# Predicting cryptic ligand binding sites based on normal modes guided conformational sampling

Wenjun Zheng

Department of Physics, University at Buffalo, Buffalo, New York 14260

Running title: Cryptic site prediction

Address: 239 Fronczak Hall, Buffalo, NY 14260

E-mail: wjzheng@buffalo.edu

Phone: (716) 645-2947

Fax: (716) 645-2507

**Key words**:

Area under the curve, cryptic site, conformational sampling, elastic network model, ligand binding, logistic regression, machine learning, neural net, normal mode analysis, random forest, receiver operating characteristic curve

# Abstract

To greatly expand the druggable genome, fast and accurate predictions of cryptic sites for small molecules binding in target proteins are in high demand. In this study, we have developed a fast and simple conformational sampling scheme guided by normal modes solved from the coarse-grained elastic models followed by atomistic backbone refinement and sidechain repacking. Despite the observations of complex and diverse conformational changes associated with ligand binding, we found that simply sampling along each of the lowest 30 modes is near optimal for adequately restructuring cryptic sites so they can be detected by existing pocket finding programs like fpocket and concavity. We further trained machine-learning protocols to optimize the combination of the sampling-enhanced pocket scores with other dynamic and conservation scores, which only slightly improved the performance. As assessed based on a training set of 84 known cryptic sites and a test set of 14 proteins, our method achieved high accuracy of prediction (with area under the receiver operating characteristic curve > 0.8) comparable to the CryptoSite server. Compared with CryptoSite and other methods based on extensive molecular dynamics simulation, our method is much faster (1-2 hours for an average-size protein) and simpler (using only pocket scores), so it is suitable for high-throughput processing of large datasets of protein structures at the genome scale.

# Introduction

A key prerequisite for successful drug design is to identify potential binding sites for small-molecule ligands to modulate target protein functions. These sites are often located in exposed and concave pockets with certain structural, dynamical, and physiochemical characteristics that favor molecular interactions. When such concave pockets are already formed in a ligand-unbound protein structure, various computational methods are available to detect them with reasonable accuracy [1,2]. However, in many cases, a binding site is "cryptic" (e.g., too open/closed or obstructed) in the absence of a ligand and only forms after binding to a ligand. A cryptic site can also form transiently in the absence of ligand, thus eluding experimental and computational structural characterization. Therefore, cryptic sites are often difficult to identify given a ligand-unbound protein structure. To computationally solve this problem, one must effectively sample protein conformational changes relevant to ligand binding. Existing methods like long-time molecular dynamics (MD) simulation [3] and flexible molecular docking [4] are highly expensive and not applicable if high throughput is required. Therefore, there is a high demand for developing accurate and efficient methods to predict the location of cryptic sites in a given ligand-free protein structure. Discovery of cryptic sites will expand the druggable genome, and offer new venues of drug design by targeting hidden allosteric sites [5] and undruggable protein-protein interfaces [6].

In a recent landmark study by Cimermancic et al [7], the CryptoSite program/server was developed based on a dataset of known examples of cryptic sites in 93 proteins (each protein has an unbound structure that harbors a cryptic site with low pocket score and another structure bound with a functionally relevant ligand at the cryptic site). This valuable dataset has proven useful for developing/assessing methods for predicting cryptic sites [8,9]. CryptoSite used conformational sampling by MD simulation to generate an ensemble of protein conformations for training a machine learning method to detect cryptic sites [7]. Cimermancic et al used an energy-landscape based MD protocol[10] which is faster than conventional all-atom MD simulation [7]. Alternative structural simulation options are available at both atomistic and coarse-grained levels. For high-throughput applications, coarse-grained models and simulations have the advantage of low computing cost and applicability to large proteins and long time scales. For example, an elastic network model (ENM) represents a protein structure as a network of springs connecting neighboring C$\alpha$ atoms [11-13]. Despite its simplicity, the normal mode analysis (NMA) of ENM can yield low-frequency modes of collective domain motions, which often capture conformational changes observed between experimentally solved protein conformations [14]. NMA provides an efficient way to generate multiple receptor conformations for ensemble docking by deforming a given protein structure along a few low-frequency modes that capture collective motions at the backbone level. NMA greatly reduces the high-dimensional conformational space to be sampled. In fact, NMA was shown to provide better coverage of the protein conformational space than MD simulation in explicit solvent for a large set of proteins [15]. Previous studies used all-atom and coarse-grained NMA to refine protein-ligand bound structure[16,17], perform flexible ligand-receptor docking [18], and sample conformational changes in ligand binding pockets [19,20]. However, the usage of NMA is subject to the following caveats: 1. Among numerous modes, the lowest few (e.g., 10-20) modes are often not adequate in describing protein functional motions[21], and it is difficult to select those modes relevant to ligand binding without the knowledge of the binding-site location (e.g, using a measure of relevance [19,20]). So it is conceivable that many modes and their numerous combinations may have to be sampled with high computational cost. 2. The ability of coarse-grained modes to adequately capture small binding-pocket changes is not established relative to their proven capacity in describing collective inter-domain motions[14]. Indeed, a recent large-scale study found the use of ENM-based NMA to be of limited applicability in small-molecule docking [22].

To address the above caveats of NMA, this study will focus on validation and optimization of normal modes guided conformational sampling to facilitate accurate prediction of cryptic sites. This task is formulated as a classification machine learning (ML) problem: each residue is labeled as either in a cryptic site or not based on their various features/scores (including sampling-enhanced pocket scores and other complimentary scores like residue conservation, B-factors, and dynamic importance, see Methods). To accurately model conformational changes at atomic resolution, we refine the Cα-only conformations from normal-mode displacement using fast programs for backbone refinement and sidechain repacking (see Methods). As a result, our protocol can model both backbone and sidechain conformational changes with high efficiency. Surprisingly, we found simple and fast sampling along each of the lowest 30 modes to be near-optimal without the burden of exploring lots of modes and their exponentially numerous combinations. We further developed ML protocols to optimize the combination of the sampling-enhanced pocket scores with other dynamic and conservation scores, which only slightly improved the performance. Overall, our protocol achieved similarly high accuracy to predict cryptic sites as CrypoSite. Compared with MD-based sampling in CryptoSite, our NMA-based protocol achieved better sampling of pocket-forming conformations, resulting in more cryptic-site residues identified (using threshold of 0.05 for pocket scores). Our protocol is also much faster (1~2 hours for an average-size protein vs 1-2 days by CrypoSite) and simpler (using just 2 pocket scores plus 3 optional scores vs 58 features in CrypoSite). Therefore, our method is suitable for high-throughput processing of large datasets at the genome scale.

# Materials and Methods

## Elastic network model (ENM) and normal mode analysis (NMA)

In an ENM, a protein is represented as a network of coarse-grained beads corresponding to the Cα atoms of protein residues. Harmonic springs link all pairs of residues within 25 Å [23] which ensures adequate local connectivity to avoid unwanted zero modes. Meanwhile, to avoid introducing unphysical long-range coupling, we use a distance-dependent nonbonded force constant that decays by 50% at $R_c$ = 10 Å, and a 10-fold larger force constant for the bonded interactions (i.e. between residue $i$ and $i$+1).

The ENM potential energy is:

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{i-2}\frac{\theta(25-d_{ij,0})(d_{ij}-d_{ij,0})^2}{1+e^{d_{ij,0}-10}} + \frac{1}{2}\sum_{i=1}^{N}10(d_{ii+1}-d_{ii+1,0})^2 , \qquad (1)$$

where $N$ is the number of residues, $\theta(x)$ is the Heaviside function, $d_{ij}$ is the distance between residue $i$ and $j$, $d_{ij,0}$ is the value of $d_{ij}$ as given by a reference structure. We tested other values of $R_c$ between 7 Å and 15 Å, and verified that the NMA results are similar but slightly worse.

