Enhancing the Performance of GNN and Utilizing 3D Instance Segmentation for Ligand Binding Site Prediction

Esha Rajesh Gavali

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2024

Committee:

Dong Si, Chair

Wooyoung Kim, Member

Jie Hou, Member

Renzhi Cao, Member

Program Authorized to Offer Degree:

Computer Science and Software Engineering

University of Washington

Abstract

Enhancing the Performance of GNN and Utilizing 3D Instance Segmentation for Ligand Binding Site Prediction

Esha Rajesh Gavali

Chair of the Supervisory Committee:
Associate Professor Dong Si, Ph.D.
Department of Computing and Software Systems

This study addresses the challenge of accurately predicting ligand binding sites (LBS) on proteins, a critical aspect of structure-based drug design. Ligand binding site prediction is crucial for designing effective drugs and understanding protein functions, benefiting pharmaceutical companies, biotechnologists, and researchers by accelerating drug discovery and improving therapeutic interventions. We employ and improve Graph Neural Networks (GNNs) and innovative 3D point cloud instance segmentation to refine and advance LBS prediction methods. This research demonstrates significant enhancements in predictive accuracy by evaluating these methods on widely used datasets. Our novel clustering algorithm, which combines density-based and fuzzy clustering, notably improves the definition and identification of ligand binding sites without prior knowledge of the number of clusters. This methodology allows for more precise

predictions, effectively managing binding sites' overlapping nature. Implementing instance segmentation further delineates individual binding pockets, offering a more granular understanding of ligand-protein interactions. The results illustrate that our approaches meet the current state-of-the-art for ligand binding site prediction and support their potential utility in real-world pharmaceutical applications. Future work will focus on refining these methods and extending their application to molecular docking studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Dong Si, for their guidance and support throughout this project. I am also grateful to my committee members for their valuable feedback.

I extend my appreciation to my peers, the faculty, and staff of the Department of Computing and Software Systems, and the University of Washington, Bothell for providing essential resources and support.

Thank you all for your contributions to my academic journey.

1 INTRODUCTION

The pivotal role of ligand binding site prediction (LBSP) within the realm of structure-based drug design (SBDD) is increasingly recognized as a cornerstone for advancing therapeutic innovations. SBDD harnesses detailed knowledge of protein structures to ideate, develop, and refine drug candidates, emphasizing the identification of active sites where ligands can bind to influence protein function (Figure 1.1). This process is fundamental for the design of new drugs and for understanding biological mechanisms and interactions at the molecular level.

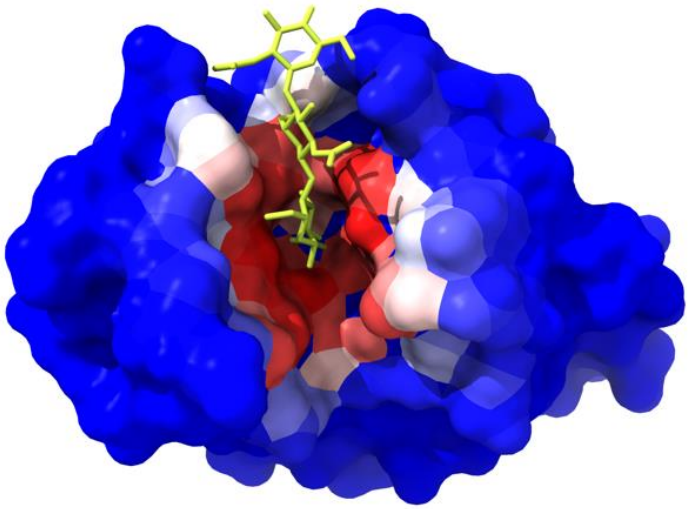

Figure 1.1 Protein 1hel in its holo (ligand-bound) state with one active site (in red).

In the broader context of drug discovery, the accurate prediction of these sites enhances the efficacy and specificity of pharmaceutical agents, directly impacting therapeutic outcomes by enabling more targeted treatments. The role of LBSP is elucidated in Figure 1.2 where the pink box signifies the ligand binding site identification step [1]. Historically, the field has evolved from

empirical methods to more sophisticated computational models, thanks to technological advancements.

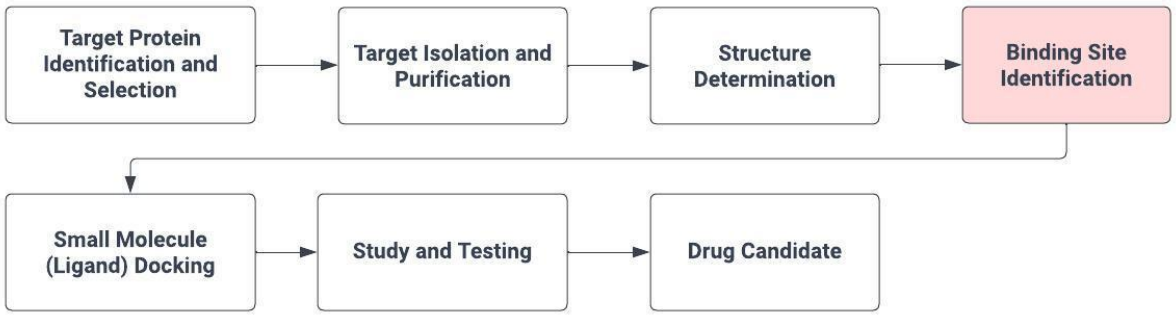

Figure 1.2 A Simplified Overview of Structure-Based Drug Design

Recent years have witnessed significant technological advancements in LBS prediction techniques, moving from traditional docking algorithms to complex machine-learning and deep-learning models that leverage large datasets of known protein-ligand interactions. These developments have substantially improved the predictive accuracy, efficiency, and scope of computational drug design tools. Techniques such as molecular dynamics simulations, and more recently, deep learning approaches have reshaped the landscape of SBDD.

1.1 Deep Learning for LBSP

The integration of machine learning (ML) and deep learning (DL) techniques into this research area marked a significant evolution, bringing about data-driven methods that could learn complex patterns from vast datasets, thereby improving prediction accuracy and efficiency. While many traditional methods were template-based methods, deep learning-based techniques fall under

template-free methods where there is no requirement for existing structural templates. These methods usually use the protein's structure and its chemical and physical features to learn the characteristics of ligand-binding pockets. Traditional machine learning techniques like Random Forrest Classifiers were among the first being utilized for ligand binding site prediction tasks. Subsequently, Convolutional Neural Networks (CNNs) were particularly successful, leveraging their spatial processing capabilities to analyze the three-dimensional structures of proteins at an unprecedented depth. Proteins are represented as 3D grids (voxelized) which are then used to train CNNs.

With the development of new deep learning architectures, Graph Neural Networks (GNNs), a class of deep learning models that treat protein structures as graphs, enabled a direct representation of their atomic interactions and spatial relationships. GNNs showed exceptional promise in accurately identifying LBS by capturing the nuanced geometries of protein surfaces. Each protein atom is considered a node and edges are created using spatial proximities to other nodes. Their ability to model proteins in a manner that mirrors their natural structure offers a powerful tool for LBSP, potentially surpassing the capabilities of traditional computational methods and CNNs alike.

Moreover, there are more solutions continuously being innovated. The latest are transformers which use attention mechanisms, and protein language models (PLMs). By understanding the 'language' of proteins, PLMs can infer function from sequence alone, often with remarkable accuracy, bridging gaps where structural data may be incomplete or unavailable.

Despite these advancements, many methods rely on clustering to convert raw predictions to relevant ligand-binding pockets. However, this comes with challenges and can also be a bottleneck

for the performance of the model. Clustering of prediction data into potential binding sites is a critical step in this process, the focus of our research transcends the clustering issue alone.

1.2 Problem Statement

Understanding how drugs interact with proteins is a foundation of modern medicine, guiding the design of new therapeutics and shedding light on protein function. Central to this understanding is the prediction of ligand binding sites on proteins—a task that blends geometry with biochemistry. A key insight from the shape theory of ligand binding is that while many factors contribute to a ligand's binding, the physical shape and size of the binding pocket play a role as well [2]. This implies that accurately identifying the location and the specific atoms or residues that define a binding pocket plays an important role in a comprehensive understanding of ligand-protein interactions.

While many tools have made strides in predicting the locations of these binding sites, focusing on identifying the centers of potential pockets with notable success, they fall short in detailing which atoms or residues precisely constitute these pockets. Many tools treat the problem as a semantic segmentation problem and require post-processing in the form of clustering to retrieve individual binding sites. The residues can be over-segmented into small but many pockets. Conversely, they can be under-segmented, combining 2 relevant pockets into a single pocket. One study [3], attempts to tackle the over-segmentation issue using graph-based clustering. This can especially be a problem when the true binding sites are too close or overlapping with each other. An example of overlapping pockets can be seen in Figure 1.3. This gap in precision, a noted limitation, underscores a challenge in the field: the detailed, residue-level prediction of individual binding sites.

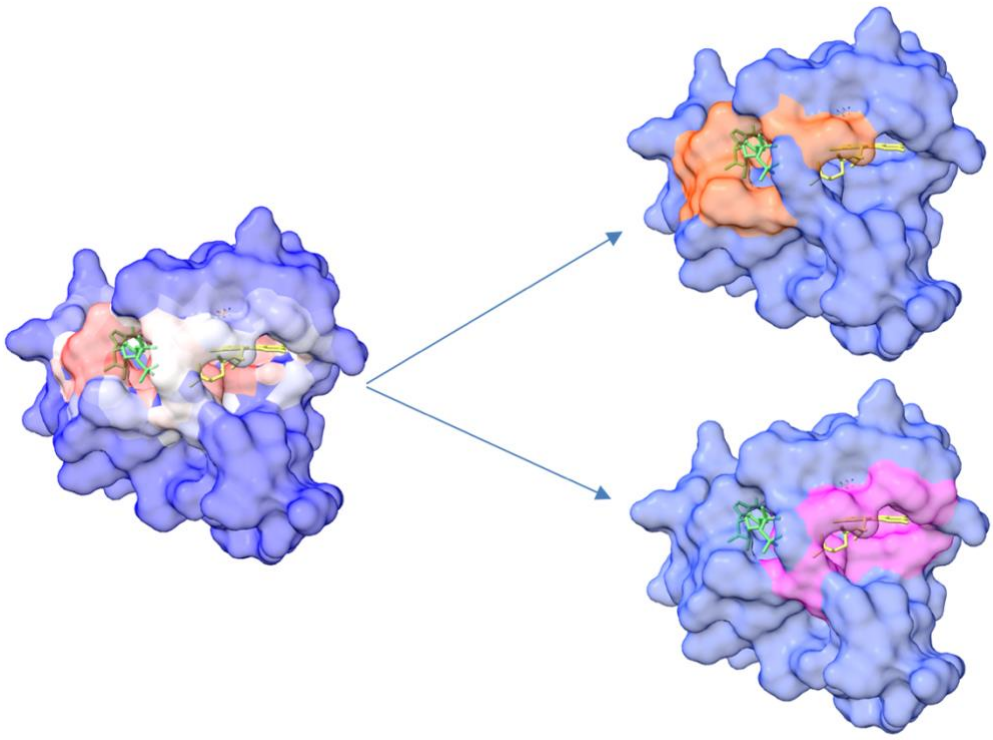

Figure 1.3 (PDB Id: 1n07A) An example of overlapping pockets. Orange (top-right) and pink (bottom-right) colors on the protein surface are ground truth pockets that have overlapping residues.

