Building a Profile Hidden Markov Model for the Kunitz-type protease inhibitor domain

Author: Martina Castellucci

For correspondence: martina.castellucci@studio.unibo.it

Master's Degree Course: Bioinformatics, University of Bologna

Laboratory of Bioinformatics I - Module 2, Academic year 2024/2025

Final submission: May 2025

Abstract

Motivation: Kunitz-type serine protease inhibitor domains (Pfam ID: PF00014) are small, structurally conserved modules involved in regulating proteolytic cascades, including inflammation, coagulation, and immune defense. They are of growing interest in host-parasite biology and therapeutic design. Despite a conserved fold stabilized by three disulfide bridges and six cysteines, sequence variability and compactness challenge reliable detection with standard sequence-based tools. This study aimed to develop a structure-guided Hidden Markov Model (HMM) for accurate identification of Kunitz domains in UniProtKB/SwissProt.

Results: The model, trained on 23 non-redundant, structurally aligned PDB entries, comprises 58 match states capturing the conserved Kunitz core. In two-fold cross-validation at an e-value threshold of 1e-6, it achieved an average MCC of 0.990, precision of 0.992, recall of 0.989, and perfect accuracy. Compared to Pfam PF00014, it showed improved specificity in borderline cases. Structural validation of false negatives revealed canonical Kunitz folds, excluded due to the model's stringent specificity.

Keywords: Kunitz domain, profile HMM, structural alignment, Cross-validation, model performance.

Supplementary materials: https://github.com/Martinaa1408/Kunitz_HMM_project/

1 Introduction

Functional domain annotation is a fundamental step in understanding protein function and evolution.

In structural bioinformatics, a domain is a compact, independently folding unit with a specific function. Its three-dimensional structure is often conserved even when sequence similarity is low, making domain-level annotation effective for detecting conserved folds like Kunitz. Among the diverse protein domain families, the Kunitz-type serine protease inhibitor represents a model system that is compact, biologically relevant, and structurally well-characterized.

1.1 The Kunitz Domain: Structural and Functional Features

Kunitz-type serine protease inhibitor domains (Pfam ID: PF00014) are short, highly conserved protein modules of ~6.5 kDa, typically composed of 50–80 amino acids. They adopt a compact α/β fold formed by a twisted β-hairpin and a short C-terminal α-helix, as illustrated in the 3D structure of BPTI (Figure 1.1), with α-helices, β-sheets, and loop regions clearly distinguished. This conformation is stabilized by three disulfide bridges connecting six conserved cysteine residues.

The specific disulfide bonding pattern—C1-C6, C2-C4, and C3-C5—is highly conserved and plays a central role in maintaining the structural integrity of the domain. The first two bridges stabilize the overall α/β framework, while the third reinforces the loop that includes the reactive site,

which is essential for inhibitory function (Figure 1.2) [1].

These domains are highly basic, often rich in lysine and arginine residues, and act as inhibitors by blocking serine protease activity through a substrate-mimicking mechanism. This involves insertion of a conserved P1 residue (typically Lys or Arg) into the active site of the target enzyme.

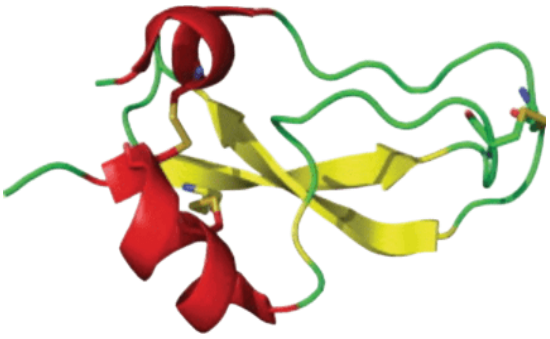


Figure 1.1: Structural representation of the Kunitz inhibitor BPTI

The 3D structure of BPTI (PDB ID: 1BPI) displays the typical Kunitz fold, with α-helices shown in red, β-sheets in yellow, and loop regions in green. The three conserved disulfide bridges, which stabilize the compact structure, are highlighted as stick models.

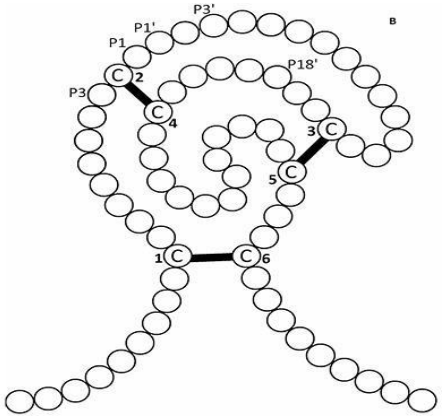


Figure 1.2: Conserved disulfide bonding pattern and folding architecture of the Kunitz domain.