NMA solves eigen decomposition for a Hessian matrix $H$ which is obtained by calculating the second derivatives of the above ENM potential energy [24]:

$$HV_m = \lambda_m V_m, \tag{2}$$

where $\lambda_m$ and $V_m$ represent the eigenvalue and eigenvector of mode $m$, respectively. After excluding six zero modes corresponding to three rotations and three translations, we number non-zero modes starting from 1 in the order of ascending eigenvalue.

For mode $m$, we use a perturbation analysis to assess how much the eigenvalue changes (represented as $\delta\lambda_m$) in response to a local perturbation at a chosen residue position $n$ [25-27] (i.e., by uniformly weakening the springs connected to this position to mimic a point mutation). Then we average $\delta\lambda_m / \lambda_m$ over the lowest $M$ modes to assess the dynamic importance of this residue position [28]:

$$D_n = \frac{1}{M}\sum_{m=1}^{M} \delta\lambda_m / \lambda_m . \tag{3}$$

This dynamic importance score measures local flexibility (i.e. total deformation of springs connecting residue $n$ to its neighbors) which is known to be enhanced moderately near a cryptic site [8].

To validate ENM-based NMA, we compare each mode (i.e., mode $m$) with the observed structural change $X_{obs}$ between two distinct protein structures (i.e., one unbound structure and another ligand-bound structure) by calculating the following overlap coefficient [29]:

$$O_m = X_{obs} \cdot V_m / | X_{obs} |, \tag{4}$$

where $|O_m|$ varies between 0 and 1 with higher value meaning greater similarity. $O_m^2$ gives the fractional contribution of mode *m* to $X_{obs}$. The cumulative overlap $CO_M = \sum_{m=1}^{M} O_m^2$ gives the fractional contribution of the lowest $M$ modes to $X_{obs}$ [29], and can be used to assess the accuracy of NMA in describing a given conformational change.

To assess local flexibility at individual residue positions as described by the lowest *M* modes, we calculate the following simulated B factors :

$$B_n = \sum_{m=1}^{M} \frac{1}{\lambda_m} \left( \left|V_{m,nx}\right|^2 + \left|V_{m,ny}\right|^2 + \left|V_{m,nz}\right|^2 \right), \quad (5)$$

where $V_{m,nx}$, $V_{m,ny}$, and $V_{m,nz}$ are the *x*, *y*, and *z* component of mode *m*'s eigenvector at residue position *n*.

The above two dynamic scores ($D_n$ and $B_n$) are used to rank residues for predicting cryptic sites, along with other scores (including pocket scores and residue conservation scores).

## NMA-guided conformational sampling

We use the following two NMA-guided sampling schemes:

**Single-mode exploration**:

We sample along the direction of each mode's eigenvector without mixing modes. For a given mode *m*, we incrementally displace the Cα atoms along a displacement vector $X_m = \pm f_k V_m$ where $f_k$ is chosen for 5 discrete root mean square deviation (RMSD) values (*k*=1, 2, 3, 4, and 5 Å). For *M* modes, a total of 10*M* displaced models are generated.

**Multi-modes combination**:

We sample along directions given by a linear combination of the lowest $M$ modes: $X = f_k \sum_{m=1}^{M} \pm g_m V_m$, where $g_m$ are random numbers from [0,1] and $f_k$ is chosen for 5 RMSD values ($k$=1, 2, 3, 4, and 5 Å). For $M$ modes, a total of 30$M$ models are generated which provides a 3-fold enhancement of sampling than the single-mode exploration scheme.

Because a linear displacement of residue atoms along their Cα eigenvector would distort the covalent bonding geometry, the protein backbone is reconstructed and refined after the displacement using the PD2_ca2main program (downloaded from http://www.sbg.bio.ic.ac.uk/~phyre2/PD2_ca2main/ ). Then we repack side chain atoms using the RASP program V1.90 (downloaded from https://sourceforge.net/projects/raspv180/files/) and the SCWRL program V4.0 (downloaded from: http://dunbrack.fccc.edu/SCWRL3.php/). Since RASP runs faster than SCWRL, the latter is only used when RASP fails. We have tried various programs for reconstructing backbones and sidechains from coarse-grained Cα models (see a review [30]), and have chosen the above options based on considerations of computing speed and structural completeness.

## Pocket detection and scoring

The conformations generated by sampling are pooled into a structure ensemble for subsequent binding pocket detection by two state-of-the-art programs (fpocket and concavity, downloaded from: http://fpocket.sourceforge.net/ and http://compbio.cs.princeton.edu/concavity/). Both programs [1,2] combine physical/chemical and geometrical features to identify ligand binding pockets in a given protein conformation. Following Cimermancic et al. [7], we define the fpocket score as the maximum druggability score among the alpha spheres within 5 Å of the residue, and the concavity score is generated by concavity directly on a per-residue basis. Each pocket score is then averaged over the conformational ensemble generated above. We also explored other ways of summarizing the ensemble distribution of the pocket scores (e.g., computing the percentage of conformations with the pocket score greater than a cutoff value like 0.5), which did not improve over the simple average score.

# Results and Discussion

## Crypto-Site datasets and initial analysis by NMA

From the CryptoSite paper [7], we obtain a representative dataset of 84 known examples of cryptic sites including pairs of protein structures (i.e. one unbound and the other ligand-bound) in which the unbound structures have low pocket scores. This dataset is used as the training set for evaluating various predictive features and training ML protocols that combine these features. A separate test set of 14 pairs of protein structures and their cryptic sites are reserved for final testing (for details, see Table S1 of Supporting Information, and Table S1 & S5 in Cimermancic et al. [7]). By using the same datasets as CryptoSite, we can objectively assess our performance relative to CryptoSite.

The CryptoSite dataset covers a diversity of conformational changes dominated by various types of backbone motions (e.g. 45% loop motion, 17% domain motion, 16% motion of secondary structural elements, and 4% N/C-terminal flexibility, see ref [7]). Only 18% of these cryptic sites are formed solely by the movement of side chains, which are thought to have limited use for drug discovery due to low ligand binding affinity [8]. These diverse and slow backbone motions are challenging for all-atom MD simulation to explore but potentially accessible to Cα-based coarse-grained modeling. Most of these changes (77/84) are relatively small with Cα RMSD <5 Å (see Fig 1a), and only 3 of them show very large structural changes (RMSD >10 Å). Therefore, we will limit conformational sampling within a maximal RMSD of 5 Å by default to cover most of these cases.

To assess how well ENM-based NMA captures these observed conformational changes, we calculated the cumulative overlaps (CO) which gives the fraction of the observed changes as described by the lowest 30 modes (see Fig 1b). The average CO is only about 30%, and it varies widely from case to case: the lowest 30 modes only explain <50% of the observed changes in most cases (68/84). This seemingly discouraging finding, however, does not necessarily invalidate NMA's application in cryptic site prediction because: our goal is to sample not global conformations but local conformations near an unknown cryptic site, so that the cryptic site is exposed and deformed (e.g., from flat to concave) to allow fpocket and concavity to detect it with high pocket score. This is potentially a less challenging task than to accurately predict the local conformation of a cryptic site, and could be achieved via optimization of NMA-guided sampling to improve subsequent performance of the pocking finding programs. To our knowledge, no prior study has explored this possibility.

## Optimization of NMA-guided conformational sampling

If the location of a ligand binding site is known (or predicted by other means), this information can be used to select a few relevant modes to focus conformational sampling at the binding site for favorable ligand binding [19,20], or introduce local perturbations and then assess their effects on dynamics [31]. However, without knowing where the binding site is, one must consider all low-frequency modes and their numerous combinations which may be highly challenging to sample. To establish the feasibility of NMA in cryptic site prediction, we must answer two key questions: How many modes need to be sampled? How should they be sampled (individually or in combination)?