This study aims to address specific gaps in current methodologies by proposing innovative solutions that leverage the latest advancements in computational power and algorithmic design to solve the problem of ensuring that predictions pinpoint the potential participants in ligand binding and how they come together to form functional pockets.

This issue can either be solved by addressing it with a focus on developing clustering methods that can accurately interpret and refine outputs, ensuring that the predictions not only reflect the complex nature of protein-ligand interactions but align with biological reality also, or the deep learning model should be able to predict the separate pockets. Ensuring correct pocket prediction is beneficial for downstream use cases such as molecular docking and de novo ligand generation.

The goal is to provide a tool that is not only academically interesting but practically applicable in real-world drug development scenarios, fulfilling the needs of pharmaceutical researchers and developers by providing them with more reliable tools for predicting drug interactions.

Therefore, this study concerns two key research questions: i) Can enhancing clustering techniques amplify the advantages of deep learning methods? ii) Is there a deep learning-based approach capable of accurately predicting individual binding pockets?

2 RELATED WORKS

This section provides an overview of previous research relevant to ligand binding site prediction (LBSP). By examining various methods and approaches, this section aims to position this study within the broader context of existing literature and situate this work within that context. The section outlines the various methods ranging from early attempts to automate LBSP to modern deep-learning techniques. This study's results as compared to the related works are detailed in Section 4

The quest for accurate ligand binding site prediction (LBSP) is foundational in bioinformatics, with implications ranging from drug discovery to understanding protein functionality. Over the years, numerous endeavors have been undertaken to streamline the identification of ligand binding pockets, spanning from algorithmic approaches to data-driven deep learning techniques.

2.1 EARLY AUTOMATION METHODS

Initial automated strategies applied various techniques for LBSP. Some were developed to detect binding pockets using geometric criteria. Tools such as Fpocket [4], which utilizes Voronoi tessellation and alpha sphere detection, and LIGSITEcs [5], which leverages Connolly surfaces, were instrumental in providing valuable insights. Despite their utility, these methods lacked the granularity needed to capture the dynamic interplay between ligands and proteins. These methods often resulted in the prediction of many pockets on the proteins, which subsequently necessitated domain knowledge to discern the actual pockets. While these methods were the more popular ones, other strategies, like energy field methods, molecular dynamics calculations, etc., were also employed by LBSP tools [2].

2.2 TRADITIONAL MACHINE LEARNING

While traditional methods have laid the groundwork for ligand binding site prediction, recent advancements in machine learning and deep learning techniques have introduced new possibilities for improving predictive accuracy and understanding protein-ligand interactions. The subsequent integration of machine learning represented a significant methodological evolution.

A notable example of such a method is P2Rank [6]. It utilizes a random forest classifier, enhancing the predictive process by accommodating complex patterns intrinsic to biological data. This shift to data-driven techniques illustrated an increased accuracy over traditional methods but still required improvements to address the intricate nature of protein structures. Despite its improved accuracy in identifying the location of the ligand binding pockets over traditional methods, P2Rank encountered limitations in predicting the shape and size of the pocket.

2.3 DEEP LEARNING

The advent of deep learning, particularly geometric deep learning, further enhanced the predictive accuracy of LBSP. Along with different architectures, tools explore different representations of the protein structures viz. grid-based, graph-based, and surface mesh-based [7][8].

2.3.1 *CNNs*

CNNs are the first examples of successful application of deep learning in the LBSP domain. The protein's structure is transformed into small units called voxels to be given as input to the CNN models. This approach treats the problem akin to an image semantic segmentation task. Subsequently, pocket voxels are extracted from the masks and clustered to identify pockets, or

they are displayed in their original form. CNN-based models, including Kalasanty [9], DeepSurf [10], and PUResNet [11], demonstrated remarkable proficiency in processing the three-dimensional spatial information of proteins. DeepSurf, especially, displayed state-of-the-art performance.

2.3.2 *GNNs*

GNNs have introduced a paradigmatic shift in LBSP by modeling proteins as graphs, thus capturing the intricate relationships between amino acids beyond mere spatial proximity. This approach offers a fresh perspective on solving the problem, closely aligning with the structural properties of proteins, and presenting numerous advantages. Several GNN-based tools exist [12][13], each bringing with it new ideas to set their solution apart.

GNNs like EquiPocket [14] have leveraged the graph structure to provide deeper insights into the geometric and structural aspects of binding sites, highlighting the potential of GNNs to understand the functional implications of protein-ligand interactions. Equipocket employs important techniques such as local geometric modeling, global structure modeling, and surface message passing to better capture the structure and features of proteins. Innovations such as the graph attention networks in GrASP [15] and transfer learning strategies in LigBind [16] illustrate GNNs' ability to focus on critical areas of the protein surface and generalize across different ligands, showcasing the versatility of GNNs in capturing diverse binding phenomena.

2.3.3 *Continuous Innovation*

Additional approaches to tackle the challenge of ligand binding site prediction exist. PeSTo [17] is based on transformers and utilizes an attention mechanism. PointSite [18] views the proteins as a point cloud and performs semantic segmentation on them. Moreover, tools like LaMPSite [19] and [20] have brought in a different perspective, predicting ligand binding pockets from protein sequence information, and eliminating the need for 3D protein structure.

Additionally, the combined protein-ligand complex prediction has garnered significant interest. Tools like AlphaFold 3 [21], RoseTTAFoldAA [22], and NeuralPLexer [23] only require protein sequence and ligand SMILEs to predict combined 3D structures with ligand docked in the binding site. While these methods are highly accurate, the binding sites are specific to the ligand input. These methods are also not suitable for applications where de novo ligand generation is the subsequent task.

2.4 Scope for Improvement

To answer the questions this study aims to answer, it is important to know the shortcomings of current methods. The limitations of traditional clustering techniques, which often lead to the fragmentation of true binding sites, underscore the need for more sophisticated approaches. GNNs excel in identifying binding residues, but their performance is constrained by the efficacy of the clustering algorithm employed, mirroring the limitations observed in CNNs. The quality of the final clustering process consequently overshadows the effectiveness of both.

The next 2 sections highlight some related methods for both the questions this study explores.

2.4.1 *Clustering Algorithms*

For the first question, it is important to understand the shortcomings of using current clustering algorithms. The two characteristics of the ligand binding pockets in proteins that needed to be addressed were that the pockets can vary in size and by number from protein to protein.

Density-based clustering models, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [24] and Mean-Shift Clustering, excel in recognizing the complex, irregular shapes characteristic of ligand binding pockets. These methods adapt well to the diverse sizes of these pockets, a crucial advantage given the variability encountered across different proteins. Unlike traditional clustering techniques, density-based approaches do not require a predetermined number of clusters, making them particularly suited for the unpredictable nature of ligand binding pocket estimation.

However, these density-based methods fall short in supporting fuzzy clustering—a technique often necessary when dealing with the overlapping and ambiguous regions frequently observed in protein structures. Fuzzy clustering allows for the assignment of a single element to multiple clusters with varying degrees of membership, reflecting the biological reality where certain residues might contribute to multiple functional sites. Fuzzy C-Means (FCM) and Gaussian Mixture Models (GMM) are examples of such fuzzy algorithms. However, a drawback of these methods is that they require the number of clusters to be given by the user which defeats the purpose of detecting pockets.

Without accurate clustering, the risk of fragmenting a unified site into disjoint parts or, conversely, merging separate sites into a unified whole becomes a significant barrier to accurate prediction. Thus, the need for advanced clustering algorithms is clear.

2.4.2 *Instance Segmentation for Individual Pocket Identification*

Studies employing CNN methods like DeepSurf [10] and point cloud segmentation techniques like PointSite [18] have demonstrated the potential of semantic segmentation as a promising tool for LBSP. This insight suggests that instance segmentation could also prove invaluable in this context. Instance segmentation is a computer vision technique that aims to identify and segment individual objects within an image. It goes beyond just detecting objects and instead assigns a unique label to each instance of an object, providing a pixel-wise breakdown of the image.

Semantic segmentation is a simpler approach than instance segmentation, employed by PointSite has already shown that segmentation techniques can be applied to this problem. PointSite provided initial evidence that this approach could work, but it didn't distinguish between different binding sites. Nevertheless, this study underscores the viability of segmentation techniques for understanding and training on protein structure data.

Despite the promising prospects of semantic segmentation, there is currently a dearth of research exploring instance segmentation as a tool for LBSP. This gap may stem from the perceived limitations of instance segmentation techniques tailored for 3D point clouds, which may not yet possess the necessary sophistication to effectively differentiate object instances. Consequently, there is a need to evaluate the effectiveness of instance segmentation in this domain. Fortunately, emerging models such as ISBNet [25] and PBNet [26] are pioneering the application of instance segmentation to point clouds, presenting an opportunity to harness these advancements for even more precise ligand binding site predictions, thus helping address the second question this study wishes to explore.

3 METHODOLOGY

This section consists of two parts: the detailing of the method to improve the performance of graph neural networks for LBSP through better clustering method and the implementation of 3D point cloud instance segmentation for the same.

3.1 IMPROVING CLUSTERING FOR GNNS

To obtain a base deep learning model for evaluating clustering algorithms, it was imperative to find models that run into the clustering issue. Since GNNs were the more recent research direction, it only made sense to evaluate their performances. Although several methods exhibited promising results, with each showcasing unique concepts contributing to their efficacy, many lacked publicly available codebases, rendering them unsuitable for evaluation in this study.

Consequently, two particularly promising methods, namely GrASP [15] and PeSTo [17], were selected for further investigation. Another method, ScanNet [27] was also selected, however, through evaluation it was determined that it was more suitable for protein-protein interaction and did not perform well for protein-ligand interactions.

To compare the selected methods, the datasets and metrics need to be consistent. The predictions for each model were made according to the manner specified by the respective methods. The models were obtained from their GitHub, the inferences were run on the dataset and finally, the metrics were calculated as consistently as possible.

3.1.1 *Datasets*

The datasets used in this study are the ones commonly used to benchmark LBSP models. These datasets have the protein structure and the corresponding ligand structures. The protein structures are contained in Protein Data Bank or PDB (.pdb) files. The ligand structure information is also contained in PDB files but can sometimes be present in Tripos Molecular File Format (.mol2).

1. BU48 [28]: This dataset contains 62 samples of protein-ligand complexes. All samples are single-chain proteins and single ligands corresponding to ground truth. This is a common dataset used for benchmarking many LBSP models.
2. COACH420 [6][28]: This is a benchmarking set containing 420 samples of protein-ligand complexes derived from a 500-protein and 814-ligand test set. The samples are single chains but can have more than one ground truth ligand.
3. COACH420(mlig) [6]: This is a subset of COACH420 where the ligand codes are validated from the MOAD 2013 database. It contains 299 proteins and 391 ligands.

