SITEFINDER: Protein Functional Site Detection Using Residue-Residue Contact Maps

Shankara Anand
Stanford School of Medicine
Biomedical Informatics Program
sanand94@stanford.edu

Abstract

Protein-ligand interactions are capable of regulating complex, biological phenomena and are often targeted in therapeutic interventions. 3D structure of protein active sites is crucial in determining chemical interaction with foreign ligands and may be used to screen proteins for new protein-ligand interactions. With the growing repository of protein:ligand co-crystallized structures in the Protein Data Bank, structural information of protein backbone coordinates is readily available for binding site prediction. Protein structure may be featurized using low-cost, residue-residue distance matrices, or protein contact maps, which act as a static image of proteins. I propose SITEFINDER, a model that uses protein contact maps for functional site detection of amino acid residues involved in drug binding. SITEFINDER yields promising results given that it approaches this highly specific, biochemical task without sequence data or sidechain geometries commonly used in the field but instead with backbone geometries alone.

1. Introduction

Biological systems are often regulated through protein-ligand binding interactions to inhibit or excite metabolic pathways. 3D protein structure is crucial in establishing specificity of binding, or functional, sites, with amino acid side-chain geometries relevant at angstrom scales. Detection of functional sites is important in understanding protein function, as ligand binding may induce full-protein conformational change [1] or alter enzymatic activity. Over seventeen computational functional site prediction tools currently exist, most of which rely on protein sequence, structural representations, or both. Prediction tools based on protein sequence fail to capture nuances in binding pocket geometries and fail to generalize beyond homology families, while structural representations are constrained by high dimensional feature spaces. [2].

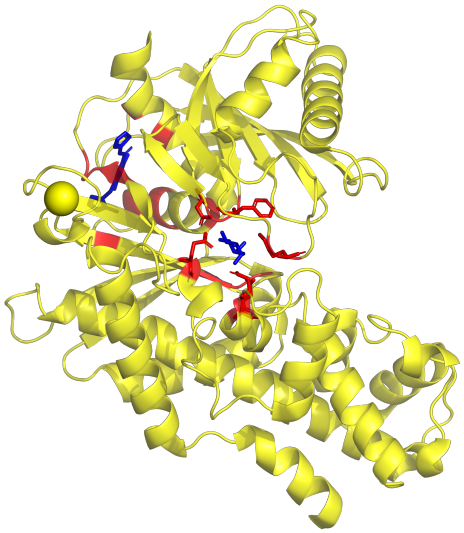


Figure 1: PYMOL Visualization of a glucokinase (PDB: 1v4s). *Yellow* indicates a "cartoon" representation of the protein with binding site residues, annotated in the **.pdb** header, colored *red*. Its co-crystallized ligand, glucose, is shown in *blue*.

Recently, deep learning methods have been applied in structural biology at the amino acid scale and proven capable of capturing complex phenomena such as a residue's micro-environment [3] and protein-ligand cavity prediction [4]. I propose SITEFINDER, a deep learning approach to functional site detection that uses a naive, backbone feature representation as input: residue-residue contact maps. These contact maps are treated as images in a semantic segmentation problem, utilizing deep, convolutional methods to downsample and upsample the image akin to Badrinarayanan *et al's* SegNet [5]. SITEFINDER outputs probabilities of each amino acid's involvement in func-

tional site despite this sequence agnostic representation. In using only protein backbone information in this prediction task, this method will prove less computationally intensive then current approaches and have potential for active site discovery at residue-level resolution in existing proteins with appropriate backbone geometries.

2. Related Work

Computational methods for function site prediction in proteins are generally classified into two domains: sequence-based approaches and structure-based approaches.

2.1. Sequence Based Approaches

Sequence-based approaches rely purely on a protein's amino acid sequence to approach this task. One key advantage of these methods is that protein sequence information is easier to obtain empirically through techniques such as Edman Degradation [6] and does not rely on crystallography to capture full 3D geometry. A common technique for leveraging sequence is homology modeling. Proteins with known biological function tend to exhibit similar biochemical function. This conservation drops rapidly for proteins sharing less than 35% to 40% sequence identity [7]. Thus, sequence-based approaches generally rely on identifying a protein-ligand target, homologous sequence search to find nearest sequence hits, a multiple sequence alignment to line up these sequences, and finally functional site prediction by finding regions on homologous proteins that line up with the input query's.