We have tested two NMA-based sampling protocols. The first (single-mode exploration) samples along each of the $M$ lowest modes separately without mixing them, making it computationally cheap (scales as $O(M)$). The second (multi-modes combination) linearly and randomly combines the lowest $M$ modes, so it is computationally more expensive (scales as $O(s^M)$ where s is the number of samples per mode).

To assess the sampling performance, we averaged fpocket and concavity scores per residue over the ensembles generated by NMA-guided sampling for the full training set, and predicted cryptic-site residues as those with high average pocket scores. Then we plotted a Receiver operating characteristic (ROC) curve to relate the sensitivity (true-positive rate) and the specificity (1 - false-positive rate) of prediction based on the selection of pocket scores with a varying threshold (set to 0.05 by default as in ref [7]). We summarized the prediction performance by calculating the area under the ROC curve (AUC).

To determine the optimal number of modes needed for the single-mode exploration scheme, we plotted the AUC of pocket scores for the full training set vs. number of modes ranging from 10 to 100. The performance of both fpocket and concavity scores is satisfactory with high AUC ~0.8. Both AUC curves peak near 30 modes, although concavity seems to perform slightly better and is less dependent on the number of modes (see Fig 2a). This encouraging finding relieves the concern of sampling hundreds of modes to cover the few modes relevant to ligand binding [19]. In this work, we use 30 modes by default.

Another key parameter for sampling is the maximal displacement (measured in Cα RMSD) explored along each mode. Given the observation of small RMSD values for most cases (see Fig 1a), one might expect the optimal RMSD to be small as well. To see this, we plotted the AUC of pocket scores for the full training set vs. five RMSD levels used for sampling (see Fig 2b). Surprisingly, the AUC increases steadily as RMSD increases to 5 Å. This finding highlights the distinction in sampling objective between accurate prediction of the ligand-bound conformation (with matching RMSD) and adequate displacement/deformation of the cryptic sites (with relatively large RMSD). It seems that the over-sampling of large conformational changes, despite causing large structural distortions, is advantageous for cryptic site prediction. Consistent with this, when we tried to filter out those conformations with high structural distortion, the prediction performance went down.

Next we test the multi-modes combination sampling protocol (using the lowest $M$=30 modes) to see if combining modes enhances sampling and improves cryptic site prediction. To this end, we plotted the AUC of pocket scores for the full training set vs. number of conformations sampled (up to 900). Compared to the single-mode exploration (with 300 conformations sampled along each of 30 modes), the performance of multi-modes exploration is very similar and insensitive to the number of conformations between 200 and 900 (see Fig 2c). We further compared the average pocket scores calculated based on the two sampling schemes and found they are highly correlated (with cross-correlation coefficient (CC) >0.98), which explains their similar performance in cryptic site prediction. Therefore, combining modes to enhance conformational sampling does not improve cryptic site prediction. It seems that favorable local changes in cryptic sites do not require specific combinations of multiple modes which largely act independently (because they are orthogonal to each other). This finding relieves the burden of sampling and ensures a very fast execution of NMA-guided sampling.

In sum, sampling along each of the lowest 30 modes with large RMSD (≤5Å) is a near-optimal sampling scheme with both satisfactory accuracy (AUC~0.8) and low computing cost (only ~300 conformations sampled). Contrary to general belief, using many more modes or linearly combining modes, although enlarging the conformal space for sampling, does not seem to improve the prediction of cryptic sites. Owning to low-cost sampling, our method runs very fast (e.g., 1-2 hours of CPU time is needed for an average-size protein of ~400 residues). In comparison, same calculation takes 1-2 days on the CryptoSite server.

## Individual evaluation of cryptic-site predicting features

After establishing the sampling scheme for generating structural ensembles, we then evaluate each of the following scores/features for predicting cryptic sites based on the structural ensembles. To this end, we plotted the ROC curve of each feature for all residues in the full training set and calculated the area under the ROC curve (AUC) to summarize its performance (see Fig 3a).

### Pocket scores: fpocket vs concavity

In CryptoSite, the most powerful features for cryptic site prediction are ensemble-averaged pocket scores from fpocket and concavity [7]. Indeed, we found both scores (used separately) can accurately predict cryptic sites (AUC=0.80 for fpocket and 0.81 for concavity, see Fig 3a). Because the fpocket and concavity scores are only moderately correlated (with CC=0.5), it is sensible to combine them (along with the other ancillary features, see below) through ML to optimize the performance (see below).

To further assess the performance of these pocket scores on a case-by-case basis, we separately plotted the ROC curves and calculated the AUC for each of the 84 cases in the training set (see Fig 3b). While concavity (AUC=0.81±0.19) performs slightly better than fpocket (AUC=0.80 ±0.14) on average, it exhibits more variation from case to case (see Fig 3b). Using a threshold of 0.05, both pocket scores can achieve an average sensitivity of ~0.8 and average specificity of ~0.6 (see Fig 3b). This is comparable to CryptoSite (average AUC = 0.77, sensitivity = 0.75, specificity = 0.66 based on the same training set).

**B factors**

Local flexibility is well known to be a key factor in ligand binding. In X-ray crystallography, protein structural flexibility is routinely characterized by the B factors which can also be calculated from MD simulation or NMA (see Methods). Cimermancic et al found a cryptic site is more flexible than a binding pocket with significantly higher normalized B-factors [7], and such conformational flexibility may enable them to readily convert from flat into concave shape. However, because binding pockets are rigid with low B factors, cryptic sites are still more rigid than non-binding-site residues. Indeed, we found B factors to be a weak negative predictor of cryptic site residues (i.e., selection of residues with low B factors can predict cryptic sites with AUC~0.59, see Fig 3a). Therefore, B factors cannot effectively exploit the local flexibility of cryptic sites for their prediction. To explain its weakness, B factors cannot distinguish global motions of the entire site and local/relative motions within the site, and only the latter are relevant to ligand binding.

**Dynamic importance**

As an alternative metric for measuring local flexibility/distortion in the context of ENM, we previously introduced the dynamic importance score based on site-specific perturbation of ENM (see Methods). It has been used to predict dynamically important hot-spot residues in previous studies [32-34]. The validity of this score assumes the lowest *M* modes accurately capture functional motions. We found this score is slightly better than B factors (with AUC=0.65, see Fig 3a) in predicting cryptic site residues, which can be attributed to its focusing on local/relative motions of a residue relative to its neighbors. Interestingly, the dynamic importance scores are weakly (yet significantly) correlated with both pocket scores (CC=0.26±0.01 for fpocket, 0.32±0.01 for concavity), supporting a possible causal linkage between NMA-enhanced local flexibility and higher pocket scores.

**Residue conservation**

Cimermancic et al. found that cryptic site residues are evolutionarily as conserved as those of binding pockets, and residue conservation is among the top three most effective features in CryptoSite[7]. By using residue conservation scores from the CONSURF webserver[35], we confirm that residue conservation is useful in predicting cryptic site residues (AUC=0.71, see Fig 3a).

In sum, based on our individual feature assessment, the performance of the above features is ranked in the following order: fpocket ~ concavity >> conservation > dynamic importance > B factor. While each pocket score can already achieve satisfactory performance by itself, it is potentially useful to combine them through ML so they complement each other and further improve the performance.