The BU48 dataset offers a more forgiving assessment of ligand binding site prediction (LBSP) methods, given that the sample complexes it contains each have only a single ligand, simplifying the prediction task. In contrast, the COACH420 dataset presents a more challenging testbed due to the presence of multiple ligands within its sample complexes, demanding a higher level of discernment from LBSP methods. Meanwhile, the COACH420(mlig) subset provides a middle ground in terms of difficulty, being somewhat more forgiving than the full COACH420 set yet still more complex than the BU48 set due to the variety of ligand interactions it includes.

3.1.2 *Metrics*

The performance of methods evaluated in this study is quantified by two categories of metrics, the ligand-centric and pocket-centric metrics. DCC_{pl} and DCA are ligand-centric, i.e., these metrics are the ones that are evaluated with the ligand as the reference ground truth, and OVR and DCC_{pp} are pocket-centric, i.e., the ground truth protein pocket residues or atoms are considered as references.

They are defined as follows:

1. **Success rate (DCC_{pl} <= 4Å) or DCC_{pl}** = $\frac{Number\ of\ predicted\ pockets\ with\ DCCpl <= 4Å}{Total\ number\ of\ ligands}$

   Here, DCC_{pl} (Distance center-center pocket-ligand) is defined as the distance between the center of the predicted pocket and the center of the atoms of the ligand. This metric is also referred to as just DCC_{pl}.

2. **Success rate (DCA <= 4Å) or DCA** = $\frac{Number\ of\ predicted\ pockets\ with\ DCA <= 4Å}{Total\ number\ of\ ligands}$

   Here, DCA (Distance center-atom) is defined as the distance between the center of the predicted pocket and any heavy atom (i.e. not a hydrogen atom) of the ligand. This metric is also referred to as just DCA.

3. **OVR** = Atom level intersection over the union of ground truth pocket and predicted pocket [10].

4. **Success rate (DCC_{pp} <= 4Å) or DCC_{pp}** = $\frac{Number\ of\ predicted\ pockets\ with\ DCCpp <= 4Å}{Total\ number\ of\ ligands}$

   Here, DCC_{pp} (Distance center-center pocket-pocket) is defined as the distance between the center of the predicted pocket and the center of the ground truth pocket. This metric is also referred to as just DCC_{pp}.

The distances measured are the Euclidean distance, and the center of the pocket is extracted by constructing a convex hull of the surface atoms of predicted pockets and finding its center [15]. If this distance is less than or equal to 4 Angstroms, the prediction is considered a success. These successes are summed over all samples and divided by the total number of ligands in the data set.

While DCC_{pl} and DCA metrics are the gold standard and are used across many other scientific studies, there are other important metrics like Success Rate (DCC_{pp} <= 4Å) where DCC_{pp} is the distance between the center of the ground truth ligand binding pocket and the center of the predicted pocket have been also calculated in this study. For OVR and DCC_{pp}, the center of the pocket is calculated as a simple centroid of all atoms in the pockets.

It was decided to have different center calculations since the two types of metrics serve different purposes. The ligand-centric metrics are more useful to downstream applications of LBSP and pocket-centric methods more effectively measure the accuracy of the prediction.

3.1.3 *Improving GNN performance using a clustering algorithm.*

To assess the impact of improved clustering on model performance, various clustering algorithms were tested. Using GrASP as the baseline GNN model, each clustering method was evaluated against standardized metrics. The selected algorithms for evaluation included agglomerative hierarchical clustering (average linkage), Gaussian Mixture Model, Mean-shift, and DBScan, all of which have been utilized in prior studies. A weighted version of hierarchical clustering that uses the confidence scores as weights.

Furthermore, as previously noted, there is a need for an advanced clustering algorithm capable of efficiently grouping predicted ligand binding residues or atoms without requiring manual specification of the number of clusters.

To address this gap and harness the benefits of both clustering approaches, this study explores the integration of density-based clustering with fuzzy clustering techniques. This comparative analysis aims to determine whether a hybrid clustering strategy could outperform traditional methods in the nuanced task of ligand binding pocket prediction. By merging the adaptability of density-based methods with the nuanced classification offered by fuzzy clustering, this study aims to create a more versatile and accurate approach to identifying ligand binding sites on proteins. The final pipeline is shown in Figure 3.1. The final pipeline employs the Mean Shift clustering algorithm in conjunction with Fuzzy C-means.

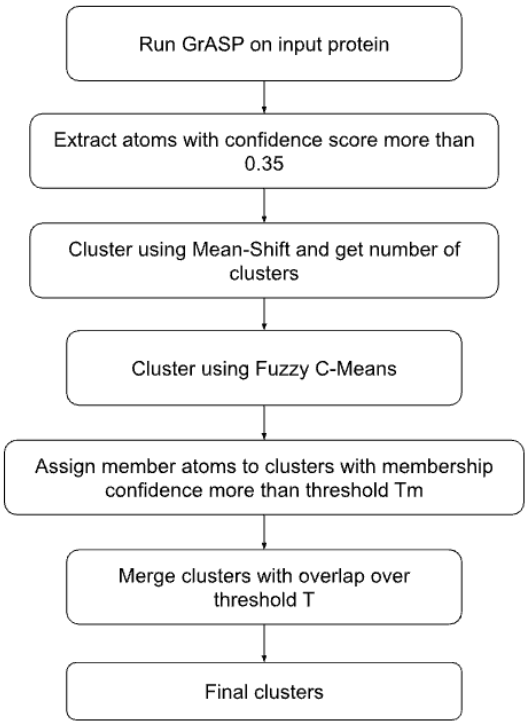

Figure 3.1 New GrASP pipeline (EnhancedGrASP) to extract pockets from predicted residues.

Additionally, Gaussian Mixture Models (GMM) were considered as an alternative to Fuzzy C-means, however, this approach was later discarded due to scattered clusters from GMM.

In the final pipeline, which will be referred to as ‘EnhancedGrASP’ hereon, the input PDB files are fed into GrASP, which predicts the confidence or "ligandability" score for each atom and embeds it into the PDB file itself. Following this, atoms with a confidence score above the threshold of 0.35 are extracted from the PDB file and inputted into the Mean Shift clustering algorithm. A bandwidth of 7.1 demonstrates optimal performance for estimating the number of clusters. However, these initial clusters do not entirely address all previously highlighted issues. Consequently, the number of clusters derived from this process is passed to the subsequent clustering algorithm, Fuzzy C-means.

Parameter values of 1.8 and 0.005 for 'm' and 'error', respectively, are determined to provide an appropriate level of fuzziness for our specific use case. Clusters are then generated, considering a threshold of 0.2 for membership. This means that atoms are assigned to clusters if their membership probability, as calculated by Fuzzy C-means, exceeds 0.2. Finally, the degree of overlap between clusters is assessed. If any two clusters overlap by more than 75%, they are considered the same cluster and are subsequently merged. These refined clusters represent the final ligand-binding pockets required for our analysis.

However, relying solely on this method did not mitigate the atom-level overlap of clusters (OVR), necessitating post-processing of the clusters. Upon examining how GrASP, P2Rank, and DeepSurf achieve remarkable accuracy in identifying pockets while still exhibiting low overlap with the ground truth, it becomes apparent that they extract surface points or atoms to delineate the pockets. While effective in identifying regions of interest, ligand binding typically occurs at

the residue level. Therefore, extracting the associated residues of the identified atoms enhances the alignment of predicted pockets with the ground truth counterparts.

3.2 3D Point Cloud Instance Segmentation

To assess the potential utility of instance segmentation for LBSP, meticulous consideration was given to the selection of each testing component. Ensuring the quality of the model, dataset, and training process was paramount.

After thorough examination and experimentation with the reusability of various models, ISBNet [25] emerged as a promising candidate for training on protein data. ISBNet is a recently published work, and it employs advanced techniques for effective instance segmentation. It introduces a cluster-free framework using Instance-aware Farthest Point Sampling (IA-FPS) and Box-aware Dynamic Convolution to generate high-recall kernels and leverage geometric cues from axis-aligned bounding boxes, achieving state-of-the-art accuracy and efficiency on multiple datasets.

3.2.1 *Training Dataset Preparation*

This study introduces a new method of representing protein atoms, as a point cloud. Considering the absence of dedicated datasets tailored for instance segmentation in LBSP, this study resorted to leveraging the sc-PDB [29] dataset. Due to the scale of the dataset, a subset of 7259 proteins. This decision stemmed from the dataset's possession of ground truth residue information. Additionally, the coach training set [30][31] contains 400 proteins and 908 ligands to help generalize the model to recognize smaller, shallower, and multiple pockets. To process the dataset,

feature engineering strategies akin to those utilized by GrASP were adopted. By incorporating both chemical and physical attributes of residues known to impact ligand binding, the dataset was enriched. These features have been used traditionally to determine the ligand binding pockets on proteins; hence it only makes sense to use these features as input to the model.

Following is the list of features used:

1. **Coordinates**: The x, y, and z coordinates of each atom.
2. **Residue Names**: The names of residues that have been one-hot encoded.
3. **Atom Type**: The atoms C, N, O, and S are considered as features. These are also on-hot encoded. Hydrogen atoms are omitted.
4. **Solvent Accessible Surface Area (SASA)**: This is the surface area of the atom that is exposed to solvent. This helps determine if the atom is in the interior of the atom or if it is on the surface where it can bind to ligands.
5. **RDKit Features**: These features are extracted using the RDKit library.
    a. **num_bonds_w_heavy_atoms**: The number of bonds an atom has with atoms other than hydrogen (heavy atoms).
    b. **formal_charge**: The formal charge of the atom.
    c. **is_in_ring**: A binary feature indicating if the atom is part of a ring structure.
    d. **is_aromatic**: A binary feature indicating if the atom is part of an aromatic ring.
    e. **hybridization**: The hybridization state of the atom (one-hot encoded).
6. **Chemical Features**: These binary features represent whether an atom has certain chemical properties:
    a. **acceptor**: Whether the atom is a potential hydrogen bond acceptor.
    b. **donor**: Whether the atom is a potential hydrogen bond donor.

c. **hydrophobe**: Whether the atom is part of a hydrophobic feature.
d. **lumped_hydrophobe**: Whether the atom is part of a larger hydrophobic region or cluster.

Radial density and mass of atoms were also considered as features initially. However, mass was later discarded as it did not contribute much to the model learning. Radial density, ideally, should be a useful feature to consider for LBSP, but may have been clashing with the model's own learnings of the densities. The removal of these features led to better convergence of the model.

Following this, a script was used to annotate the samples to mark their semantic and instance labels using the ground truth residues provided. The semantic labels indicate whether an atom is a pocket atom (1), or a non-pocket atom (0) and the instance labels are continuous integers indicating the pocket number the atom belongs to. Through this method, I had a comprehensive dataset for training and testing.