La *et al* demonstrated in 2004 that phylogenic motifs (PM), or sequence regions conserved across protein families, are capable of capturing important functional site information. These methods were eventually developed to the tool MINER [8]. However, such methods are limited both by their reliance on curated protein-ligand binding information for a proteins in a given family and by the selection of homologous sequences to the query protein that generalizes to different protein families. Finally, sequence methods also struggle in performance when a binding site is non-local or non-contiguous in sequence.

2.2. Structure Based Approaches

Structure based approaches rely on solved protein structures deposited in publicly available databases *see Dataset*. These are broadly characterized as alignment-based approaches, geometric approaches, and energetic approaches. Structural approaches generally leverage sequence information in some manner, notably using side chain geometries in the computational pipeline.

Structural alignment models. These approaches leverage similar techniques to those used for sequence-based models, but use structural data instead. In 1995, a database of protein structural classification named SCOP was published. Analysis of this data showed that specific protein folds tend to to bind substrates at similar locations [9]. This motivated global structural alignment approaches to functional site detection, such as FINDSITE in 2008 [10]. FINDSITE aims to detect and correctly rank ligand-binding regions in weakly homologous protein models ($< 35\%$). It identifies template structures from a query protein using a threading algorithm, superimposes these templates to a predicted target protein structure, and uses the clustered centers of mass of ligands bound to identify putative binding sites. This method uniquely incorporates structure in a way that does not require high sequence identity, but structural alignment modeling, like sequence-based methods, has drawn critique due to its reliance on general homology trends with known binding sites.

Geometry-based methods. Geometric methods approach functional site prediction by searching for binding pockets or cavities in 3D structure given crystallographic information. POCKET was one of the first algorithms that incorporated a geometric approach to functional site detection. It maps a protein to a 3D grid along with solvent levels for each point, with regions that exceeded a level of protein-solvent-protein interactions as pocket binding regimes [11]. SURFNET is another algorithm that identifies binding sites by placing spheres between all pairs of atoms [12]. Each sphere is reduced in size if any other toms intersect it until it intersects with no other atoms or its radius drops below a given size. Thus, clefts on the surface of a protein are filled with these spheres, which may be used as a feature for binding site prediction , achieving up to 52% accuracy. These methods are computationally intensive and require assumptions of the interaction's chemical mechanism.

Energetic-based methods. Energetic methods approach this task using computed interaction energies between specific residues and ligands to inform site prediction. Q-SiteFinder (2008) approaches this problem by harnessing chemical properties implicated in these interactions [13]. This model uses molecular modeling to coat a protein with a layer of methyl (-CH3) probes, model van der Waals interaction energies, or innate electronic forces at an atomistic level, to elucidate drug-binding sites. This method also is subject to computational constraints, as they claim 80% accuracy, but only on 35 structures. FTSite [14] is another energetics approach based on the observation that many ligand binding sites also bind small organic molecules of various shapes and polarity at their functional site as noted by Ha-

jduk *et al* in 2005 [15]. This approach, uniquely, does not rely on evolutionary (used in homology modeling) or statistical information, and places multiple molecular probes in a molecular simulation suite to optimize a free energy function. Hajduk *et al* claim 94% accuracy, but are limited to unbound protein structures and use the same small test set also used by Q-SiteFinder (n=35).

2.3. DeepSite

DeepSite, released in 2017, applies 3D convolutional neural networks to this long-standing problem [4]. DeepSite trains on the scPDB v.2013 dataset, a database of high-quality, non-redundant druggable binding sites extracted from the Protein Data Bank [16]. A set of 8 chemical properties are used for channels and the protein coordinates are converted to a grid of 1x1x1 Angstrom voxels. DeepSite performs at the state of the art level, predicting occupancy of binding-cavity sites at 50% accuracy for predictions within 4 Angstroms up to 82 % within 10 Angstroms. This approach is state-of-the art in its use of a large, curated dataset and full-protein, 3D convolutional approach to identify protein binding pockets.

3. Dataset

3.1. Data Source

The Protein Data Bank (PDB) is a public repository of crystallized protein structure coordinates including ligands bound, functional data (i.e. enzyme class), and protein metadata from a peer-reviewed depositing source. [17] Each **.pdb** file contains relevant metadata along with coordinate information for each atom (*see figure 2*). I opted to scrape data directly from the PDB rather than use scPDB (*see DeepSite*) to curate residue-level information and create a generalizable, scalable data curation pipeline for future work with SITEFINDER.