## Optimal combination of features through machine learning

To optimally combine the above scores, we train ML to predict cryptic sites in the training set of proteins with 84 known cryptic sites [7]. Unlike CryptoSite which used 58 features for ML, we only use a small set of features (i.e. two pocket scores plus three additional scores, see above), so there is no need to select features to avoid overfitting. Instead, we input all scores and let the ML algorithm decide how they should be optimally weighted. The hyper-parameters of each ML algorithm is tuned by maximizing the cross-validated AUC of ROC using five-fold cross validation: the training set of 27648 residues from 84 cases are roughly evenly divided into five subsets (with 5426, 5677, 5523, 5653, and 5369 residues, respectively), and four subsets are used for training while the fifth subset is used for validation and calculating AUC. For later comparison, the five-fold AUC for each pocket score was also calculated: AUC = 0.79±0.07 (fpocket), 0.81±0.09 (concavity)

After experimenting with various ML methods, we have achieved satisfactory and comparable performance using the following three methods:

1. **Logistic regression with LASSO regularization**:

Using the Glmnet package of R, we found the hyper-parameter lambda = 0.02 attained a maximal cross-validated AUC=0.83. The LASSO regularization removed B factors and dynamic importance while the remaining three scores were kept with coefficients 2.66 (fpocket), 3.47 (concavity), and 0.38 (conservation). This is consistent with our individual assessment of these scores which ranked the pocket scores higher than the conservation score. If only the two pocket scores are used, logistic regression can still achieve a similar performance (with cross-validated AUC=0.83 and lambda=0.005).

2. **Random forest**:

Using the RandomForest package of R (with 2000 trees), we found the hyper-parameter mtry = 1 attained a maximal cross-validated AUC=0.81. The relative importance of scores was ranked by Mean Decrease in Gini as follows: concavity (597) > fpocket (579) > conservation (527) > dynamic importance (475) > B factors (465). This is fully consistent with our individual assessment of the scores. If only the two pocket scores are kept, the random forest performance would go down (with cross-validated AUC=0.76).

3. **Neural net**:

Using the Neuralnet package of R (with a single hidden layer), we found the hyper-parameter of number of nodes in the hidden layer = 3 attained a maximal cross-validated AUC=0.83. The relative importance of scores is ranked by Garson's algorithm as follows: concavity (0.37) > fpocket (0.28) > conservation (0.24) > B factors (0.07) ~ dynamic importance (0.05). If only the two pocket scores are kept, the neural net can still achieve similar performance (with cross-validated AUC=0.83) using just one hidden node.

In sum, our training of ML that combined the scores has achieved slightly better performance on the training set than single pocket scores (with AUC increasing from 0.80 to 0.83). The combination of two pocket scores seems to be sufficient for the optimal performance of logistic regression and neural net, while the random forest requires the addition of other axillary scores.

## Final performance assessment on the test set

So far, we have only used the training set of 84 known cryptic sites and associated protein structures for optimization of sampling, evaluation of individual scores, and ML training. To rigorously assess the general performance of our method, we applied it to a held-out test set of 14 unbound structures with known cryptic sites (see Table S5 of ref[7]). For performance assessment, we plotted the ROC curves using the pocket scores calculated for all residues in 14 test cases, and calculated the AUC as a summary score. Encouragingly, we obtained an AUC of 0.83 for the fpocket score, and 0.77 for the concavity score. This is comparable to AUC=0.83 by CryptoSite on the same test set[7]. Using a threshold of 0.05, we attained a false positive rate (FPR) of 0.35 (fpocket) and 0.37 (concavity), and a true positive rate (TPR) of 0.87 (fpocket) and 0.83 (concavity), which are higher than CryptoSite [7] (FPR = 0.29 and TPR = 0.79). Therefore, our method can find more cryptic-site residues than CryptoSite, which implies that our NMA-guided sampling is more effective in promoting concave pocket formation at both cryptic sites (true positives) and nonbinding sites (false positives).

Next, we applied the above three cross-validated ML protocols to the test set (see Fig 4a), and obtained AUC = 0.84 (logistic regression), 0.84 (neural net), and 0.83 (random forest), which are comparable to the cross-validated

AUC on the training set. Although our trained ML protocols generalize well, their performance does not improve markedly than simply using the fpocket score for prediction (see Fig 4a). After comparing the ROC curves (see Fig 4a), we found the most improvement was on the region of ROC curve with low sensitivity (~40%) and high specificity (~90%).

To assess the usefulness of our method for drug design, we adopt the following criterion for accurate prediction of the cryptic site in a given protein: at least one-third of cryptic-site residues are identified (i.e. sensitivity ≥33%, following ref [7]). Reassuringly, all 14 proteins in the test set and all but 4 cases in the training set have met this accuracy criterion if residues are selected with either fpocket or concavity score >0.05. This gives an overall 96% success rate which is comparable to CryptoSite [7].

## Examples of successful cryptic site prediction

To illustrate our cryptic site prediction results (see Table S1 for full details), we discuss the following three examples from the test set which were also discussed by CryptoSite [7]:

**TEM1 β-lactamase**

This is a classical example of cryptic sites extensively studied by MD simulations and Markov state models [36,37]. In the unbound structure (PDB id: 1JWP), an allosteric cryptic site is buried by a short helix 11 and a long helix 12 (see Fig 4b). In the bound structure (PDB id: 1PZO), helix 11 moves by >3 Å to open a crevice for ligand binding. No such opening of the cryptic site is seen in any unbound structure despite its flexibility [8] . After NMA-guided sampling (see supplemental movie in Supporting Information), fpocket accurately identified this site (sensitivity=77%) with AUC=0.80 (compared to AUC=0.72 by CryptoSite). The predicted site includes three residues (A232, S249, and L286) validated by the thiol-labeling experiment [37].

**Exportin 1**

Exportin 1 is a large multi-domain protein with 1017 residues. In the unbound structure (PDB id: 4HB2), two helices (residues 52-541 and 570-585) form the ligand binding site but are too close to each other (see Fig 4b). In the bound structure (PDB id: 4HAT) they move slightly apart. The global conformational changes are small (RMSD=0.7 Å) with slight helix reorientation and sidechain rearrangements. After NMA-guided sampling, fpocket predicted the cryptic site (sensitivity=73%) with AUC=0.83 (compared to AUC=0.85 by CryptoSite).

**Interleukin-2**

This is a prototype example of cryptic sites at the difficult-to-drug protein–protein interaction interfaces. Near the main site is a disordered loop (residues 74-80) [8], which may contribute to high flexibility of the main site. The global conformational changes from the unbound structure (PDB id: 1Z92) to the bound structure (PDB id: 1PY2) are small (RMSD=1.4 Å). After NMA-guided sampling, fpocket predicted the cryptic site (sensitivity=80%) with AUC=0.67 (compared to AUC=0.65 by CryptoSite).

We next review more examples from the training set to cover various types of cryptic sites with different conformational changes.

**PTP1B**

Protein tyrosine phosphatase 1B (PTP1B) has only a single known cryptic allosteric site close to its C-terminus (Fig S1). The unbound structure (PDB id: 2F6V) has a well-resolved C-terminal helix (residues 285-299) that covers the allosteric site. In the bound structure (PDB id: 1T49) this helix is absent, so the allosteric site is accessible to ligand binding. After NMA-guided sampling, fpocket predicted the known cryptic site (sensitivity=71%) with AUC=0.75 (compared to AUC=0.56 by CryptoSite). Additionally, we also found another possible cryptic site (same as the CryptoSite-predicted site in Fig 3B of ref [7]), which is relatively close to the main active site (Fig S1). This new site was experimentally validated [7] and offers a promising target for drug design.