3.2.2 *Model Training*

The model training process necessitated substantial modifications to both its configuration and hyperparameters. Originally tailored for dense point clouds representing 3D scenes, the model initially struggled with the relatively sparse protein data. Extensive experimentation was required to discern the significance of each configuration parameter and conduct further trials to optimize values for the dataset at hand.

Additionally, ISBNet encountered challenges stemming from class imbalance, wherein the size of the ligand binding pockets was much smaller than the whole protein. This posed a hurdle

to effective learning. With far fewer atoms comprising the pockets compared to thos e outside, a class imbalance in semantic labels.

Some important configurations were changed as follows (non-exhaustive list):

1. Weights for cross-entropy loss used in the semantic segmentation backbone were set to 1 for class 0 (non-pocket) and 15 for class 1 (pocket). This helps overcome the class imbalance problem.
2. Standard non-max suppression (nms) was used instead of Matrix nms which helped eliminate potentially duplicate pockets. Standard nms is a more aggressive approach to reduce highly overlapping pocket predictions but proved to be suitable for this use case.
3. In the instance segmentation configurations, the radius considered by the local aggregation layer to collect neighboring points around each sampled point was set to 5. Also, the number of neighbors to be considered was 15.

Furthermore, the model used an aggregate of several loss functions during training. However, some loss functions were not effectively contributing to the learning, and may also be hindering the convergence of the model.

Following is the list of loss functions used for the final model:

1. Pretrain: Binary cross entropy, mean square error loss.

2. Fine-tuning loss functions used with their respective weights: Dice loss (weight = 1), focal loss (weight = 1), binary cross-entropy loss (weight = 1), classification loss (weight = 0.5) and intersection over union (iou) loss (weight = 0.5).

Initially, it appeared that filtering the atoms according to their SASA values so we only get surface atoms would work as effectively as it did for other models. However, in this case, considering whole proteins helped improve the performance of the model in multiple metrics.

The effectiveness of attention layers was tested in this study; however, it did not make any significant contribution. It also increased the resource demand of the model which was not ideal.

Finally, the training of the model is a 2-step process. First, the pre-training part of the model focused on the semantic segmentation of protein. Secondly, the fine-tuning part of the training learned to identify the individual instances of ligand binding pockets on the protein.

The model resulting after all these changes is referred to as ISBNet-Pocket from hereon. The final architecture is shown in Figure 3.2.

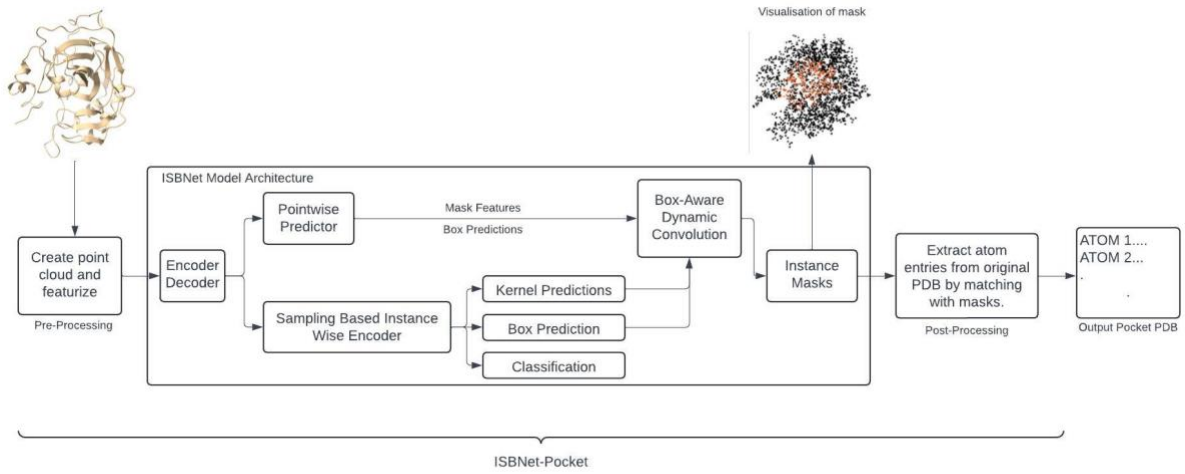

Figure 3.2 ISBNet-Pocket Architecture Diagram.

3.2.3 *Inferences and Metrics*

Post-processing was necessary to isolate the atoms within the pockets identified by ISBNet-Pocket. The model's output comprises distinct binary masks for each instance, necessitating scripts to read and apply these masks to the original PDB file, facilitating pocket retrieval. Only the masks with confidence higher than 0.65 were considered. The extracted pockets are saved in .pdb files. Subsequently, these extracted pockets serve as valuable inputs for various downstream applications or metric calculations. The metrics used for performance assessment are the same as the ones defined in section 3.1.2.

Additionally, the base model ISBNet reports semantic segmentation metrics for pretraining and instance segmentation metrics for fine-tuning. Those are defined as follows:

1. Pretrain/Backbone reported metrics (semantic segmentation):

   a. **Class-wise mIoU**: This metric calculates the mean IoU for each class. Higher values indicate better performance.
   b. **mIoU**: This is the overall mean IoU across all classes.
   c. **Acc**: This is the overall accuracy.

2. Fine-tuning reported metrics (instance segmentation):
   a. **AP (Average Precision)**: Measures the precision of the model for different instance classes.
   b. **AP_50%**: Measures the average precision at a 50% IoU threshold.
   c. **AP_25%**: Measures the average precision at a 25% IoU threshold.
   d. **AR (Average Recall)**: Indicates how well the model can identify all instances of a class.
   e. **RC_50% and RC_25%**: These recall metrics at different IoU thresholds indicate the proportion of true positive instances correctly identified by the model.

4 RESULTS

This section summarizes the results and observations from the comparative analysis of the PeSTo and GrASP models, the performance of clustering techniques, and the outcomes of the ISBNet-Pocket approach, alongside a review of notable findings and methodological impacts in that order.

4.1 BASE MODEL EVALUATION RESULTS

The results of the preliminary GNN comparison gave good insight into how the models worked and what the limitations of their performance were. On inspection, it was observed that pockets predicted by GrASP were larger and more precise than the ones predicted by PeSTo. Table 4.1 shows that GrASP shows superior performance across the test datasets. However, GrASP shows even better performance visually than these metrics reflect (Figure 4.1). The ground truth ligand appears to have a considerable amount of overlap with the predicted pocket in most cases. The natural conclusion to draw is that the clustering method used is not able to correctly segment the individual pockets to give correct pocket centers.

Table 4.1 Comparison between PeSTo and GrASP on BU48 and COACH (a superset of COACH420 with 500 samples) dataset, measuring DCC_{pl} and DCA.

Model	**BU48**		**COACH**	
	DCC_{pl}	**DCA**	**DCC_{pl}**	**DCA**
PeSTo	39.7%	44.8%	23.4%	39.7%
GrASP	72.88%	89.83 %	41.36%	59.09%

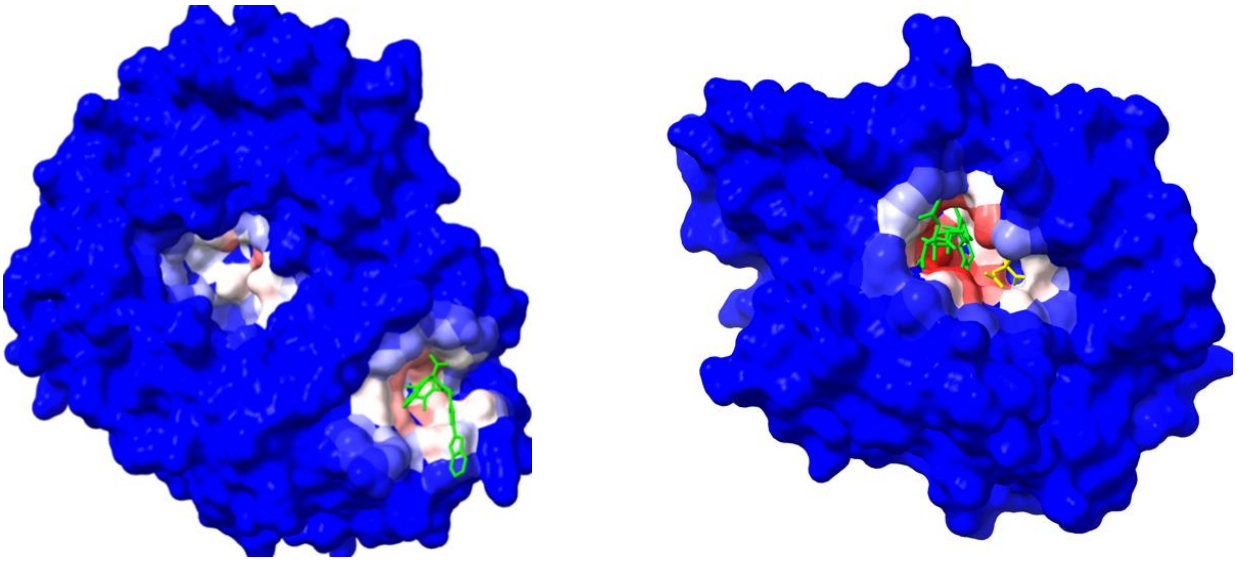

Figure 4.1 Visualizations of GrASP's predictions. (Left) Protein 1nhvA with 1 ground truth ligand (green). (Right) Protein 3b4yA with 2 ground truth ligands (green and yellow). Predicted regions with high confidence scores are shown in a gradient of blue, white, and red. Blue being the lowest confidence score(0) and red being the highest confidence score (1).

4.2 Evaluating EnhancedGrASP

With the base model selected, the next step was to evaluate whether better clustering can improve metrics. By strategically integrating Mean-Shift clustering with either Fuzzy C-means or Gaussian Mixture Models, the benefits of both density-based clustering and fuzzy clustering could be reaped. This combined approach aimed to accurately estimate the number of clusters in the unpredictable landscape of ligand binding sites while also accommodating the reality of overlapping clusters.

Table 4.2 shows that Mean Shift and Fuzzy C-means can capture the characteristics of density-based better than Mean Shift + Fuzzy C-means. The average number of pockets predicted

by it is 1.83 which is slightly over the average number of pockets in the Coach420 dataset (1.597). However, we can see that the clusters created are more accurate. Gaussian mixture models were also evaluated in place of Fuzzy C-means in the clustering pipeline (Figure 3.1); however, the inclusion of GMM causes the clustering to be too scattered, and, in some cases, completely nonsensical.

While the proposed clustering method shows comparable performance with other clustering performances, it does not have the best performance in terms of metrics. Yet, it is still the best method due to other high-performing methods having their disadvantages. Agglomerative Hierarchical clustering underestimated the number of pockets present. While this works in its favor for the BU48 dataset, it only estimates an average of 1.01 pockets for the Coach420 dataset which has an average of 1.67 pockets. From this, we can conclude that agglomerative clustering clubs together clusters that are very close. The mean-shift clustering algorithm also exhibits similar behavior. GMM shows high performance in terms of DCC_{pl} and DCA for both datasets; however, as mentioned before, it requires the number of pockets to be input by the user and gives scattered clusters. The scattered clusters remain an issue when testing GMM in place of FCM in the proposed post-processing clustering. Other methods, i.e., DBScan and Weighted hierarchical clustering performed sub-optimally.

