are different fro LB-chain A as is often seen by the orientation of the ligand in the binding site. This illustrates that the conformational changes sampled by each of



D

	Hyperparameter	Value
0	Batch Size	64
1	Optimizer	RMSprop
2	Affine Widths	[32]
3	Output Activation	Tanh
4	Loss Function	mse
5	Latent Dimension	8
6	Learning Rate	0.001
7	Weight Decay	0.00001

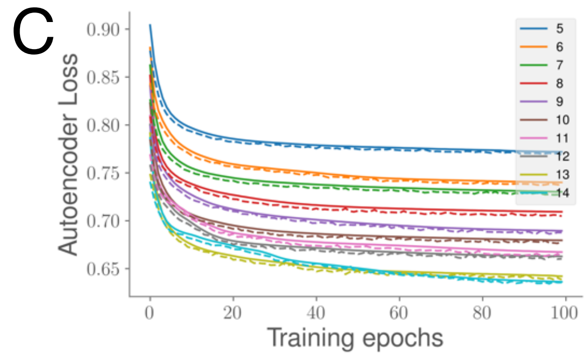


Figure S7: Characterizing how ANCA-AE identifies conformational substates from long timescale simulations. (A) Mapping the autoencoder loss versus the latent dimensions for training (red lines) and valiation (black) data based on 80:20 split of the simulation data. The solid lines represent the simulation datasets from the LB and LF states where the ligand was bound to both protomer units whereas the dotted line represents the loss based on including the simulations where the ligand was just bound to one of the protomers. (B) and (C) represent the training curves for ANCA-AE based on the number of latent dimensions (from 5-14) for both set of simulations as in (A). The solid lines track the training loss where as the dotted lines represent the validation loss as tracked by ANCA-AE. (D) summarizes the hyperparameter settings we used in training the ANCA-AE model.

the LB-states are different and gives rise to unique set of conformational states sampled. These motions can be attributed to the additional flexibilty observed in the loops surrounding the ligand binding pocket(s) in the LF-state (and when the ligand is absent from one of the chains), giving rise to the collective motions that were shown in Fig. 4 of the main text.

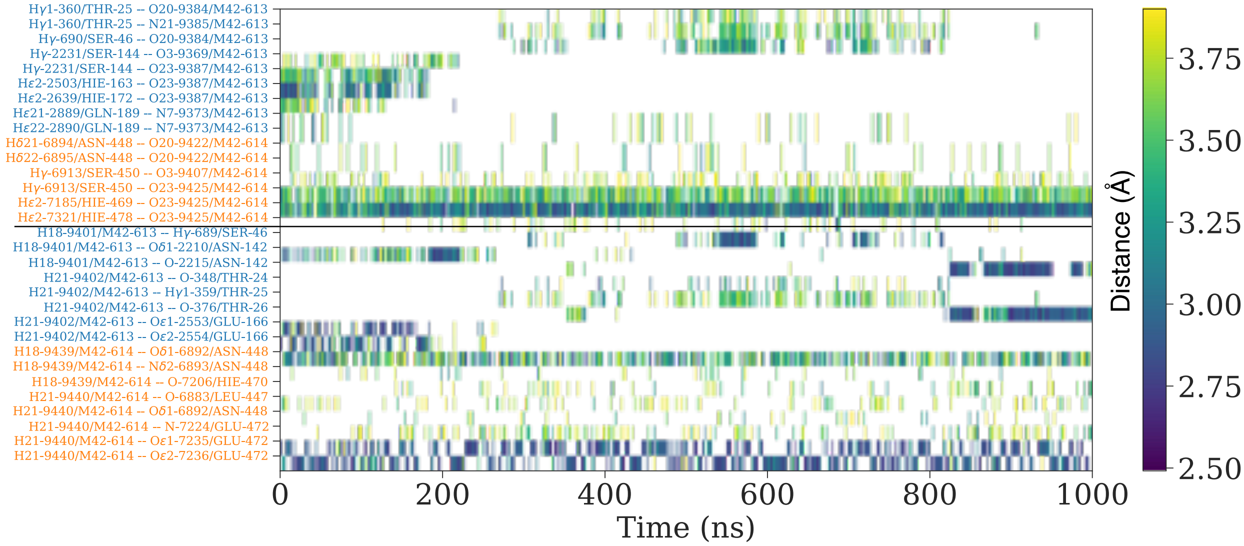


Figure S8: Hydrogen bond analysis in the LB simulations (bound to both protomers) show distinct behavior of the interaction patterns between the two protomers. The hydrogen bonding distance between the ligand and the protomer A is shown on the top with a solid line depicting the separation between the two protomers. While we observe stable hydrogen bonding patterns (distance between the heavy atoms ¡ 2.75 $\mathring{A}$) in protomer B, this is not observed in protomer A, where many hydrogen bonds are transient. Given that we observe a similar behavior with a second set of force-field parameters (calculated using NWChem), we posit that such fluctuations across the different protomers are unique.