3.2. Data Scraping

.pdb files were scraped for coordinate information of residue α-carbons and then converted to distance-distance matrices, or contact maps (*see Figure 2*). Labels were generated for each protein through text parsing of the **.pdb** files to find residues interacting with a bound ligand. Each residue has a default label of 0, and incremented by one if the metadata outlined its interaction with a ligand. Importantly, the Protein Data Bank considers a wide range of molecules as "ligands" which are not relevant to the problem statement, such as ions, which required further filtering of proteins considered.

Because there is a vast range of protein lengths, padding and centering was also required at this step with an upper length threshold of 512 used (*see Figure 2*). If proteins were longer than 512 but less than or equal to 520 residues, they were truncated at each respective end to fit into a 512x512 image. Residues greater than 520 residues in size were not considered. I initially extracted 104,733 proteins the PDB designates as proteins with a ligand deposited with its crystal structure. Following the filtering steps above, I was left with 47,968 structures.

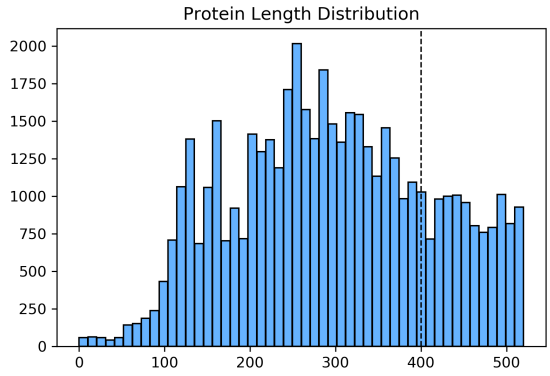
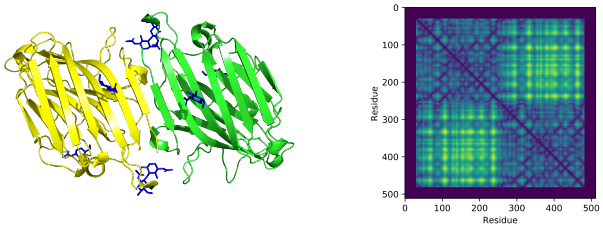


Figure 2: Protein length distribution of ligand-bound proteins from PDB following filtering.



(a) PYMOL Structure

(b) α-carbon Contact Map

Figure 3: PYMOL Visualization (a) and contact map (b) of Arcelin-1, a glycoprotein (PDB ID: 1avb). The PYMOL visualization shows protein chains in *yellow* and *green*, with the ligands bound in *cyan*. The contact map is a distance matrix with low values indicating lower distances between residue α-carbons. The map displayed is post-processing: padding and centering have both been applied.

Future directions in data curation would include splitting proteins greater than the 520 residue cutoff into separate chains to be considered separately for this model. This may act as a form of data augmentation provided each segment contained at least one ligand-binding site.

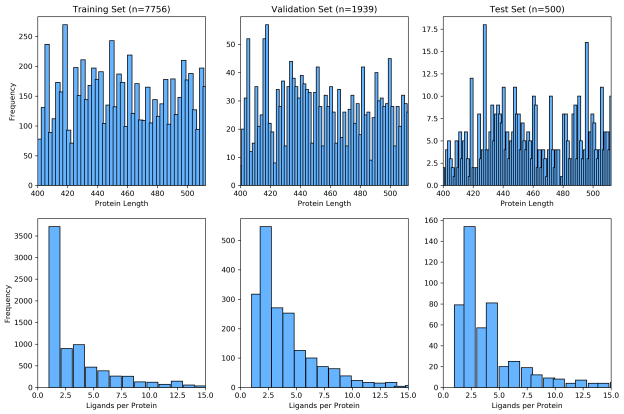


Figure 4: Protein length and ligand bound statistics for training, testing, and validations ets used in SITEFINDER.

3.3. Data Smoothing

An important facet of this hand-curated dataset is its sharpness. The novelty of SITEFINDER is its residue-level approach to drug-binding prediction. However, while there may exist biological instances where an interaction may be limited to a single amino acid side chain, generally the region surrounding a residue-ligand interaction may contribute to this process mechanistically, whether in providing electron density or steric, crowding, effects. Additionally, this initial data curation results in a strong class imbalance of 96% to 4% non-functional site to functional site involvement per residue. This motivated label "smoothing:" a Gaussian smoothing kernel was applied to the 1x512 label array that had 0's for non-ligand binding proteins and the number of times an amino acid residue was involved in a drug binding event.

