**Results**

**All-atom molecular dynamics simulations rapidly reveal known cryptic pocket openings**

There have been few studies which systematically evaluate whether MD simulations using modern force fields recapitulate known cryptic pockets. In previous work, MD simulations captured opening at known cryptic sites in TEM Beta-lactamase, Interleukin-2, and several other protein drug targets<citation(s) >. MD simulations have also identified novel cryptic pockets whose existence was subsequently experimentally verified <citation>. However, these studies focused on a narrow set of proteins or did not evaluate whether MD simulations accurately recapitulated the *holo* protein conformation. We were also interested in generating additional training data to train machine learning models to identify the locations of cryptic pocket formation.

Hence, we conducted unbiased adaptive sampling MD simulations of 12 proteins known to form cryptic pockets from *apo*, or ligand-free, starting structures in the CryptoSite dataset. Eight of the pairs have ligand-binding residues which were closer together in the *apo* structure than in the *holo* structure (i.e. they required an opening motion to form the pocket). Multiple different types of motion are represented, including three cases of secondary structure change. After launching 10 parallel simulations from the *apo* structure, we constructed a Markov State Model (MSM) of the conformational ensemble and prioritized structures for the next round of simulations using a ranking function that balances exploitation (i.e. prioritizing states with large pockets) with exploration <Cite FAST>. This procedure was repeated four times to generate 5 ‘swarms’ of simulations. We hypothesized that cryptic pockets would open only after several rounds of adaptive sampling.

To our surprise, we found that the majority of cryptic pockets opened in just 10 parallel simulations of 40 nanoseconds. Pockets were counted as having opened in simulation when the pocket volume of a simulated structure reached or exceeded the *holo* crystal structure pocket volume (see Methods). Eight out of the 12 proteins simulated opened in the first 10 simulations of 40 nanoseconds. Two others opened after 4 and 5 40 ns rounds of adaptive sampling (each also consisting of 10 parallel simulations). The other two, which were not distinguished by particularly complex conformational changes, remained closed even during adaptive sampling simulations. The rapid pocket opening we observe suggests that for most smaller proteins a modest amount of simulation data may be enough to discover cryptic pockets. Additionally, this finding suggests that machine learning models trained to predict cryptic pocket formation over small simulation time windows (i.e. 40 ns) would also be able to identify cryptic sites in ligand-free experimental structures.