Docking Pipeline Result Comparison

Overlap between X-Chem fragments and MCULE-5948770040 We retrieved the fragments for 3CL-M^{pro} from X-Chem fragment crystalgraphic screens.[?] Fragments did not have a high fingerprint similarity, with two common fragments appear in the highest 0.4 bin (see Fig. S12a). The crystals were aligned in PyMol and visualized in Fig. S12b.

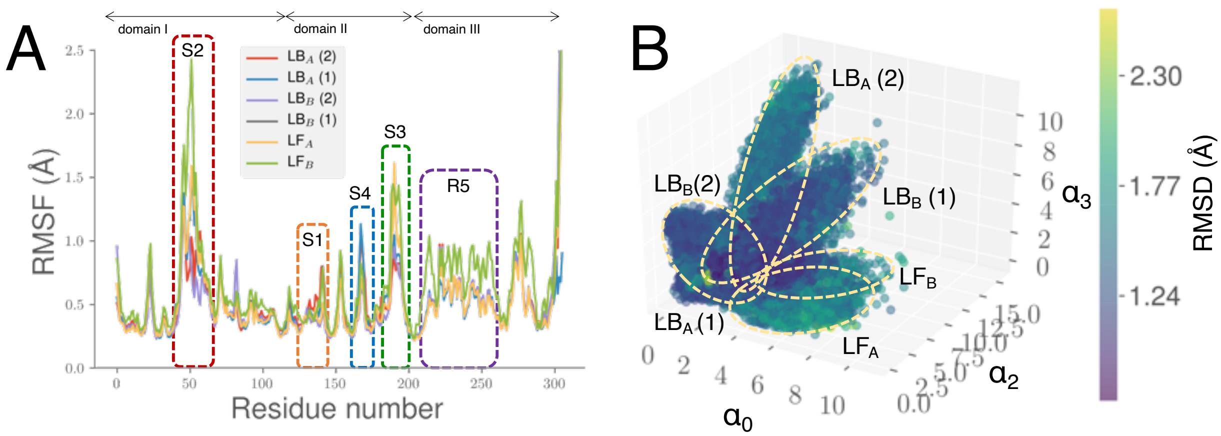


Figure S9: Conformational landscape of the LB-states as a consequence of binding MCULE-5948770040 to one of the protomers and its comparison to the other simulations. (A) Root mean squared fluctuations from MD simulations of M^{pro} the LB- and LF-states. The number in the brackets for the LB-states indicate the number of ligands bound to the complex. While we find similar regions affected by the ligand (Fig. 4 of the main paper), there are subtle differences in the overall profiles, indicating that a single molecule of MCULE-5948770040 bound to one of the protomers does not alter the overall conformational fluctuations as two molecules of the ligand bound to the complex. (B) Collective conformational fluctuations as characterized by ANCA-AE demonstrate the presence of distinct conformational states between the LB- and LF-states. The ellipses indicate a mixture of Gaussians taht depict how clusters are elucidated; further, simulations can be separated as shown in Fig. S11B.