Following smoothing, residues were then re-binned to 0's or 1's if they reached above a certain threshold probability (0.25) from the smoothing kernel. Following smoothing, we were left with a 92% to 8% split. Because of the importance of site-specific information per protein in this classification task, upsampling data from one class was not an option. Thus, class frequencies were computed before training the model and passed to the loss function for weighting (*see Methods*).

4. Methods

4.1. Architecture

Following data processing, SITEFINDER is given a 512x512 image of each protein and a binarized array of labels (0,1) for each residue. These images are fed through three rectangular convolutional layers downsampled to a 2x64x128 size. At each layer, a 2D batch normalization is applied followed by a parametric ReLU for a nonlinear activation layer. This low resolution feature space of the protein is then upsampled using convolutional transpose layers, 3 layers, up to a 2x512x512. Motivation for using this encoder:decoder format for the first part of this network architecture comes from Badrinarayanan *et al's* SegNet [5]. The novelty of Badrinarayanan *et al's* approach lies in the ability to map low resolution features from the encoder network back to full size input feature space where each channel represents a different segmentation. Opting to use rectangular convolutions in encoding these lower level features is inspired by the data structure itself. Each row or column of each input corresponds to one amino acid and its neighboring interactions with each other amino acid.

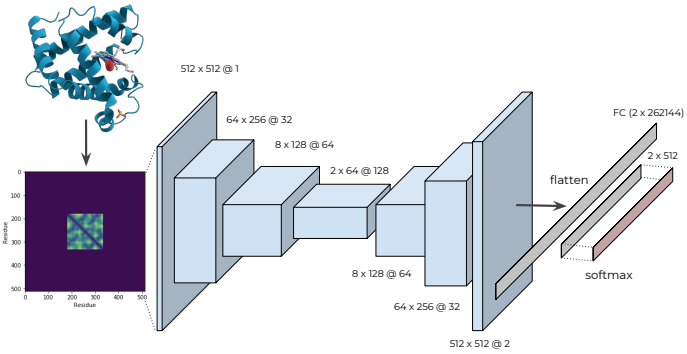


Figure 5: SITEFINDER Network Architecture

By taking rectangular convolutions, I bias the downsampling to force some notion of locality in a given direction. SITEFINDER uses (2x8) convolutions, with the number 2 chosen to incorporate information from each amino acid row and an adjacet residue along a chain. The decision to choose row or column is arbitrary given the matrix is symmetric. Motivation for using a 2D convolution rather than a 1D convolution per column/row stems from a need to capture distance-distance information of a residues neighboring amino acids. If a residue in a binding pocket has neighbors that are also surrounded by backbone motifs, I assume there might be a relationship to ligand binding I hope to capture using this data representation. Following encoding and decoding, the network flattens the 2x512x512 image to a 2x262,144 image which is then fed to a fully connected layer with 512 neurons and another layer of dropout is applied for regularization. This Nx2x512 image is then fed to a log-Softmax function for the final output.

4.2. Loss

SITEFINDER uses a 2-D Cross Entropy Loss function during training, which subjects the softmax function from

the model to a negative log likelihood loss. This is averaged over each amino acid residue (r = 512) and each sample (i) per mini-batch (size N):

$$L^r_{i,c^*} = \frac{exp(y^r_{i,c^*})}{\sum_C exp(y^r_{i,c})}$$

$$L = \frac{1}{N}\sum_i \frac{1}{512}\sum_r -log\left(L_i^{r,c^*}\right)\mathbf{1}[y^r_{i,c^*} = target^r_{i,c^*}]$$

The softmax function computes the probability that each amino acid in the protein is implicated in a drug binding site - two possible classes. These values are passed to the negative liklihood loss function to minimize the log probabilities in our distribution throughout training. A loss value is computed per protein's (i) amino acid (r). From here, I average over all amino acids and proteins per mini-batch. An important facet of the loss function for this task was the use of class weights. In the data loading process, frequencies for each class (drug binding or not) are computed *(see Data Smoothing)*. These weights then contribute to the loss function when computing log probabilities. Intuitively, this tells the classifier to up-weight every correct classified instance of a drug binding site to combat the heavy class imbalance in this dataset.