To confirm the effectiveness of the proposed clustering pipeline, we can further analyze the results and understand how the pockets look by visualizing them in ChimeraX [32]. Figure 4.2 shows the predicted pocket on a protein with a single ligand. Most of the samples in datasets have single ground truth ligands.

To take another example that can better display fuzzy clustering, we can look at Figure 4.3. where the protein has 2 ligands that are situated close together; hence, their pockets overlap a bit.

Table 4.2 Comparison of various clustering methods on GrASP for BU48 and COACH420 datasets, measuring DCC_{pl}, DCA and input number of clusters. Highlighted are the best performances in each metric.

Clustering Method	**BU48**		**COACH420**		**Input number of clusters**
	DCC_{pl}	**DCA**	**DCC_{pl}**	**DCA**	
Agglomerative hierarchical clustering (average linkage)	72.88%	89.83 %	41.36%	59.09%	**non-parametric**
Weighted hierarchical clustering (average linkage)	67.79%	88.14%	41.57%	57.68%	**non-parametric**
Gaussian Mixture Model	**75.41%**	88.52%	**52.09%**	**63.30%**	5
Mean-shift	62.30%	**93.44%**	36.95%	59.98%	**non-parametric**
DBScan	59.32%	88.13%	40.95%	58.68%	**non-parametric**
Mean Shift + GMM	69.49%	86.44%	45.20%	58.87%	**non-parametric**
Mean Shift + FCM + post-processing (EnhancedGrASP)	71.18%	91.52%	46.40%	62.48%	**non-parametric**

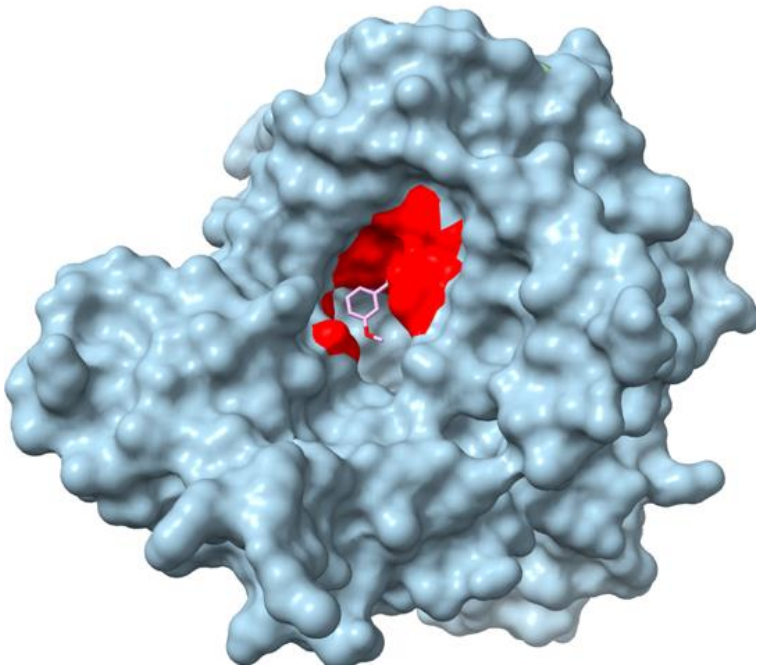

Figure 4.2 Protein 1efyA is a protein with 1 ground truth ligand. EnhancedGrASP predicts one pocket.

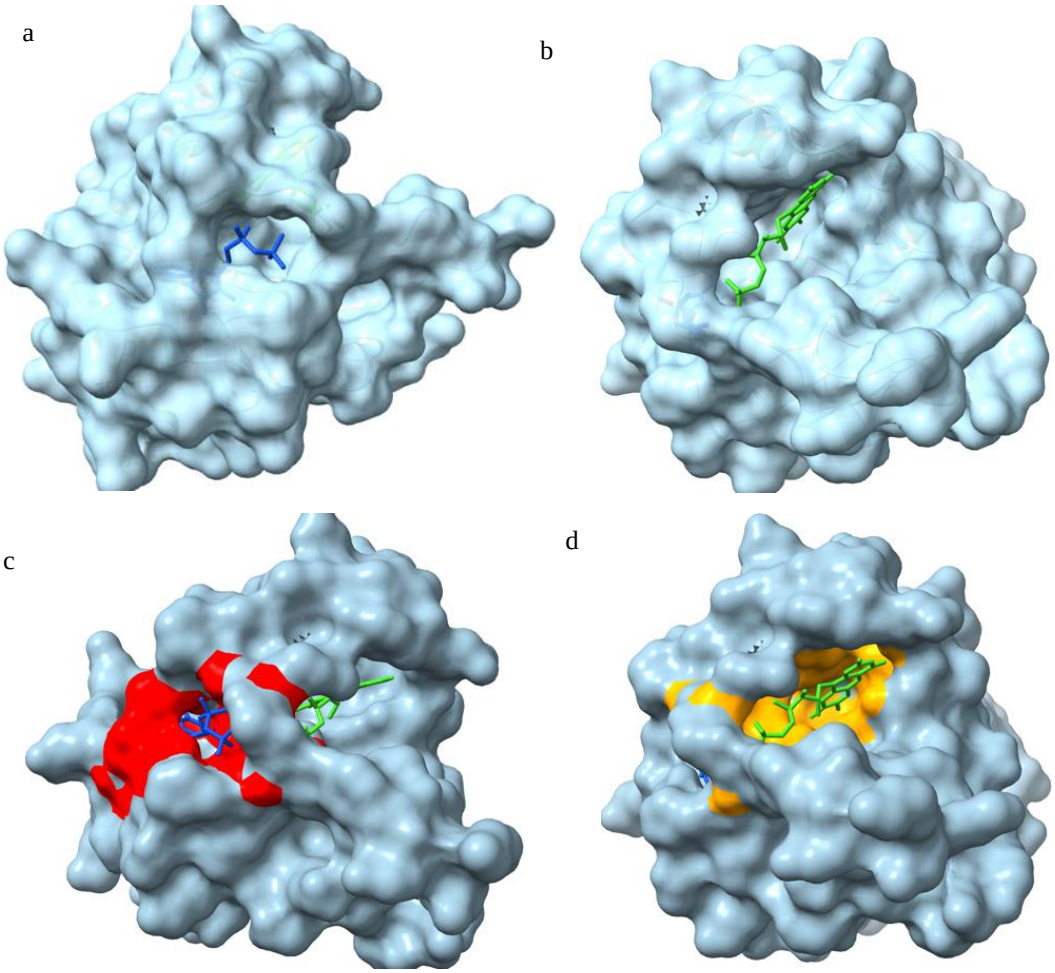

Figure 4.3 (PDB Id: 1n07A) Predicted overlapping pockets. Top row shows the ground truth ligands of protein 1n07A and the second row shows the predicted pockets against the ground truth ligands. a) The ligand ADP in blue partially occluded by the transparent surface. b) The ligand FMN in green partially occluded by the protein's transparent surface. c) Pocket 2 in red behind ligand ADP. d) Pocket 1 in orange behind ligand FMN. The 2 predicted pockets are overlapping in nature.

4.3 ISBNET-POCKET

While we can see that a custom clustering method can improve the estimation of the number of ligand binding pockets present in a protein, it is not based on any shape prior to the pocket. It

only considers the spatial proximities of atoms. ISBNet-Pocket, developed in this study can estimate the same without having to rely on clustering methods.

The proposed method, 3D point-cloud instance segmentation by representing the protein as a point cloud proved as an effective method for ligand binding site prediction. The model internally uses semantic and instance segmentation metrics to quantify its performance, these are summarized in Table 4.3 and Table 4.4. These values are calculated for the cross-validation set used during training. In terms of LBSP metrics on benchmark testing datasets, it was able to show comparable performance to other methods created by other researchers (Table 4.5).

The segmentation model outputs the masks for each instance which can be visualized in 3D space. Figure 4.4 shows a protein as a point cloud where the black (or dark) points are the non-pocket atoms and the orange (or light) points are the pocket atoms.

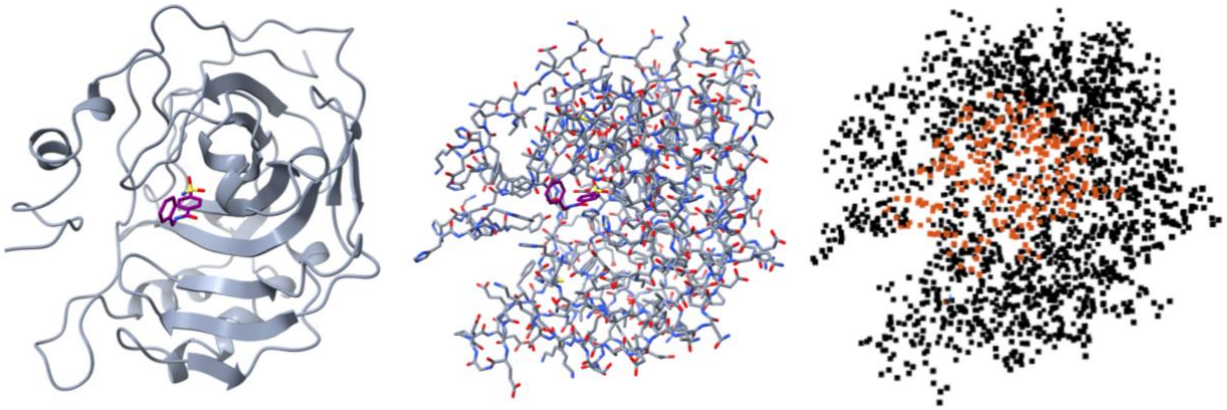

Figure 4.4 (PDB Id: 1g4oA) Point cloud representation used in ISBNet-Pocket (left) Protein structure and ligand BSB. (middle) The same protein structure in atom view. (right) Segmented protein represented as point cloud, orange points signify ISBNet-pocket predicted pocket.

Table 4.3 Semantic segmentation metrics on the validation set.

Metric	Value	
Class-wise mIoU	Non-pocket	Pocket
	86.2	32.0
mIoU	59.1	
Acc	87.1	
Offset MAE:	6.312	

Table 4.4 Instance segmentation metrics on the validation set.

Instance Class	AP	AP_50%	AP_25%	AR	RC_50%	RC_25%
Non-pocket	0.973	1.000	1.000	0.981	1.000	1.000
Pocket	0.212	0.560	0.807	0.329	0.750	0.953

A notable observation during training was that both the pretraining and fine-tuning parts of the training process did not need many epochs to converge.

4.4 Comparison with Other Models

To understand how EnhancedGrASP and ISBNet-Pocket performed compared to other available LBSP methods, the Coach420(mlig) dataset was used as a benchmark, and predictions were run for each. Since many methods use different subsets of the Coach420(mlig) dataset, it was determined that running predictions help keep comparison uniform. The findings are summarized in Table 4.5. DeepSurf [10] being one of the most successful LBSP models was a good choice for comparison.