Diagram

Description automatically generated

* For figure caption: each simulation round consists of 10 40 ns simulations

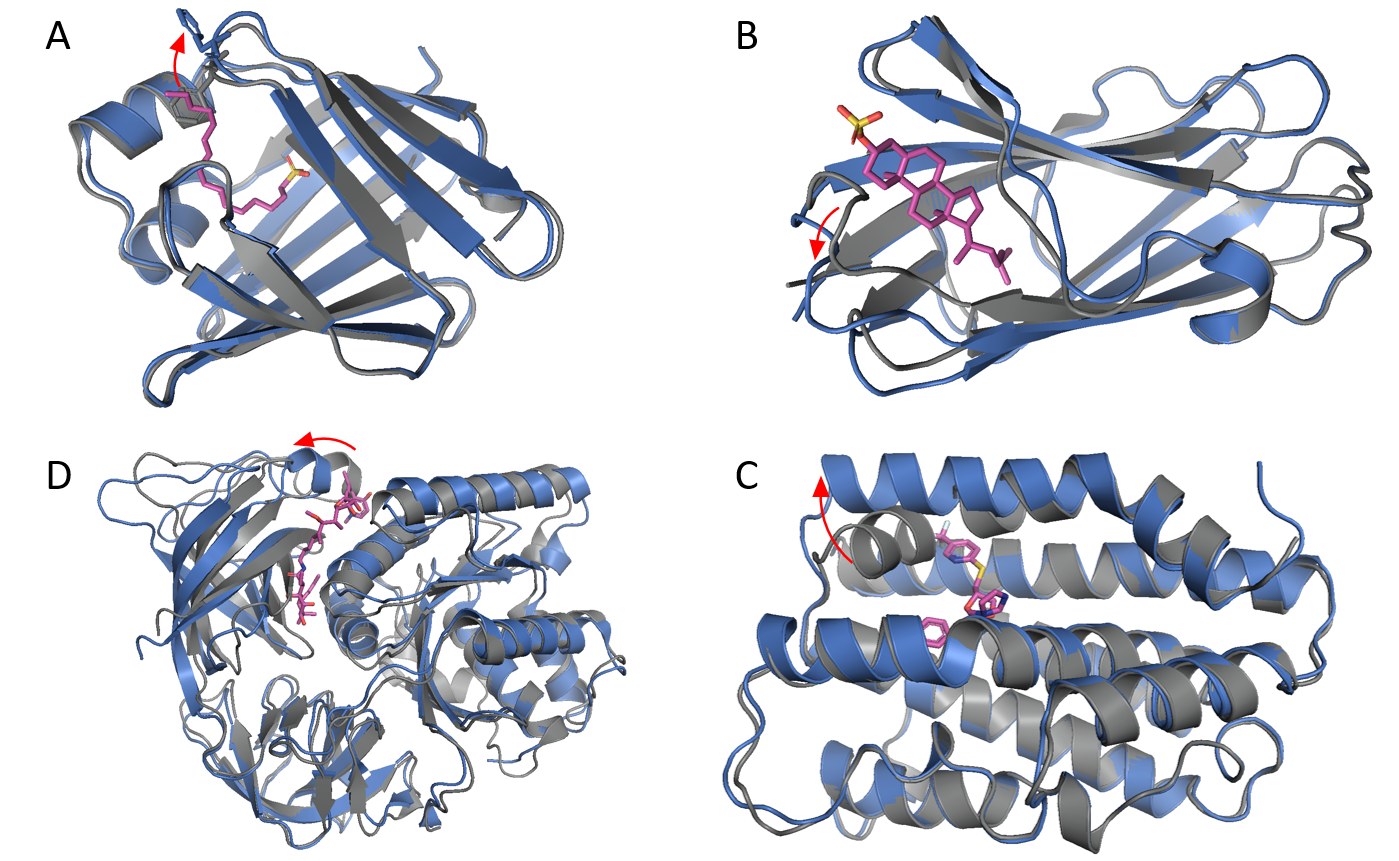


Figure 1: Four classes of pocket opening motion, with apo structures in grey, holo structures in blue, and ligands in magenta: A. sidechain motion only (PDB 1ALB *apo*, 1CIB *holo*), B. loop motion (PDB 1NEP *apo*, 2HKA *holo*), C. secondary structure change (PDB 1NI6 *apo*, 3HOK *holo*), D. interdomain motion (PDB 1EXM *apo*, 1HA3 *holo*).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| number of proteins by motion and opening time | | aggregate simulation time required for pocket opening | | |
| 0.4 µs | 2 µs | did not open within 2 µs |
| pocket opening type | sidechain motion | 1 | 0 | 0 |
| loop motion and/or secondary structure element bending | 2 | 1 | 1 |
| ~~secondary structure element motion~~ | ~~0~~ | ~~1~~ | ~~0~~ |
| interdomain motion | 3 | 0 | 1 |
| secondary structure change | 2 | 1 | 0 |

Table 1: Structures with multiple kinds of secondary structure change are counted in the category in the lowest applicable row (i.e. a structure with sidechain and interdomain motions would be counted in the interdomain motion row).

If we want to be systematic here (though not free of arbitrary thresholds since one RMSD cutoff has to be set) we can define:

* secondary structure change with DSSP,
* interdomain motion by defining the domains with CATH (<https://www.cathdb.info/>) and an active site C-alpha RMSD cutoff, secondary structure element motion and loop motion with the same minimum C-alpha RMSD cutoff, and with the secondary structure assignments of the mobile elements to separate loop motion from secondary structure motion (i.e. loop-helix or loop-loop would be loop motion and helix-helix or helix-beta sheet would be secondary structure motion
* sidechain motion with the same C-alpha RMSD cutoff but as an upper limit
* Chart, scatter chart

  Description automatically generated
* Figure 2: Cryptic site pocket volume (Å ^3) vs cryptic site c-alpha RMSD to holo (Å) for simulations of PDB 1MY0 (apo). The lower and upper horizontal green lines are the apo and holo (PDB 1N0T) pocket volumes respectively, and the vertical green line is the apo-holo c-alpha cryptic site RMSD. A. k-means cluster centers after 10 40 ns simulations. B. k-means cluster centers assembled from 5 rounds of adaptive sampling with a total of 50 40 ns simulations. C. The difference in normalized densities between 400 and 2000 ns aggregate simulation

**A simple geometric definition of pocket opening accurately identifies residues which bind cryptic ligands from MD simulations**