**GluR2**

Glutamate receptor subunit 2 (GluR2) has a cryptic site within a cleft between two domains. In the unbound structure (PDB id: 1MY1), several loops protrude into the cryptic site (Fig S1), but they move away upon ligand binding in the bound structure (PDB id: 1FTL). Therefore, binding to the cryptic site of GluR2 requires domain opening (RMSD=1.9Å) which is well described by NMA (CO=0.45). After NMA-guided sampling, fpocket predicted the cryptic site (sensitivity=100%) with AUC=0.90 (compared to AUC=0.89 by CryptoSite).

**MAP p38 kinase**

In the unbound structure of MAP p38 kinase (PDB id: 2ZB1), a helix (residue 253-261) closes down the cryptic site which would clash with a bound ligand (Fig S1) [8]. In the bound structure (PDB id: 2NPQ), this helix moves outward to accommodate the ligand. After NMA-guided sampling, fpocket predicted the cryptic site (sensitivity=87%) with AUC=0.81 (compared to AUC=0.67 by CryptoSite).

**Beta-secretase 1 protease**

In the unbound structure of beta-secretase 1 protease (PDB id: 1W50), the pocket is too open for binding ligand (Fig S1). In the bound structure (PDB id: 3IXJ), a loop (residues 71-74) closes down on the ligand. After NMA-guided sampling, fpocket predicted the cryptic site (sensitivity=94%) with AUC=0.80 (compared to AUC=0.69 by CryptoSite).

**cAMP-dependent protein kinase**

In the unbound structure of cAMP-dependent protein kinase (PDB id: 2GFC), an activation loop (residues 51-56) protrudes into the cryptic site and would clash with a bound ligand (Fig S1). In the bound structure (PDB id: 2JDS), as well as many unbound structures, the loop is farther from the site, leaving it wide open [8]. After NMA-guided sampling, fpocket predicted the cryptic site (sensitivity=96%) with AUC=0.96 (compared to AUC=0.97 by CryptoSite).

**1-deoxy-D-xylulose-5-phosphate reductoisomerase**

In the unbound structure of 1-deoxy-D-xylulose-5-phosphate reductoisomerase (PDB id: 1K5H), a loop (residues 208-215) is highly flexible and open at the cryptic site (Fig S1). In the bound structure (PDB id: 2EGH), it closes down on the bound ligand[8]. After NMA-guided sampling, fpocket predicted the cryptic site (sensitivity=100%) with AUC=0.83 (compared to AUC=0.80 by CryptoSite).

**Kynurenine/alpha-aminoadipate aminotransferase**

In the unbound structure of kynurenine/alpha-aminoadipate aminotransferase (PDB id: 2QLR, a homotetramer), the ligand binding site is between two chains. In the bound structure (PDB id: 3DC1, a homodimer), binding causes a large conformational change of the N-terminal loop (residues 15-33) to accommodate the bound ligand. After NMA-guided sampling (using only chain C of 2QLR), fpocket predicted the cryptic site (sensitivity=100%) with AUC=0.92 (compared to AUC=0.56 by CryptoSite).

**Hepatitis C virus RNA polymerase**

There are two known cryptic sites for inhibitor binding in Hepatitis C virus RNA polymerase. In one unbound structure (PDB id: 3CJ0, see Fig S1), the 1$^{st}$ site is occluded by a loop (residues 22-35). This loop becomes partially disordered in an inhibitor-bound structure (PDB: 2BRL), with slight opening between the thumb and fingers domains. The 2$^{nd}$ site is near the polymerase active site between a loop (residues 364-369) and the core palm domain. Upon binding to an inhibitor in a bound structure (PDB: 3FQK), this loop slightly moves outward to accommodate the inhibitor. After NMA-guided sampling, fpocket predicted the 1$^{st}$ site with AUC=0.59 (sensitivity=50%) and the second site with AUC=0.82 (sensitivity=91%) (compared to AUC=0.73 for both sites by CryptoSite).

As shown by the above examples, our method successfully predicted various types of cryptic sites with an accuracy better than or comparable to CryptoSite. In particular, our method worked well on three false-negative cases of CryptoSite (kynurenine aminotransferase II, HCV RNA polymerase, and PTP1B, see Fig. S9 of ref [7]).

## Examples of unsuccessful cryptic site predictions

Despite general success, there are a few false negative cases in which our sampling method did not help to locate the cryptic sites.

In the first case (Ca-ATPase), ligand binding is associated with large conformational changes (RMSD=13.5 Å) beyond by our NMA-guided sampling (RMSD < 5Å). In the unbound structure (PDB id: 1SU4), the cryptic site is split between two distant domains (Fig S1), which move closer in the bound structure (PDB: 3FGO).

In the second case (Exodeoxyribonuclease I), ligand binding is associated with small local changes (RMSD=1.2Å) instead of global inter-domain motions (Fig S1). The cryptic site is distant from the flexible hinge region between two domains where most high-score residues are located.

The above negative cases expose limitations of NMA-guided sampling in describing some conformational changes required to form or expose cryptic sites. On the other hand, the ‘false-positive’ predictions in those cases may suggest new binding sites for further study.

# Concluding Remarks

We have developed a fast and accurate method for predicting cryptic sites based on a given unbound structure, which can be readily applied at high-throughput to explore the druggable proteome space. This has been done in ref

[7] by CryptoSite (but using a faster and less accurate version of their ML model) on 4421 human proteins, and cryptic sites were predicted in ~3300 (74%) proteins. Given the finding that our method obtained higher FPR and TPR than CryptoSite, we expect an even larger number of crypto sites to be identified by our method. Therefore, small molecules might be used to target significantly more proteins than previously thought to be druggable. However, to exploit those numerous predicted sites, it will be critical yet challenging to distinguish true binding sites from false positives, which requires low-throughput experimental/computational characterization of the predicted sites. We envision future applications of our method focusing on a few high-priority target proteins rather than exploring the proteome space.

In the induced fit model, the presence of ligand is required to induce conformational changes that expose/form a cryptic site. As an alternative model, conformational selection postulates that unbound proteins can transiently sample bound conformations given adequate sampling time. Our method supports the merit of conformational selection, and points to a fast and simple way of sampling by exploring coarse-grained normal modes individually. While such sampling may be too crude to accurately describe the detailed conformational changes, it seems to be sufficient in deforming/exposing potential ligand binding sites so other pocket detection programs can discover them. This method can be applied to difficult cases involving large/flat interfaces between interacting proteins and unknown allosteric sites.

A caveat of the CryptoSite dataset was raised recently that about half of the cases have already formed binding sites in at least one unbound structure so they are no longer qualified as cryptic sites [8]. To address this caveat, we focused on a subset (46 cases) of the CrypticSite set after excluding those cases with only side chain motions and cases that do not qualify for cryptic site based on the more strict criterion (i.e. there exists at least one unbound structure that could accommodate the ligand without any conformational change [8]). The performance of our method on this reduced dataset is virtually unchanged (with AUC~0.8). So our method is applicable to more strictly defined cryptic sites.

Thanks to fast growing computing powers, all-atom MD simulation has been increasingly employed in exploring transient pockets[3]. However, it remains computationally infeasible to scale it up. Another caveat of MD simulation is its focusing on relatively small conformational changes rich in side chain motions, resulting in mostly weak-affinity ligand binding sites unsuitable for drug design [9]. It may be advantageous to incorporate our NMA-guided sampling into the ensemble docking [38] pipeline (e.g. as a preprocessing step prior to MD simulation).