While Deepsurf does a great job in predicting precise pockets, metrics for DeepSurf are calculated including the proteins where no pockets were predicted. This decision does bring the

metrics for DeepSurf lower; however, these samples are important aspects of DeepSurf's performance.

Through the results, we can conclude that having a hybrid clustering technique combining the powers of density-based clustering and fuzzy clustering can improve the prediction results of underlying GNNs. When compared to the performance of the base model, it is evident that even though the number of pockets predicted is higher, the predicted pocket centers are closer to the ground truth pockets. This can be observed through the DCC_{pp} metric, which is higher for EnhancedGrASP than the base model GrASP. The average number of pockets in the Coach420(mlig) dataset is 1.28.

Table 4.5 Comparison of LBSP models with EnhancedGrASP and ISBNet-Pocket on Coach420(mlig) dataset, measuring DCC_{pl}, DCA, DCC_{pp}, OVR, the average number of pockets predicted, and % proteins with no predicted pockets. Highlighted rows are methods introduced in this study.

Model	**COACH420(mlig)**					
	DCC_{pl}	**DCA**	**DCC_{pp}**	**OVR**	**Avg. num of Pockets predicted**	**% of proteins with no predicted pockets**
DeepSurf	40.20%	67.62%	46.21%	0.16	1.04	3%
GrASP	51.34%	73.11%	52.68%	0.26	1.14	0
EnhancedGrASP	**55.06%**	**76.16%**	**58.07%**	**0.24**	**1.79**	**0**
ISBNet-Pocket	**42.03 %**	**71.54 %**	**48.83%**	**0.37**	**1.57**	**0**

ISBNet-Pocket was also able to perform competitively with the state-of-the-art models, if not better in some regards. Its predictions have a higher overlap with the ground truth pocket

because all atoms of the protein are utilized. It has a fair estimate of the center of the ground truth pocket as we can infer from 48.83% DCC_{pp}.

However, these high metrics are not reflected by the DCC_{pl} metric. DCA shows that 71.54% of pockets are less than 4 angstroms close to an atom of the binding ligand. This begs the question of why DCC_{pl} shows such a substantial decrease compared to DCA. To investigate this discrepancy, this study visualized some proteins where the pockets were correctly predicted according to DCC_{pp} and DCA but failed according to DCC_{pl} metrics. It was observed that due to considering all atoms, unlike other models that consider only surface atoms, the center calculated was dragged down towards the protein surface. These metrics worsen when using simple centroid calculation to the center point. Some examples of this observation are shown in Figure 4.5, Figure 4.6, and Figure 4.7. There is a good overlap between the predicted pocket and ligand with most of the surface area covered. Yet the centers of the ligand and centers of the predicted pocket are more than 4 Å apart.

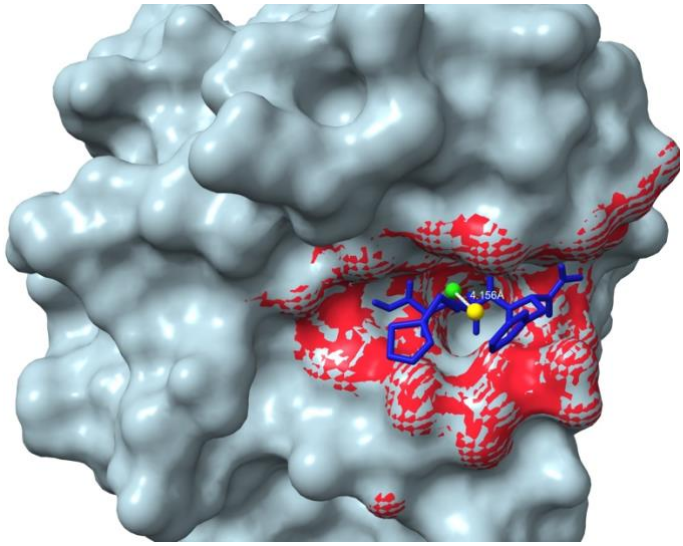

Figure 4.5 (PDB Id: 1mmbA) The center of the predicted pocket (green) and center of ligand (yellow) are 4.1 Å apart.

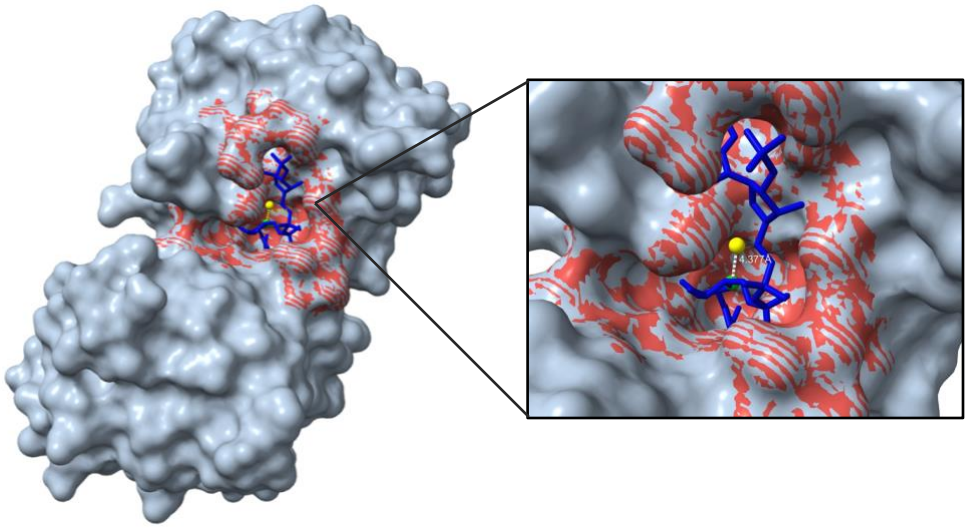

Figure 4.6 (PDB Id: 3bazA) Center of the predicted pocket (green) and center of ligand (yellow) are 4.377 Å apart.

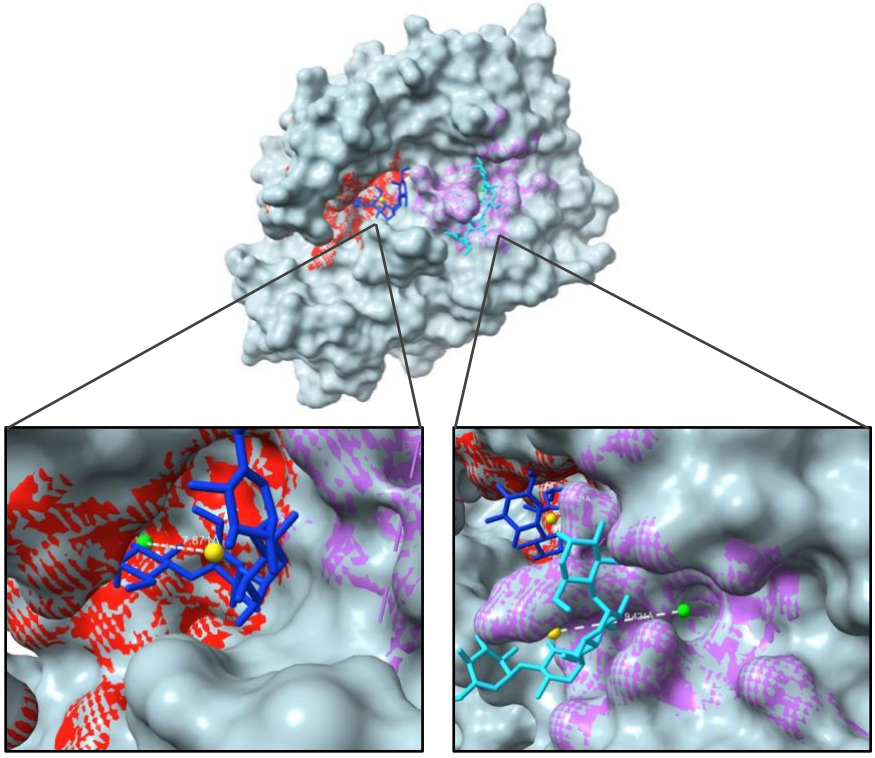

Figure 4.7 (PDB Id: 1eswA) Protein and its predicted pockets (top) Entire protein with 2 predicted pockets. (bottom left) The predicted pocket (green) and ligand (yellow) centers are 7.871 Å apart. (bottom right) Predicted pocket (green) and ligand (yellow) centers are more than 9.431 Å apart.

Finally, a notable observation for some examples was that the metrics may not reflect the success of the predicted pocket correctly when the pocket shapes are not circular (or spherical). An example of this is (Figure 4.8), where the pocket shape is irregular and even convex in nature when viewed from the side. In this example, the distance between the ligand and predicted pocket centers is 4.762 Å.

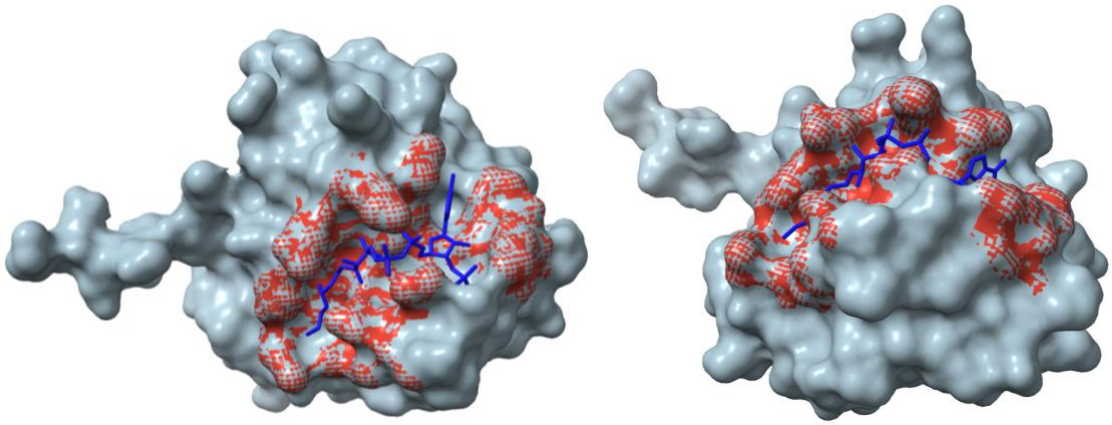

Figure 4.8 (PDB Id: 2e6uX) A special case of ligand-bound protein with ISBNet-Pocket predicted pocket which was only successful in the DCC_{pp} metric.

4.5 CASP16

The capabilities of ISBNet-Pocket were tested in Critical Assessment of Structure Prediction (CASP16), a biennial scientific competition in the field of computational biology. Researchers from around the world participate in CASP to assess and improve the methods for predicting the three-dimensional structure of proteins. CASP16 refers to the 16th iteration of this competition.