Though cryptic pocket opening in MD simulations can easily be evaluated when the cryptic site is known, it is not clear what criteria should be used to prospectively determine whether a residue participates in a cryptic pocket. Several approaches have been developed to identify potential binding sites on the surfaces of protein structures. These include geometric algorithms that identify concavities within a protein structure, clustering-based algorithms that identify groups of residues with cooperative changes in solvent exposure <cite Porter 2019 exposons paper>, and ML-based approaches that determine the druggability of potential binding sites <cite TACTICS (https://pubs.acs.org/doi/pdf/10.1021/acs.jcim.1c00204)>. Some of these approaches specifically assign pockets to residues while others do not. We tested two common methods of identifying pockets, the LIGSITE algorithm and solvent accessible surface area (SASA), to determine if they accurately labeled sites of known cryptic pocket formation in MD simulations of the 12 proteins described above. The LIGSITE algorithm identifies points in space which are large enough to fit a small molecule and are surrounded by protein residues in multiple directions. Additional steps are then needed to assign these points to protein residues, involving decisions and parameters which require optimization.

We found that the LIGSITE algorithm accurately identified which residues participate in known cryptic sites within MD simulations. We calculated LIGSITE pockets and SASA for each frame in 10 parallel simulations of 40 ns of proteins known to form cryptic pockets. We then evaluated two different procedures for assigning LIGSITE pocket elements to nearby residues (either counting all pocket elements within 5 Angstroms of a residue or assigning each pocket element to the nearest residue). For each residue, we found the maximum increase in either SASA or residue-level pocket volume in a simulation time window (see Fig 3). We tested 40, 20, and 10 ns simulation windows in order to determine which produced the most accurate pocket volume predictions. We then evaluated which featurization scheme most accurately selected those residues known to bind a cryptic ligand. For our true labels, we considered residues within 5 Angstroms of the cryptic ligand as positive examples and assessed the performance of each featurization approach as a classifier. Our single best featurization scheme where each residue was assigned all LIGSITE pocket elements within 5 Angstroms and volume increases were calculated over a 40 ns window achieves an AUC of 0.44 under a Precision-Recall Curve and an AUC of 0.85 under the Receiver Operating Curve (ROC) across the 10 Cryptosite proteins which opened during molecular dynamics simulations. Thus, ML classifiers trained with this featurization scheme are likely to identify the locations of cryptic sites.

Fig 2

A picture containing diagram

Description automatically generated

Figure 2: The maximum pocket volume increase for each residue is calculated by scanning across all saved simulation frames in the window and comparing the pocket volume assigned to each residue in each frame to the volume assigned to that residue at t0. The volume changes are then binarized at a fixed threshold to produce labels for training.

**Materials for methods**

* Molecular dynamics simulations were performed in GROMACS using the Amber03 force field and TIP3P water with 100mM NaCl.
* Cryptic pocket lining residues were defined as all residues within 5 Angstroms of the cryptic pocket’s ligand in the holo structure.
* Cryptic pocket volume was calculated by identifying points in space surrounded by protein residues using the LIGSITE algorithm <citation>, assigning each such point to the nearest protein residue, and summing the number of LIGSITE pocket points assigned to all cryptic pocket lining residues.
* The 12 apo-holo structure pairs selected from CryptoSite had no gaps or non-canonical residues in the apo structure, and the structures were not parts of tightly packed multimers.
* No two of the twelve structures selected have more than 40% sequence identity. All are soluble globular proteins as expressed, and none have cryptic pockets lined by multiple protein chains.

This may be of exclusively internal interest

* ~~Our featurization schemes assigned pocket volume changes to protein residues based on the maximum increase in pocket volume over a 40 ns simulation window. Using a change in pocket volume rather than the total pocket volume is important because it provides a chance to distinguish between residues lining large static pockets and those lining cryptic pockets. Our use of only increases in pocket volume would not be justified at thermodynamic equilibrium, where one would expect to observe equal numbers of pocket volume increases and decreases for each residue as the cryptic pocket opened and closed. However, as cryptic pockets are by definition closed in the apo crystal structures from which our simulations are started and our simulations are not nearly long enough to have converged, the trajectories which we featurize are actively diffusing away from the starting structure and should therefore be enriched in pocket opening events with pocket volume increases~~.