# References


1. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS computational biology 2009;5(12):e1000585.
2. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC bioinformatics 2009;10:168.
3. Kuzmanic A, Bowman GR, Juarez-Jimenez J, Michel J, Gervasio FL. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. Accounts of chemical research 2020;53(3):654-661.
4. Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophysical reviews 2017;9(2):91-102.
5. Nussinov R, Tsai CJ. Allostery in disease and in drug discovery. Cell 2013;153(2):293-305.
6. London N, Raveh B, Schueler-Furman O. Druggable protein-protein interactions--from hot spots to hot segments. Current opinion in chemical biology 2013;17(6):952-959.
7. Cimermancic P, Weinkam P, Rettenmaier TJ, Bichmann L, Keedy DA, Woldeyes RA, Schneidman-Duhovny D, Demerdash ON, Mitchell JC, Wells JA, Fraser JS, Sali A. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. Journal of molecular biology 2016;428(4):709-719.
8. Beglov D, Hall DR, Wakefield AE, Luo L, Allen KN, Kozakov D, Whitty A, Vajda S. Exploring the structural origins of cryptic sites on proteins. Proceedings of the National Academy of Sciences of the United States of America 2018;115(15):E3416-E3425.
9. Vajda S, Beglov D, Wakefield AE, Egbert M, Whitty A. Cryptic binding sites on proteins: definition, detection, and druggability. Current opinion in chemical biology 2018;44:1-8.

10. Weinkam P, Pons J, Sali A. Structure-based model of allostery predicts coupling between distant sites. Proceedings of the National Academy of Sciences of the United States of America 2012;109(13):4875-4880.
11. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophysical journal 2001;80(1):505-515.
12. Tama F, Sanejouand YH. Conformational change of proteins arising from normal mode calculations. Protein Eng 2001;14(1):1-6.
13. Zheng W, Doniach S. A comparative study of motor-protein motions by using a simple elastic-network model. Proc Natl Acad Sci U S A 2003;100(23):13253-13258.
14. Krebs WG, Alexandrov V, Wilson CA, Echols N, Yu H, Gerstein M. Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins 2002;48(4):682-695.
15. Ahmed A, Villinger S, Gohlke H. Large-scale comparison of protein essential dynamics from molecular dynamics simulations and coarse-grained normal mode analyses. Proteins 2010;78(16):3341-3352.
16. Lindahl E, Delarue M. Refinement of docked protein-ligand and protein-DNA structures using low frequency normal mode amplitude optimization. Nucleic acids research 2005;33(14):4496-4506.
17. Venkatraman V, Ritchie DW. Flexible protein docking refinement using pose-dependent normal mode analysis. Proteins 2012;80(9):2262-2274.
18. May A, Zacharias M. Protein-ligand docking accounting for receptor side chain and global flexibility in normal modes: evaluation on kinase inhibitor cross docking. Journal of medicinal chemistry 2008;51(12):3499-3506.
19. Cavasotto CN. Normal mode-based approaches in receptor ensemble docking. Methods in molecular biology 2012;819:157-168.
20. Cavasotto CN, Kovacs JA, Abagyan RA. Representing receptor flexibility in ligand docking through relevant normal modes. Journal of the American Chemical Society 2005;127(26):9632-9640.
21. Petrone P, Pande VS. Can conformational change be described by only a few normal modes? Biophysical journal 2006;90(5):1583-1593.
22. Dietzen M, Zotenko E, Hildebrandt A, Lengauer T. On the applicability of elastic network normal modes in small-molecule docking. Journal of chemical information and modeling 2012;52(3):844-856.
23. Hinsen K, Petrescu A-J, Dellerue S, Bellissent-Funel M-C, Kneller GR. Harmonicity in slow protein dynamics. Chemical Physics 2000;261(1–2):25-37.
24. Zheng W, Auerbach A. Decrypting the sequence of structural events during the gating transition of pentameric ligand-gated ion channels based on an interpolated elastic network model. PLoS computational biology 2011;7(1):e1001046.
25. Zheng W, Brooks BR, Doniach S, Thirumalai D. Network of dynamically important residues in the open/closed transition in polymerases is strongly conserved. Structure 2005;13(4):565-577.
26. Zheng W, Brooks BR, Thirumalai D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci U S A 2006;103(20):7664-7669.
27. Zheng W, Tekpinar M. Large-scale evaluation of dynamically important residues in proteins predicted by the perturbation analysis of a coarse-grained elastic model. BMC structural biology 2009;9:45.
28. Zheng W. Probing the structural dynamics of the SNARE recycling machine based on coarse-grained modeling. Proteins 2016.
29. Zheng W. Coarse-grained modeling of the structural states and transition underlying the powerstroke of dynein motor domain. The Journal of chemical physics 2012;136(15):155103.
30. Badaczewska-Dawid AE, Kolinski A, Kmiecik S. Computational reconstruction of atomistic protein structures from coarse-grained models. Computational and structural biotechnology journal 2020;18:162-176.
31. Greener JG, Sternberg MJ. AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. BMC bioinformatics 2015;16:335.
32. Zheng W, Wen H, Iacobucci GJ, Popescu GK. Probing the Structural Dynamics of the NMDA Receptor Activation by Coarse-Grained Modeling. Biophysical journal 2017;112(12):2589-2601.
33. Zheng W. Probing the structural dynamics of the CRISPR-Cas9 RNA-guided DNA-cleavage system by coarse-grained modeling. Proteins 2017;85(2):342-353.

34. Zheng W. Toward decrypting the allosteric mechanism of the ryanodine receptor based on coarse-grained structural and dynamic modeling. Proteins 2015;83(12):2307-2318.
35. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, Ben-Tal N. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic acids research 2016;44(W1):W344-350.
36. Bowman GR, Geissler PL. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. Proceedings of the National Academy of Sciences of the United States of America 2012;109(29):11681-11686.
37. Porter JR, Moeder KE, Sibbald CA, Zimmerman MI, Hart KM, Greenberg MJ, Bowman GR. Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling. Biophysical journal 2019;116(5):818-830.
38. Amaro RE, Baudry J, Chodera J, Demir O, McCammon JA, Miao Y, Smith JC. Ensemble Docking in Drug Discovery. Biophysical journal 2018;114(10):2271-2278.

Acknowledgement

This study was supported by a grant from American Heart Association (#17GRNT33690009). The simulations were conducted using the supercomputing cluster of the Center for Computational Research at the University at Buffalo.

# Figure Legends

**Figure 1.** Analysis of unbound-to-bound conformational changes in the training set: (a) Histogram of RMSD of C$\alpha$ atoms from the unbound to the bound structures dominated by small backbone motions (< 5 Å). (b) Histogram of cumulative overlap (CO) which gives the fraction of observed conformational changes as described by the lowest 30 modes calculated for the unbound structures.

**Figure 2**. Optimization of NMA-guided sampling assessed by the AUC of ROC for the average pocket scores from fpocket (blue) and concavity (red). (a) AUC peaks when ~ 30 modes are used in the single-mode exploration; (b) AUC increase as the maximal RMSD increases in the single-mode exploration; (c) AUC saturates after more than 200 conformations are sampled by the multi-modes combination.

**Figure 3**. Assessment of individual predictive scores on the training set. (a). ROC curves for pocket scores from fpocket (blue) and concavity (red), residue conservation scores from CONSURF server (green), dynamical importance (cyan) and B factors (magenta) calculated from NMA of ENM. (b). Histograms of AUC of ROC, sensitivity, and specificity for pocket scores from fpocket (blue) and concavity (red), which were calculated for 84 individual cases of the training set. The sensitivity/specificity values were determined at a score threshold of 0.05.

**Figure 4**. Assessment of cryptic site predictions on the test set. (a). ROC curves from three machine learning protocols: linear regression (red), random forest (green), and neural net (blue), in comparison with the pocket score from fpocket (black). (b) Examples from the test set (TEM1 β-lactamase, Exportin 1, and Interleukin-2), where the unbound structure is colored by average fpocket score (blue for low score and red for high score), the cryptic-site residues are shown as Cα spheres and boxed.