One of the tracks of the competition is to predict protein-organic ligand complexes which entails ligand binding site prediction. The residue-level predictions for targets are shown in Figure 4.9 and Figure 4.10. However, as expected the predicted pockets are large, hence the final

predictions were a consensus between several models. Nevertheless, ISBNet-Pocket's predictions do demonstrate its capabilities to predict pockets on unseen and challenging proteins.

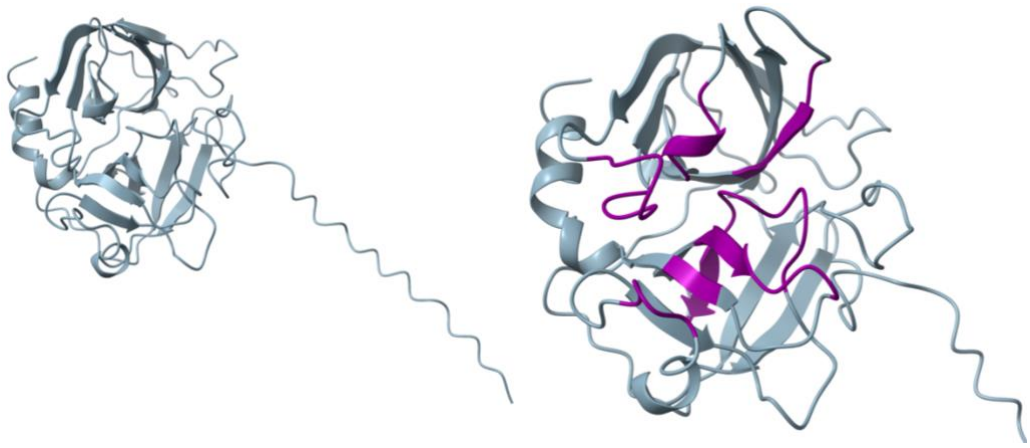

Figure 4.9 (Target L1000: Chymase) (left) Predicted structure. (right) Predicted pocket in purple.

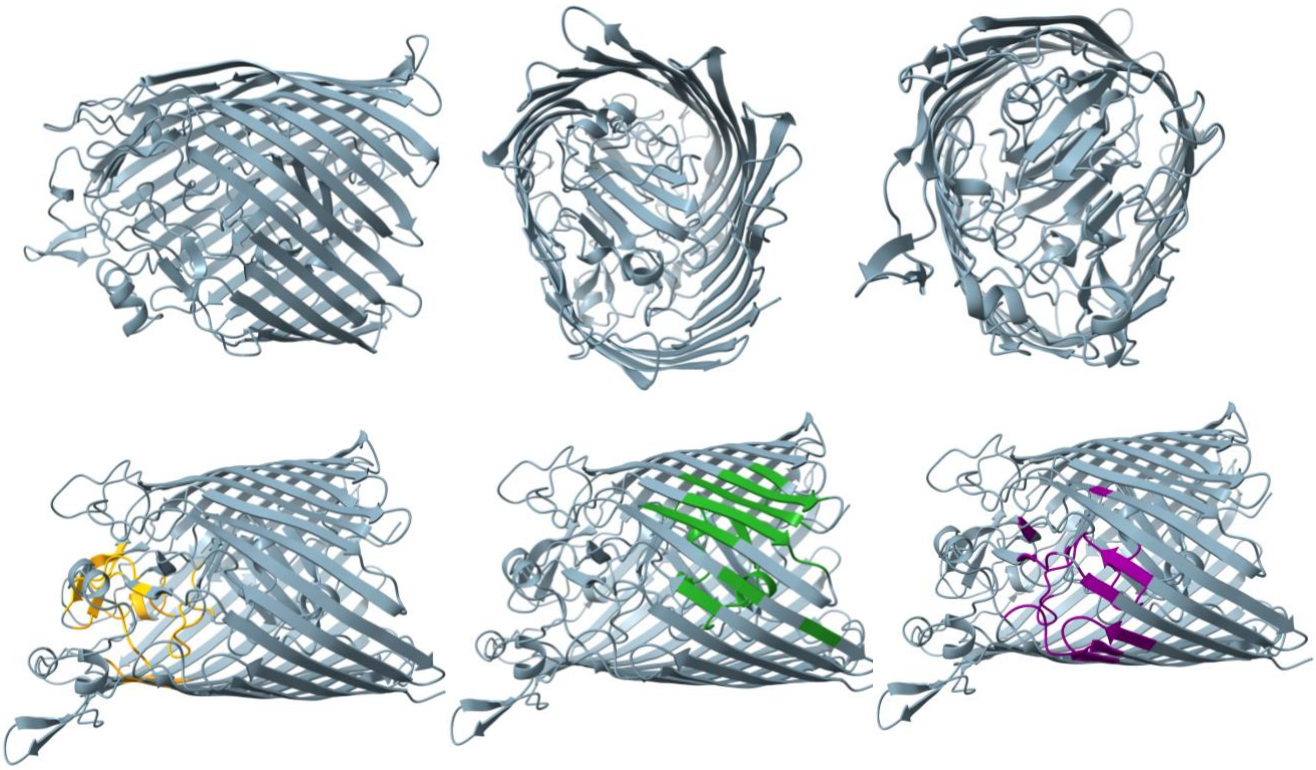

Figure 4.10 (Target T1214: YncD (NP_415968)) (top row) Predicted protein structure. (bottom row) 3 predicted pockets.

5 DISCUSSION

This section delves into the implications of the findings obtained from the analysis of EnhancedGrASP, ISBNet, and other related methods, which aimed to address the research questions concerning ligand binding site prediction.

5.1 ENHANCEDGRASP VS. ISBNET-POCKET

The results from this study indicate that the hybrid clustering approach, EnhancedGrASP, performs better than ISBNet-Pocket in some key areas. However, it's important to note that we can't conclusively say one method is superior to the other at this point. EnhancedGrASP's performance largely depends on the underlying model it builds on, and its improvements through hybrid clustering have their limits like sensitivity to the choice of parameters and poor generalization across different scenarios.

On the other hand, ISBNet-Pocket shows a lot of potential as it considers features other than distance, especially with its performance in the OVR metric. There's room to improve ISBNet-Pocket by using better training data, which could help overcome some of the limitations that affect clustering methods.

5.2 NEED FOR STANDARD METRICS

In this study, the DCC metric is uniquely divided into DCC_{pl} and DCC_{pp}, whereas typically, it is referenced simply as DCC; potentially representing either metric depending on the context. DCC and DCA are widely recognized metrics within the literature; however, their interpretations can vary significantly across studies. For instance, P2Rank [6] interprets DCC as the distance between

the center of the pocket and any atom of the ligand, whereas SiteRadar [13] considers a DCC_{pl} threshold of 5 Angstroms. The lack of uniformity in these metrics hampers effective comparison across models, underscoring the need for standardized metrics to enhance the robustness of future studies.

Additionally, a deeper understanding of which metrics most accurately reflect the correct prediction of pockets would be beneficial. Notably, many studies, including this one, use a threshold of 4 Angstroms for the DCC_{pl} distance. It is important to recognize that DCC_{pl} values exceeding 4 Å can still represent correct predictions, indicating that a re-evaluation of the commonly accepted thresholds may be necessary to better align with empirical observations.

As a result of this observation, it is apparent that DCC_{pp} is a more accurate metric to measure the success LBSP methods. However, this metric requires the ground truth pocket information to be available. Due to the unavailability of ground truth information in many datasets, studies often rely on DCC_{pl} to measure results. This further solidifies the need for better metrics for LBSP.

6 CONCLUSION

This study contributes to the advancement of ligand binding site prediction methodologies by leveraging the capabilities of GNNs combined with innovative clustering techniques and introducing a new method utilizing instance segmentation. Through the clustering pipeline we addressed the limitations of traditional clustering methods, we have demonstrated that a hybrid approach of density-based and fuzzy clustering improved the precision of binding site predictions. The developed method accurately estimated the number of clusters without prior knowledge and effectively managed the overlapping nature of ligand-binding sites, presenting a substantial improvement over existing techniques.

Additionally, this study also tested and proved the transferability of 3D point cloud instance segmentation for ligand binding site prediction and displayed its potential to compete with state-of-the-art methods. This technique captures the shape of the pocket and is also able to learn to differentiate between different pockets, unlike most other methods that require post-processing clustering to identify individual pockets.

Further study would be required to improve ISBNet-Pocket and subsequently determine whether ISBNet-Pocket surpasses the performance of EnhancedGrASP or other such clustering methods.

Finally, his study’s findings underscored the potential of machine learning and deep learning technologies in enhancing our understanding of protein-ligand interactions, which is crucial for drug discovery and the elucidation of protein functions.

6.1 LIMITATIONS

Both, EnhancedGrASP and ISBNet-Pocket are very effective in learning deep, well-defined pockets on the protein surface. However, in many cases, the ligands bind to proteins in even small and shallow pockets or clefts. These pockets are underrepresented in most datasets and often get overlooked. This could be due to the contents of sc-PDB, where the pockets are large well-defined, deep pockets. This also caused the pockets predicted by ISBNet-Pocket to be slightly larger than the ground truth. The predictions contain the ground truth residues but contain some peripheral residues as well.

To mitigate the issue, this study included samples from Coach test data as described in Section 3.2.1. That did not fix the problem entirely and identifying shallow pockets is still an outstanding issue. The only benefit of including the Coach training dataset was that the ISBNet-Pocket model predicted some shallow pockets. Their accuracy, however, is not up to par at the current state as the number of training samples in Coach was too less as compared to sc-PDB samples. Figure 5.1 shows examples of the same.

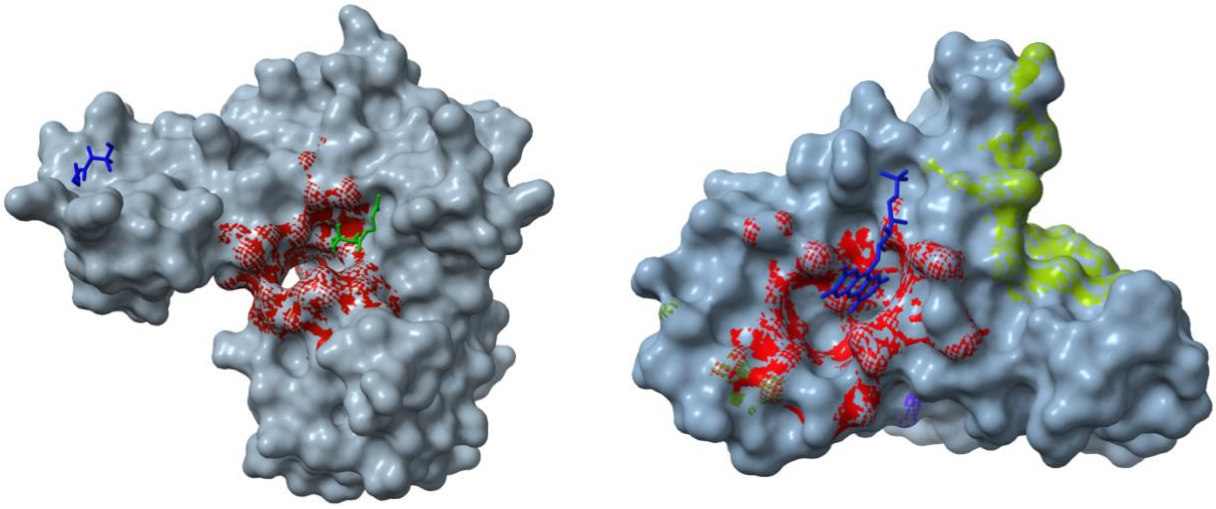

Figure 6.1 (left) (PDB Id: 2f9wA) Missed shallow pocket. (right) (PDB Id: 1y30A) Partially correct predicted shallow pocket.