4.3. Optimizer

SITEFINDER uses ADAM optimization [18]. ADAM adapts parameter learning rates using both the first moment and second moments of the gradients taken at each time step. ADAM does this by calculating an exponential moving average of the gradient and square gradient to control rate decay. Initial tests used Stochastic Gradient Descent with Nesterov momentum, but better results were found using ADAM.

5. Results

5.1. Metrics

This task approaches site detection at the residue level rather than by pockets, making direct comparison to current methods difficult (*see Related Work*). Additionally, manually curation directly from the PDB to find specific residues of interest embedded in metadata result in using generic classification metrics to evaluate SITEFINDER's performance. I use both quantitative metrics and qualitative metrics to evaluate the performance of this algorithm.

Quantitative. Standard classification metrics used for this task are precision/recall and F1 score. Precision defined for this task is the ratio of amino acids correctly classified as a drug binding site to the number of amino acids SITEFINDER believes are drug binding sites. Recall, is the ratio of amino acids correctly classified to all drug binding sites in the protein. F1 score is a weighted contribution score of both Precision and Recall.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

For this task, I am most concerned with optimizing Recall. I am aiming to create a model that can recapitulate all amino acids charted in the PDB to have drug-site interaction. Precision is also as important, but false positives for this task may encompass neighboring amino acids to the drug pocket site that were not highlighted in the stringent metadata available.

Qualitative. Qualitative analysis of SITEFINDER was done using PYMOL, an open source platform for molecular visualization [19]. PYMOL allows python scripting to interface with the visualization software. A script was written to take outputted results from SITEFINDER and visualize them in PYMOL.

5.2. Experiments

SITEFINDER was trained for 200 epochs using a learning rate of 0.001. Every 55 epochs, the learning rate was decayed by a factor of 0.8. Additionally, the mini-batch used was 250 protein contact maps. This was limited by computational resources as 2 K80 GPU's were used for training. Random grid search was used to tune hyperparameters. Additionally, the amount of dropout used at each layer was also optimized by trial and error following overfitting to a tenth of the dataset (*see Architecture*). SITEFINDER's most recent results are as follow.

SITEFINDER is tested on a set of 500 proteins in the same 400 residue threshold range not yet seen before to yield the following results:

Set	n	Precision	Recall	F1	Acc
Train	9695	0.8291	0.7204	0.77	96.3%
Test	500	0.5451	0.3772	0.45	91.5%

To qualitatively assess SITEFINDER's performance, scripting was conducted in PYMOL and ground truth active site labels were compared with SITEFINDER's predictions. Figure 7 shows results for 3 proteins, 1V4S (training set) and 4JGE and 5EYI (testing set) using visualization software built.