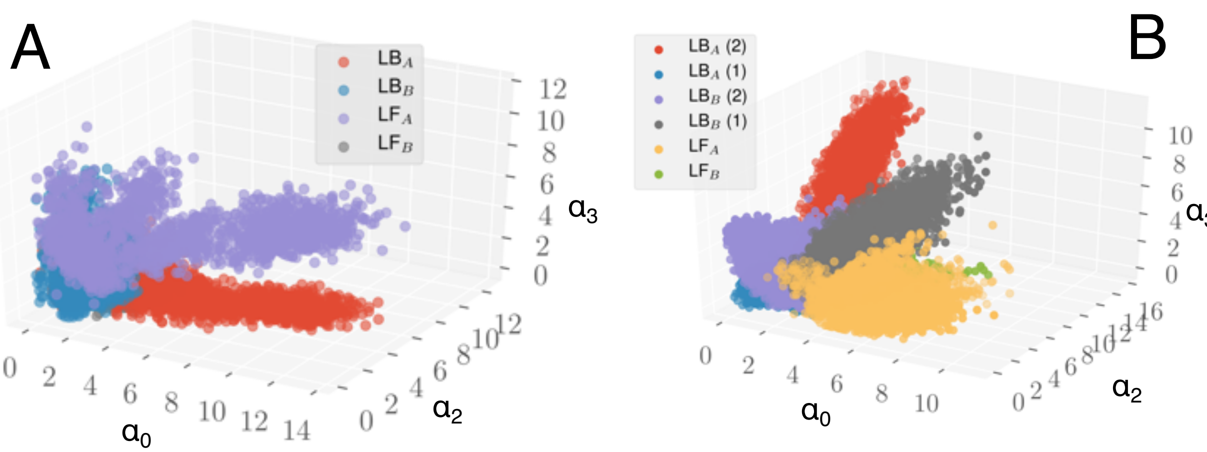


Figure S10: Conformational states sampled by the LB- and LF-states indicate distinct conformational transitions. (A) The projections from three ANCA-AE dimensions depicting distinct conformational transitions in the LB- and LF-states; notably the fluctuations in the LB-states in the two chains are quite distinct. (B) Same information as in (A), but includes projections from the two additional simulations where the ligand was bound only to one protomer units.

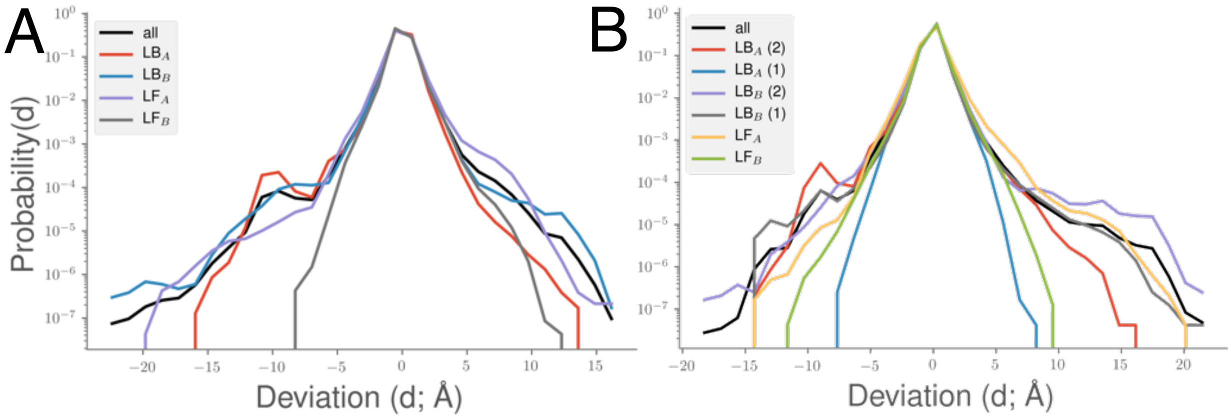


Figure S11: Presence of significant long-tails in the atomic fluctuations of the M^{pro} simulations, as quantified by the histograms of the positional deviations of the C$^{\alpha}$ atoms in our simulations. (A) depicts the long tails from the LB-(ligand bound to both protomers) and LF-states; (B) depicts the same information along with the simulations where the ligand was bound to only one of the two chains. Note that the fluctuations when including the single ligand bound to the protomer actually increases the overall fluctuations as observed from the deviation profiles.

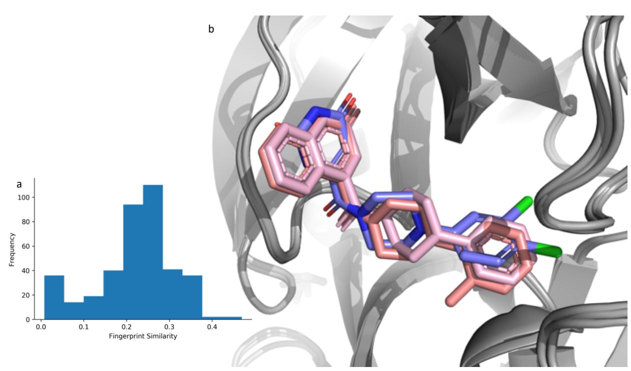


Figure S12: (a) Histogram outlining similarity of X-Chem 3CL-M^{pro} fragments with MCULE-5948770040 utilizing 2D fingerprint similarity. (b) X-Chem fragments x3303 and x11366 overlain with MCULE-5948770040.[?] There is an overlap on the P1 and linker, but notice the difference in the P2 region.