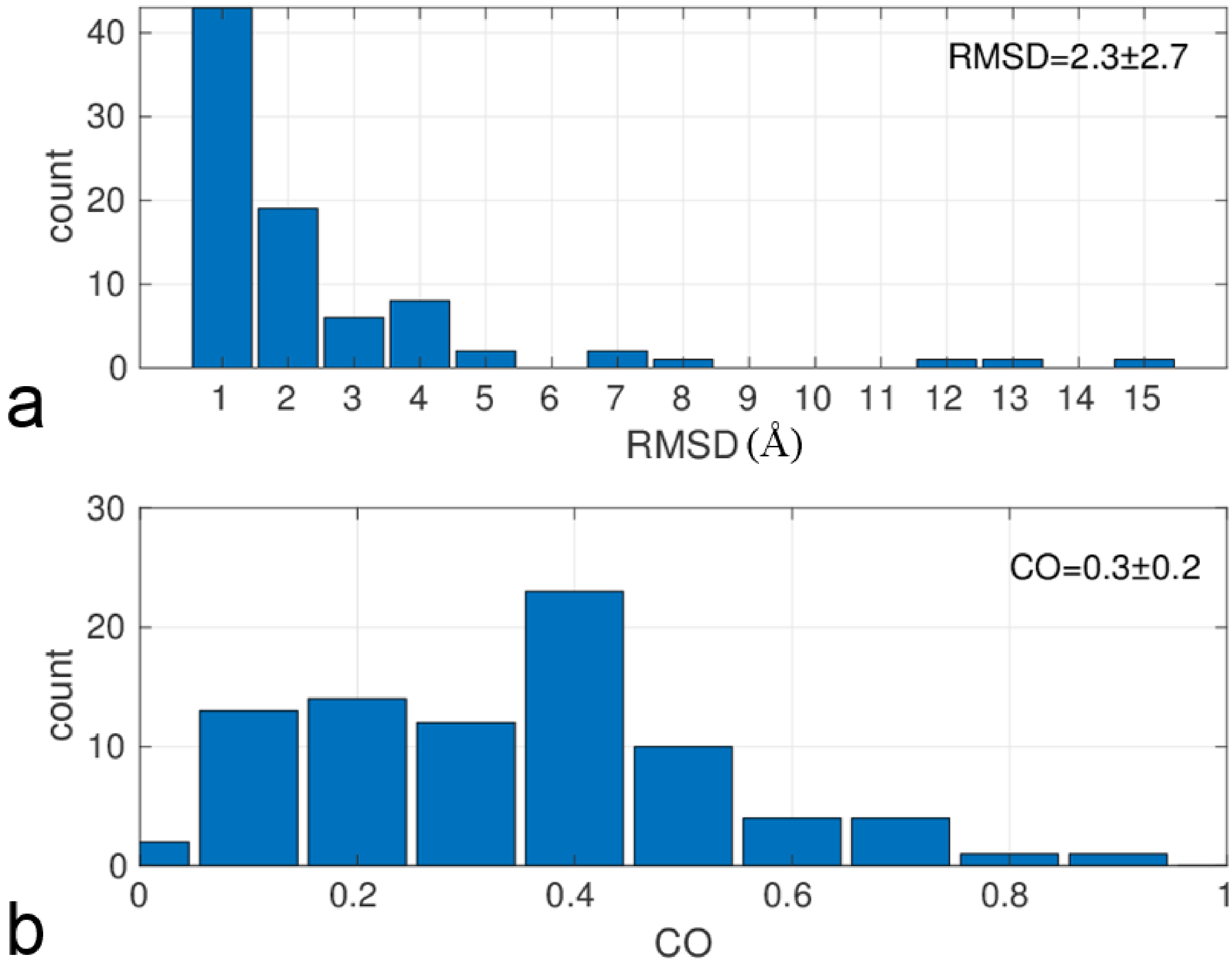
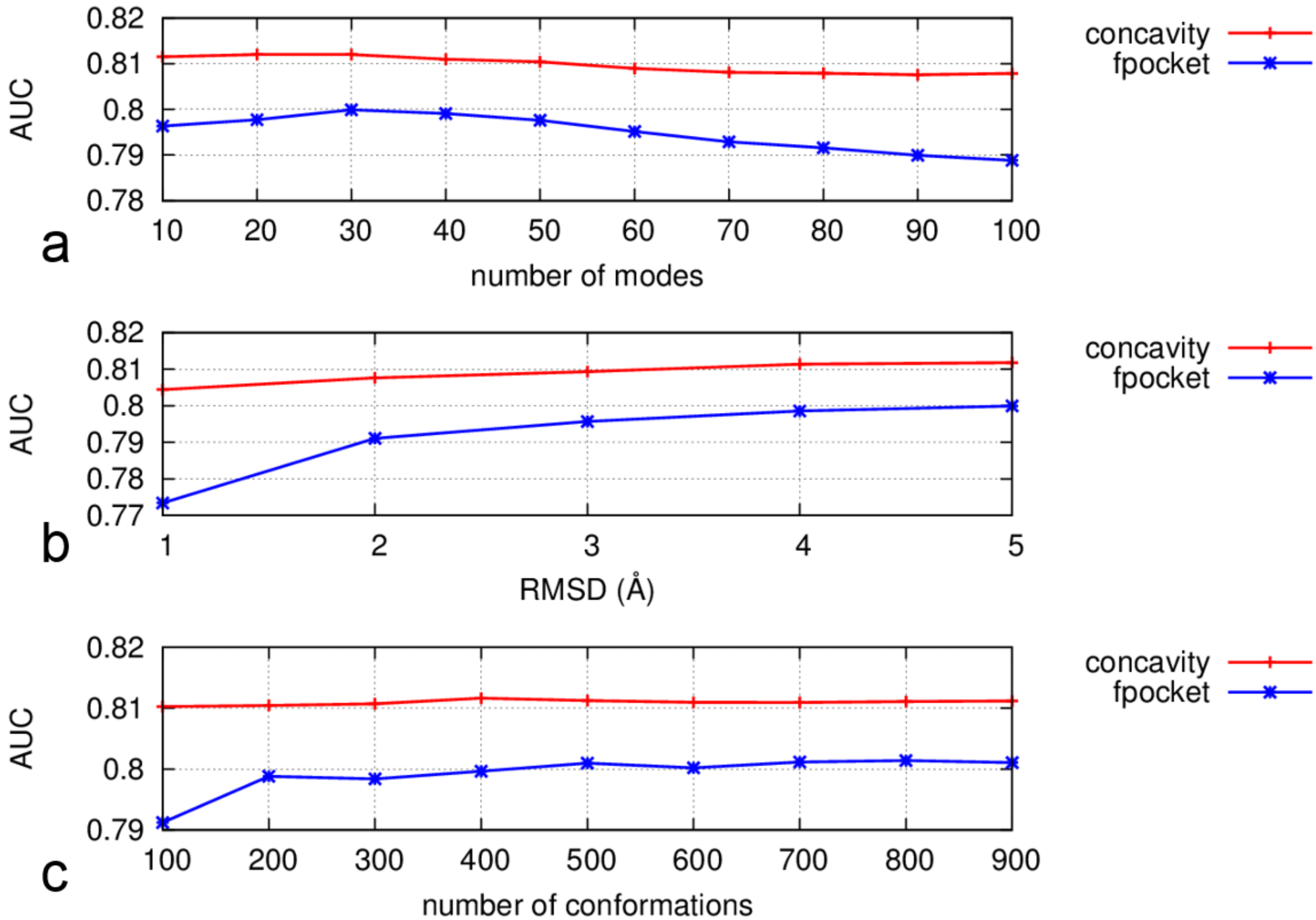