**Materials for the SI**

**table S#1**

Results of FAST simulations of cryptosite proteins

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **apo id** | **holo id** | **ligand** | **simulation round at which pocket opens (x if it does not)** | **length (residues)** | **notes** |
| 4AKE | 1ANK | ANP | 1 | 214 | pocket forms when two domains move towards each other |
| 1BSQ | 1GX8 | RTL | 1 | 162 | beta barrel |
| 1ALB | 1LIC | HDS | 1 | 131 | beta barrel |
| 1EX6 | 1GKY | 5GP | 1 | 186 |  |
| 2BLS | 3GQZ | GF7 | 1 | 357 |  |
| 2QFO | 2WI7 | 2KL | 1 | 207 |  |
| 1ADE | 1CIB | IMP | 1 | 431 |  |
| 1MY0 | 1N0T | AT1 | 1 | 258 | expressed protein is a deletion mutant with three internal transmembrane helices removed |
| 1NI6 | 3HOK | Q80 | 4 | 213 | secondary structure change involved in pocket formation in vitro; not recapitulated in silico  expressed protein is a truncation mutant missing the membrane anchor terminal domain |
| 3F74 | 3BQM | BQM | 5 | 180 | adaptive sampling partially pursues a second pocket  expressed protein is the extracellular domain of an integrin |
| 1NEP | 2HKA | C3S | x | 130 | beta barrel; relatively small pocket |
| 1EXM | 1HA3 | MAU | x | 403 | adaptive sampling pursues a second pocket |

**Table S#2**

Results of FAST simulations of proteins with large conformational changes

If we include this it should probably be as a spreadsheet (or at least rotated 90 degrees) for readability

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **apo id** | **holo id** | **ligand name(s)** | **apo-holo RMSD (nm)** | **force field** | **FAST rounds** | **opening motion** | **opening** | **local maximum pockets larger than the crystallographic one** | **relevant secondary structure change** |
| 6rvm | 5xdt | ZI7-Ca | 1.0590 | amber03 | 0; equilibrated | helix splits in two, changes its alignment, and pulls away from beta sheet; appears to be allosterically controlled | n/a | n/a | n/a |
| 2fjy | 2p70 | 2xPRZ | 0.5606 | amber03 | 10 | helix pulls out of core of protein | a more superficial pocket opened off to one side of the experimental pocket | a pocket forms between the core helix and a flexible outer helix | no |
| charmm36m | 5 | a pocket opened off to one side of the experimental pocket close to the experimental location | a pocket forms between the core helix and the other large helices as the core helix moves towards the flexible helix | central helix unwinds about halfway in state 000136, other ss changes not observed (although one involves terminal residues missing from the apo structure) |
| 3p53 | 6i11 | H0H\* | 0.3979 | amber03 | 5 | opening between two domains without significant secondary structure change (one loop hairpin forms half a turn of a helix but it's unclear from the crystal structures if that's essential) | did not open, one loop adopted a holo-like conformation without any domain separation |  | no |
| 1y1a | 1y1a | GSH | 0.2661 | amber03 | 5 | secondary structure change (helix->loop) combined with the outward motion of that loop | did not open, the two domains are connected by a flexible loop and moved a lot relative to one another | very large pockets form between the domains | no |
| charmm36m | 5 | did not open, the two domains are connected by a flexible loop and moved a lot relative to one another | very large pockets form between the domains | no |
| 2w9t | 2w9s | NDP:TOP | 0.1626 | amber03 | 5 | large loop movement; no secondary structure change | opened, although loop shape matches neither apo nor holo | none observed | n/a |
| charmm36m | 5 | did not open; adjacent loop moved instead | yes; semi-overlapping | n/a |
| 3fvj | 2b03 | TUD | 0.1470 | amber03 | 5 | loop-helix-loop movement, set of residues comprising the helix on the loop-helix-loop that moves to open the pocket shifts, there are a pair of apo-holo backbone dihedral rotations of 90-180 degrees which appear to be important | opened partially in restart 000613 | shallow cavity on face of protein around one end of pocket | no |

We did not capture cryptic pocket openings that required substantial secondary structure rearrangements.

* MD simulations using Amber03 and Charmm36m molecular dynamics force fields rarely reproduce secondary structure changes involved in pocket openings, limiting their ability to detect cryptic pocket openings involving such changes.
* As most of the pockets in CryptoSite involve relatively small conformational changes <RMSD histogram in SI or citation of cryptosite figure if necessary> and the PDB has roughly doubled in size since the compilation of the CryptoSite protein set <cite PDB inventory page>, we prepared a second set of apo-holo protein structure pairs using a similar method to identify additional cryptic pockets (see methods).
* We ran FAST simulations of five of these structures with the largest-RMSD pocket opening motions, four of which involve significant secondary structural changes.
* Only the cryptic pocket which opens via a loop motion opened in simulation.
* The other four pockets did not open and the necessary secondary structural changes did not occur for all but a few residues <cite table S# of these results in the SI>.