Another limitation specific to ISBNet-Pocket is that it gave confused and large number of pocket predictions on small protein structures. Figure 6.2 shows an example small protein with 4 ligands. This example also demonstrates the failure to identify very shallow pockets as only 1 pocket (in green) is close to the ground truth.

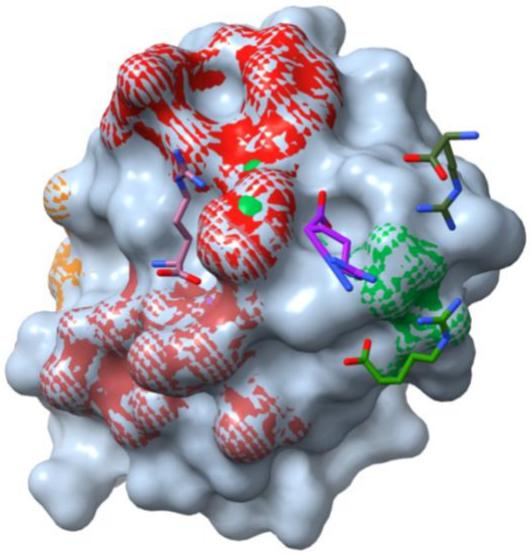

Figure 6.2 (PDB Id: 3cagA) Too many pockets predicted on a small protein.

6.2 Future Work

While ISBNet-Pocket is already performing at competitive levels, there are several promising avenues for further enhancement. Firstly, addressing the current shortcomings discussed in Section 6.1 can significantly elevate the model's accuracy and reliability for ligand binding site prediction (LBSP). These improvements can be achieved by expanding the training set to include a greater variety of proteins, particularly those with shallow ligand binding patterns. This will provide the model with a broader spectrum of examples, enabling it to generalize better across different types

of binding sites. Additionally, adapting the model to accommodate multi-chain proteins and interfacial binding sites will enhance its versatility and applicability to more complex biological systems.

Secondly, the role of conformational changes in protein-ligand interactions presents a rich area for future exploration. Proteins are dynamic entities, and their conformational flexibility can open up new binding pockets or occlude existing ones. By integrating data on protein conformational changes the model could predict static binding sites and those that emerge or disappear under different physiological conditions. This dynamic perspective would make ISBNet-Pocket an even more powerful tool for drug discovery and other applications.

Further exploration of the LBSP metrics could provide a deeper understanding of which metrics most effectively quantify the efficacy of the methods evaluated. This could also help push efforts to have a standard metric that can be used by all studies pertaining to LBSP.

Lastly, integrating ISBNet-Pocket with downstream tasks such as molecular docking and de-novo ligand generation can create a comprehensive and robust pipeline for drug discovery. Molecular docking can validate the predicted binding sites by simulating how ligands interact with these sites, while de-novo ligand generation can propose new molecules that fit the predicted pockets. Combining these capabilities would streamline the drug development process, from identifying potential binding sites to designing novel therapeutic compounds.

.

BIBLIOGRAPHY

[1] A. C. Anderson, “Review The Process of Structure-Based Drug Design speed at which drug leads can be identified and evalu-ated in silico. Structure-based drug design is most powerful when,” *Chem Biol*, vol. 10, pp. 787–797, 2003, doi: 10.1016/j.

[2] I. Ramírez-Velásquez, Á. H. Bedoya-Calle, E. Vélez, and F. J. Caro-Lopera, “Shape Theory Applied to Molecular Docking and Automatic Localization of Ligand Binding Pockets in Large Proteins,” *ACS Omega*, vol. 7, no. 50, pp. 45991–46002, Dec. 2022, doi: 10.1021/acsomega.2c02227.

[3] J. Degac, U. Winter, and V. Helms, “Graph-Based Clustering of Predicted Ligand-Binding Pockets on Protein Surfaces,” *J Chem Inf Model*, vol. 55, no. 9, pp. 1944–1952, Sep. 2015, doi: 10.1021/acs.jcim.5b00045.

[4] V. Le Guilloux, P. Schmidtke, and P. Tuffery, “Fpocket: An open source platform for ligand pocket detection,” *BMC Bioinformatics*, vol. 10, May 2009, doi: 10.1186/1471-2105-10-168.

[5] B. Huang and M. Schroeder, “LIGSITEcsc: Predicting ligand binding sites using the Connolly surface and degree of conservation,” *BMC Struct Biol*, vol. 6, Sep. 2006, doi: 10.1186/1472-6807-6-19.

[6] R. Krivák and D. Hoksza, “P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure,” *J Cheminform*, vol. 10, no. 1, Dec. 2018, doi: 10.1186/s13321-018-0285-8.

[7] Z. Zhang, J. Yan, Q. Liu, E. Chen, and M. Zitnik, “A Systematic Survey in Geometric Deep Learning for Structure-based Drug Design,” Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.11768

[8] C. Isert, K. Atz, and G. Schneider, “Structure-based drug design with geometric deep learning,” Apr. 01, 2023, *Elsevier Ltd*. doi: 10.1016/j.sbi.2023.102548.

[9] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, “Improving detection of protein-ligand binding sites with 3D segmentation,” *Sci Rep*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-61860-z.

[10] S. K. Mylonas, A. Axenopoulos, and P. Daras, “DeepSurf: A surface-based deep learning approach for the prediction of ligand binding sites on proteins,” *Bioinformatics*, vol. 37, no. 12, pp. 1681–1690, Jun. 2021, doi: 10.1093/bioinformatics/btab009.

[11] J. Kandel, H. Tayara, and K. T. Chong, “PUResNet: prediction of protein-ligand binding sites using deep residual neural network,” *J Cheminform*, vol. 13, no. 1, Dec. 2021, doi: 10.1186/s13321-021-00547-7.

[12] N. Abdollahi, S. A. M. Tonekaboni, J. Huang, B. Wang, and S. MacKinnon, “NodeCoder: a graph-based machine learning platform to predict active sites of modeled protein structures,” Feb. 2023, [Online]. Available: http://arxiv.org/abs/2302.03590

[13] S. A. Evteev, A. V. Ereshchenko, and Y. A. Ivanenkov, “SiteRadar: Utilizing Graph Machine Learning for Precise Mapping of Protein-Ligand-Binding Sites,” *J Chem Inf Model*, vol. 63, no. 4, pp. 1124–1132, Feb. 2023, doi: 10.1021/acs.jcim.2c01413.

[14] Y. Zhang, W. Huang Í Mí±Ž , Z. Wei Í Mí±Ž , Y. Yuan, Z. Ding, and Z. Mí±• , "EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction ; EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction," 2018. [Online]. Available: https://doi.org/

[15] Z. Smith, M. Strobel, B. P. Vani, and P. Tiwary, "Graph Attention Site Prediction (GrASP): Identifying Druggable Binding Sites Using Graph Neural Networks with Attention", doi: 10.1101/2023.07.25.550565.

[16] Y. Xia, X. Pan, and H. Bin Shen, "LigBind: Identifying Binding Residues for Over 1000 Ligands with Relation-Aware Graph Neural Networks," *J Mol Biol*, vol. 435, no. 13, Jul. 2023, doi: 10.1016/j.jmb.2023.168091.

[17] L. F. Krapp, L. A. Abriata, F. Cortés Rodriguez, and M. Dal Peraro, "PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces," *Nat Commun*, vol. 14, no. 1, Dec. 2023, doi: 10.1038/s41467-023-37701-8.

[18] X. Yan *et al.*, "PointSite: A Point Cloud Segmentation Tool for Identification of Protein Ligand Binding Atoms," *J Chem Inf Model*, vol. 62, no. 11, pp. 2835–2845, Jun. 2022, doi: 10.1021/acs.jcim.1c01512.

[19] S. Zhang and L. Xie, "Protein Language Model-Powered 3D Ligand Binding Site Prediction from Protein Sequence," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.03016

[20] H. Gamouh, M. Novotny, and D. Hoksza, "Hybrid protein-ligand binding residue prediction with protein language models: Does the structure matter?", doi: 10.1101/2023.08.11.553028.

[21] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, May 2024, doi: 10.1038/s41586-024-07487-w.

[22] R. Krishna *et al.*, "Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom", doi: 10.1101/2023.10.09.561603.

[23] Z. Qiao, W. Nie, A. Vahdat, T. F. Miller, and A. Anandkumar, "State-specific protein–ligand complex structure prediction with a multiscale deep generative model," *Nat Mach Intell*, Feb. 2024, doi: 10.1038/s42256-024-00792-z.

[24] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996. [Online]. Available: www.aaai.org

[25] T. D. Ngo, B.-S. Hua, and K. Nguyen, "ISBNet: a 3D Point Cloud Instance Segmentation Network with Instance-aware Sampling and Box-aware Dynamic Convolution," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.00246

[26] W. Zhao, Y. Yan, C. Yang, J. Ye, X. Yang, and K. Huang, "Divide and Conquer: 3D Point Cloud Instance Segmentation With Point-Wise Binarization." [Online]. Available: https://github.com/weiguangzhao/PBNet.

[27] J. Tubiana, D. Schneidman-Duhovny, and H. J. Wolfson, "ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction," *Nat Methods*, vol. 19, no. 6, pp. 730–739, Jun. 2022, doi: 10.1038/s41592-022-01490-7.

[28] C. Lu *et al.*, "Protein-Ligand Binding Site Prediction and de Novo Ligand Generation from Cryo-EM Maps", doi: 10.1101/2023.11.16.567458.

[29] J. Desaphy, G. Bret, D. Rognan, and E. Kellenberger, “Sc-PDB: A 3D-database of ligandable binding sites-10 years on,” *Nucleic Acids Res*, vol. 43, no. D1, pp. D399–D404, Jan. 2015, doi: 10.1093/nar/gku928.

[30] J. Yang, A. Roy, and Y. Zhang, “Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment,” *Bioinformatics*, vol. 29, no. 20, pp. 2588–2595, Oct. 2013, doi: 10.1093/bioinformatics/btt447.

[31] J. Yang, A. Roy, and Y. Zhang, “BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions,” *Nucleic Acids Res*, vol. 41, no. D1, Jan. 2013, doi: 10.1093/nar/gks966.

[32] E. C. Meng *et al.*, “UCSF ChimeraX: Tools for structure building and analysis,” *Protein Science*, vol. 32, no. 11, Nov. 2023, doi: 10.1002/pro.4792.