Figure 1



concavity
fpocket
AUC
0.82
0.81
0.8
0.79
0.78
10
20
30
40
50
60
70
80
90
100
a
number of modes
0.77
1
2
3
4
5
b
RMSD (Å)
100
200
300
400
500
600
700
800
900
c
number of conformations


**Figure 2**

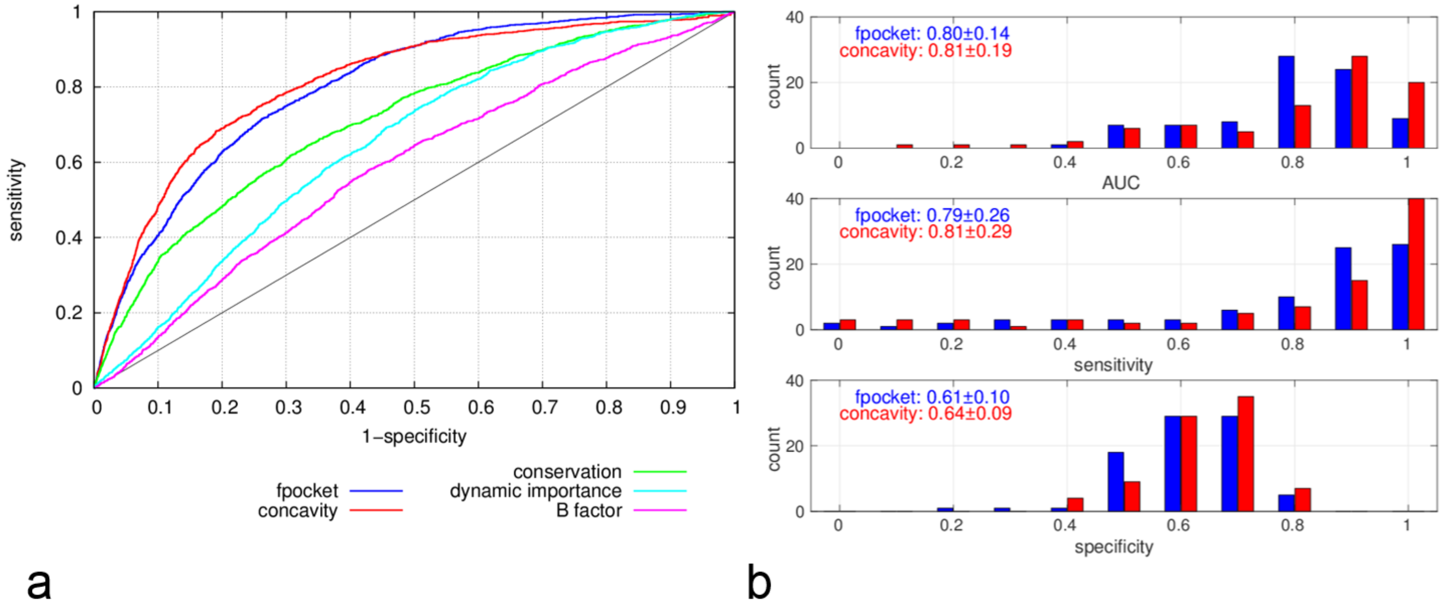

sensitivity
1-specificity
fpocket
concavity
conservation
dynamic importance
B factor
count
AUC
fpocket: 0.80±0.14
concavity: 0.81±0.19
sensitivity
fpocket: 0.79±0.26
concavity: 0.81±0.29
specificity
fpocket: 0.61±0.10
concavity: 0.64±0.09
a
b

**Figure 3**

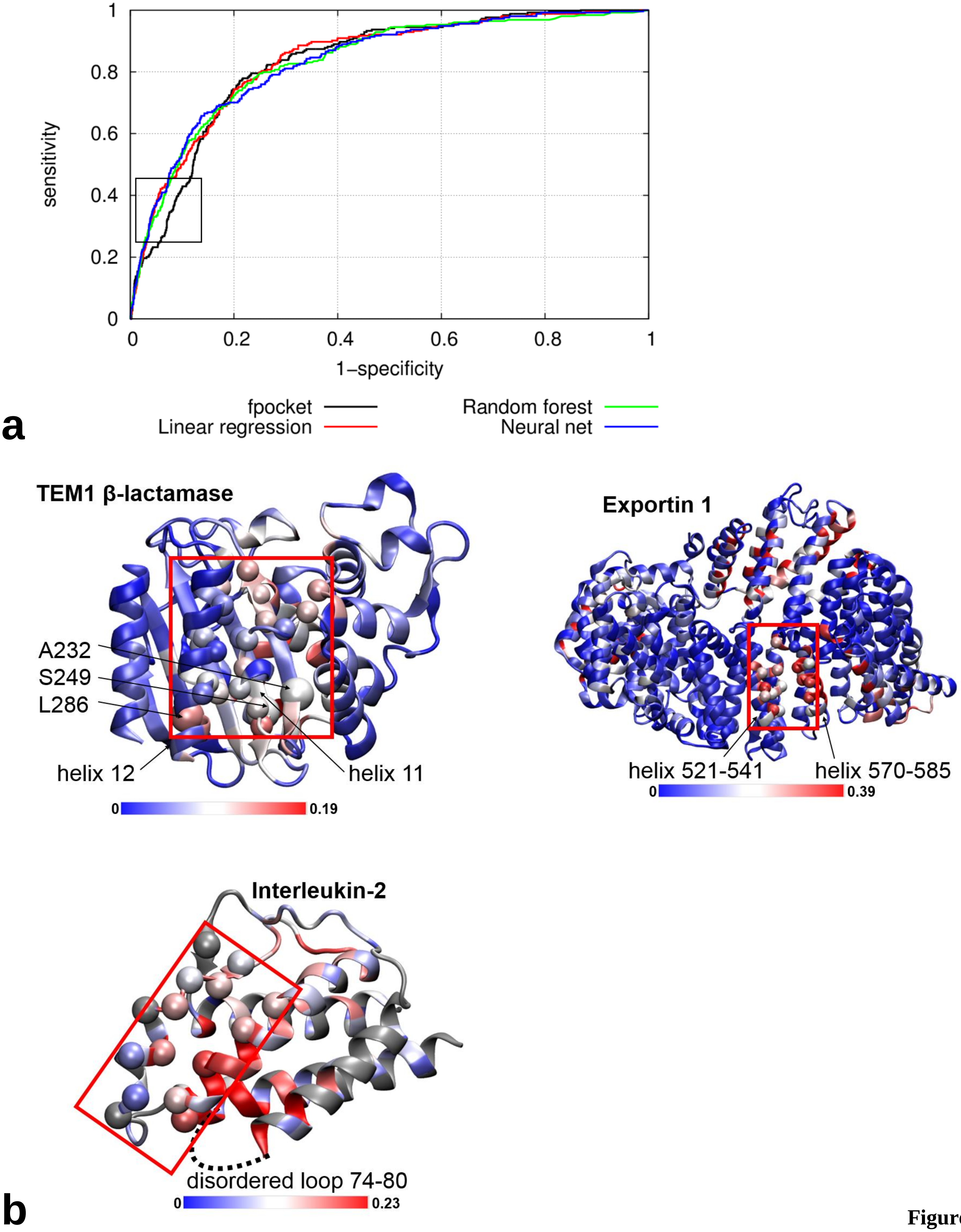

sensitivity
1−specificity
fpocket
Linear regression
Random forest
Neural net
a
TEM1 β-lactamase
A232
S249
L286
helix 12
helix 11
0
0.19
Exportin 1
helix 521-541
helix 570-585
0
0.39
Interleukin-2
disordered loop 74-80
0
0.23
b

**Figure 4**