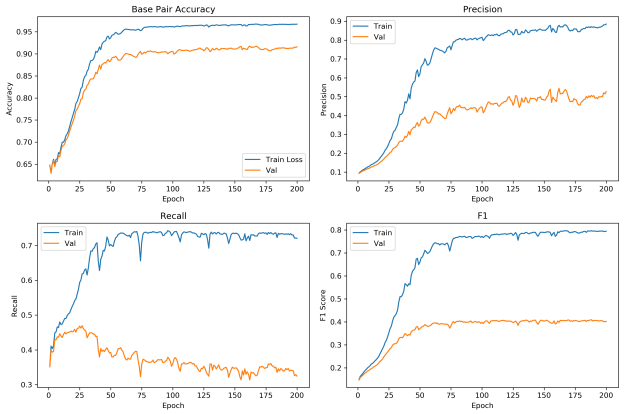


Figure 6: SITEFINDER metrics. SITEFINDER achieves above 90%

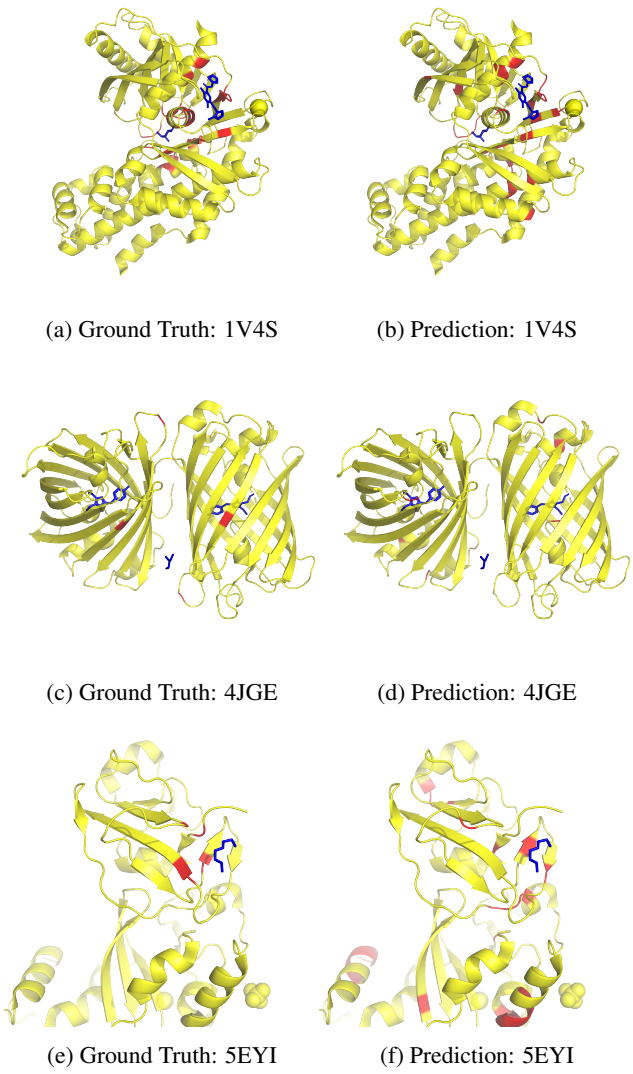
5.3. Discussion

SITEFINDER achieves 91.5% base-pair accuracy for test set of 500 proteins. However, following label smoothing, the dataset still has a class imbalance with only 8% of amino acids implicated in drug binding sites across all proteins. Thus, precision and recall are more appropriate metrics to understand how SITEFINDER is performing at this stage. I find that SITEFINDER is prone to overfitting to the training set, achieving a precision of 54.5 % and Recall of 37.8 %. This indicates that roughly half the amino acids SITEFINDER currently classifies as drug binding are truly involved in drug binding and it is only able to highlight 37.8% out of all drug binding sites. These results fall short in capturing base-pair level relevance of binding site pocket, but are promising initial steps towards the completion of this task.

The challenge of this task lies in using a minimalist feature representation of proteins to predict the possibility of a specific, biochemical event. Values achieved for precision and recall are promising in that a more extensive hyperparameter search and adjustments to the model, such as increasing dropout rates at different layers, may boost SITEFINDER's performance in future work. Notably, qualitative inspection of SITEFINDER's results on proteins selected from the testing set also show promise. They seemingly capture similar regions to binding sites near the ligands of interest with no ligand information input into the model.

6. Conclusion

The novelty of SITEFINDER is twofold. Approaching drug pocket site detection at residue-level scales is an approach not done before (*see Related Work*). Often,



(a) Ground Truth: 1V4S (b) Prediction: 1V4S

(c) Ground Truth: 4JGE (d) Prediction: 4JGE

(e) Ground Truth: 5EYI (f) Prediction: 5EYI

Figure 7: SITEFINDER Visualized using PYMOL.

techniques rely on protein microenvironments, local geometries, or sequence to classify functional site regions. SITEFINDER trains on a manually curated dataset that extracts residue-level information from the PDB, outputting how likely an amino acid is involved in an interaction given no sequence information. It takes only backbone geometries of the protein and uses deep vision architectures to approach this task. In exploring this problem, hyperparameter tuning and overfitting were quite challenging. A number of architectures were experimented with, but approaching the problem using semantic segmentation *a la* SegNet produced promising results. Notably, overfitting to a training step is a promising first step given that the feature representation used from this task is so bare boned, indicating there may be latent information within protein backbones that are capable of guiding functional site detection and design in

research endeavors to come.

7. Future Work

Architecture redesign of the model would be the first step. Aiming to achieve higher precision and recall are guiding goals for this task. Additionally, attempting to use SITEFINDER on pre-curated database such as scPDB (*see DeepSite)* would be an important sanity check. Scraping residue information from this database is possible and may prove to be a more effective dataset to train with for this task.

If future iterations of SITEFINDER are successful at functional site detection, a next step would be to explore site design. Because this model is only taking in residue backbone information, and not side-chains, it forgoes additional residue information for the opportunity to learn ligand-binding by backbone metrics alone. Thus, future work may include seeing if this model may be applied to proteins with no known ligand binding sites to discover new ones. If this model proves successful, incorporation into current protein modeling software, such as ROSETTA Remodel [20], may allow us to design new sidechains on backbone positions found.

Finally, if this model works well, the final and most substantial metric for functional site detection is empirical testing. One may conduct binding assays of a ligand to an engineered protein with no previous known interaction to see if the binding site ”generated” does, in fact, result in a binding event.

8. Contributions

This project was completed by Shankara Anand. Scraping the database, model testing, and PYMOL visualization efforts were all completed for this project. Raphel Eugichi (*see Acknowledgments)* provides two functions for initial contact map generation that are used in SITEFINDER. All work was done on Google Cloud Computing.

9. Acknowledgements

This work was inspired by many of the talented scientists in Dr. Possu Huang’s Lab. Raphael Eugichi was a great asset in discussing the nuances of using protein contact maps for a new problem following his success with domain segmentation. Without his help and vote of confidence, I could not have started a project so daunting. Namrata Anand was the first scientist in Dr. Huang’s lab to use protein contact maps in a novel way. Her work with GANs for in-painting of proteins was an inspiration for using this feature representaton to approach functional site detection. [21].

References

[1] Omar NA Demerdash, Michael D Daily, and Julie C Mitchell. Structure-based predictive models for allosteric hot spots. *PLoS computational biology*, 5(10):e1000531, 2009.

[2] B KC Dukka. Structure-based methods for computational protein functional site prediction. *Computational and structural biotechnology journal*, 8(11):e201308005, 2013.

[3] Wen Torng and Russ B. Altman. 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinformatics*, 18(1):302, dec 2017.

[4] J Jiménez, S Doerr, G Martínez-Rosell, AS Rose, and G De Fabritiis. Deepsite: protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics*, 33(19):3036–3042, 2017.

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[6] Richard A Laursen. Solid-phase edman degradation. *The FEBS Journal*, 20(1):89–102, 1971.

[7] Annabel E Todd, Christine A Orengo, and Janet M Thornton. Evolution of function in protein superfamilies, from a structural perspective1. *Journal of molecular biology*, 307(4):1113–1143, 2001.

[8] David La, Brian Sutch, and Dennis R Livesay. Predicting protein functional sites with phylogenetic motifs. *Proteins: Structure, Function, and Bioinformatics*, 58(2):309–320, 2005.

[9] Alexey G Murzin, Steven E Brenner, Tim Hubbard, and Cyrus Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540, 1995.

[10] Michal Brylinski and Jeffrey Skolnick. A threading-based method (findsite) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of sciences*, 105(1):129–134, 2008.

[11] David G Levitt and Leonard J Banaszak. Pocket: a computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of molecular graphics*, 10(4):229–234, 1992.

[12] Roman A Laskowski. Surfnet: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of molecular graphics*, 13(5):323–330, 1995.

[13] Mizuki Morita, Shugo Nakamura, and Kentaro Shimizu. Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. *Proteins: Structure, Function, and Bioinformatics*, 73(2):468–479, 2008.

[14] Chi-Ho Ngan, David R Hall, Brandon Zerbe, Laurie E Grove, Dima Kozakov, and Sandor Vajda. Ftsite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics*, 28(2):286–287, 2011.

[15] Philip J Hajduk, Jeffrey R Huth, and Stephen W Fesik. Druggability indices for protein targets derived from nmr-based screening data. *Journal of medicinal chemistry*, 48(7):2518–2525, 2005.

[16] Jérémy Desaphy, Guillaume Bret, Didier Rognan, and Esther Kellenberger. sc-pdb: a 3d-database of ligandable binding sites10 years on. *Nucleic acids research*, 43(D1):D399–D404, 2014.

[17] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank, 1999–. In *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, pages 675–684. Springer, 2006.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

[20] Po-Ssu Huang, Yih-En Andrew Ban, Florian Richter, Ingemar Andre, Robert Vernon, William R. Schief, and David Baker. RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLoS ONE*, 6(8):e24109, aug 2011.

[21] Namrata Anand and Possu Huang. Generative modeling for protein structures. 2018.