Schematic folding model of a single-domain Kunitz inhibitor illustrating the six conserved cysteine residues (C1–C6) and the formation of three disulfide bridges (C1–C6, C2–C4, C3–C5), which stabilize the characteristic α/β fold of the domain. The diagram also highlights the protease-binding loop (P3–P3') containing the reactive site, typically inserting into the active site of target serine proteases.

Kunitz domains are broadly distributed across metazoans and some microorganisms and are involved in essential physiological processes including blood coagulation, inflammation, synaptic function, and immune regulation [2]. They are found in proteins such as BPTI (Bovine Pancreatic Trypsin Inhibitor), APP (Amyloid Precursor Protein), TFPI (Tissue Factor Pathway Inhibitor), and a variety of venom neurotoxins and anticoagulants. In parasitic helminths like *Schistosoma mansoni* and *Fasciola hepatica*, Kunitz-type domains help modulate host immune responses, facilitating immune evasion and supporting parasite survival. These properties underscore their biomedical and therapeutic potential, both as targets and tools for drug development.

1.2 Hidden Markov Models and Their Relevance in Remote Homology

Profile Hidden Markov Models (HMMs) are probabilistic frameworks widely employed to describe the conservation and variability of protein domains based on multiple sequence alignments. Each HMM encodes match (M), insert (I), and delete (D) states, which model the probability of observing a residue, an insertion, or a deletion at each position in the alignment—allowing for a detailed and position-specific description of sequence variability (Figure 2).

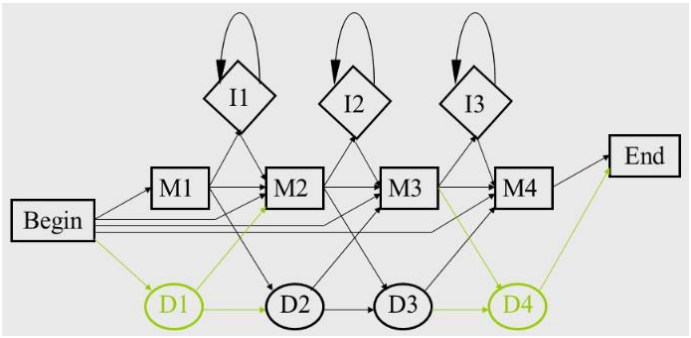


Figure 2: Profile Hidden Markov Model (HMM) architecture

Simplified representation of the structure-based profile Hidden Markov Model (HMM), showing the first four match states (M_1–M_4) along with their corresponding insert (I) and delete (D) states. Arrows indicate possible state transitions, following the standard left-to-right architecture typically used to model conserved and variable regions of the Kunitz domain.

Unlike BLAST, which relies on local pairwise similarity, HMMs enable the detection of remote homologs by capturing position-specific substitution patterns and structural constraints [3]. Moreover, profile-HMMs can

recognize complex patterns of conservation and indel events that are typically missed by alignment-based tools, which are limited to direct sequence similarity rather than evolutionary modeling.

This advantage becomes especially important in the so-called "twilight zone" of sequence identity (20–30%), where homology is often undetectable by traditional methods. In such cases, training HMMs on structurally aligned sequences—rather than on sequence alignments alone—enhances both specificity and biological interpretability. Structure-based models focus on conserved core residues essential for domain stability and function, filtering out noise from poorly aligned or redundant regions. These models are generative and compute scores as log-odds ratios between the probability of the observed sequence under the trained model and a background null model. This ensures that domain predictions are statistically grounded and biologically meaningful. Several studies have confirmed that structure-informed HMMs significantly improve the detection of functionally relevant domains in complex or highly divergent protein families [4].

1.3 Aim of the Project

This project aims to construct a structure-guided profile Hidden Markov Model (HMM) for the accurate detection of Kunitz-type protease inhibitor domains in protein sequences. The classification strategy relies on experimentally solved protein structures annotated with the Pfam domain (PF00014), which are filtered to remove redundancy and structurally aligned to preserve the conserved three-dimensional fold. This alignment serves as the basis for building a probabilistic profile that captures the core features of the domain.

The model is tested on curated datasets of annotated sequences to evaluate its ability to generalize beyond the training data. Performance is assessed through cross-validation using standard metrics such as accuracy, precision, recall, and Matthews Correlation Coefficient (MCC). Threshold selection and classification decisions are guided by score distributions and ROC curves.

A detailed diagram of the full pipeline, including dataset construction, filtering criteria, model building, and evaluation, is provided in Figure 3 and fully reproduced in the project GitHub repository.

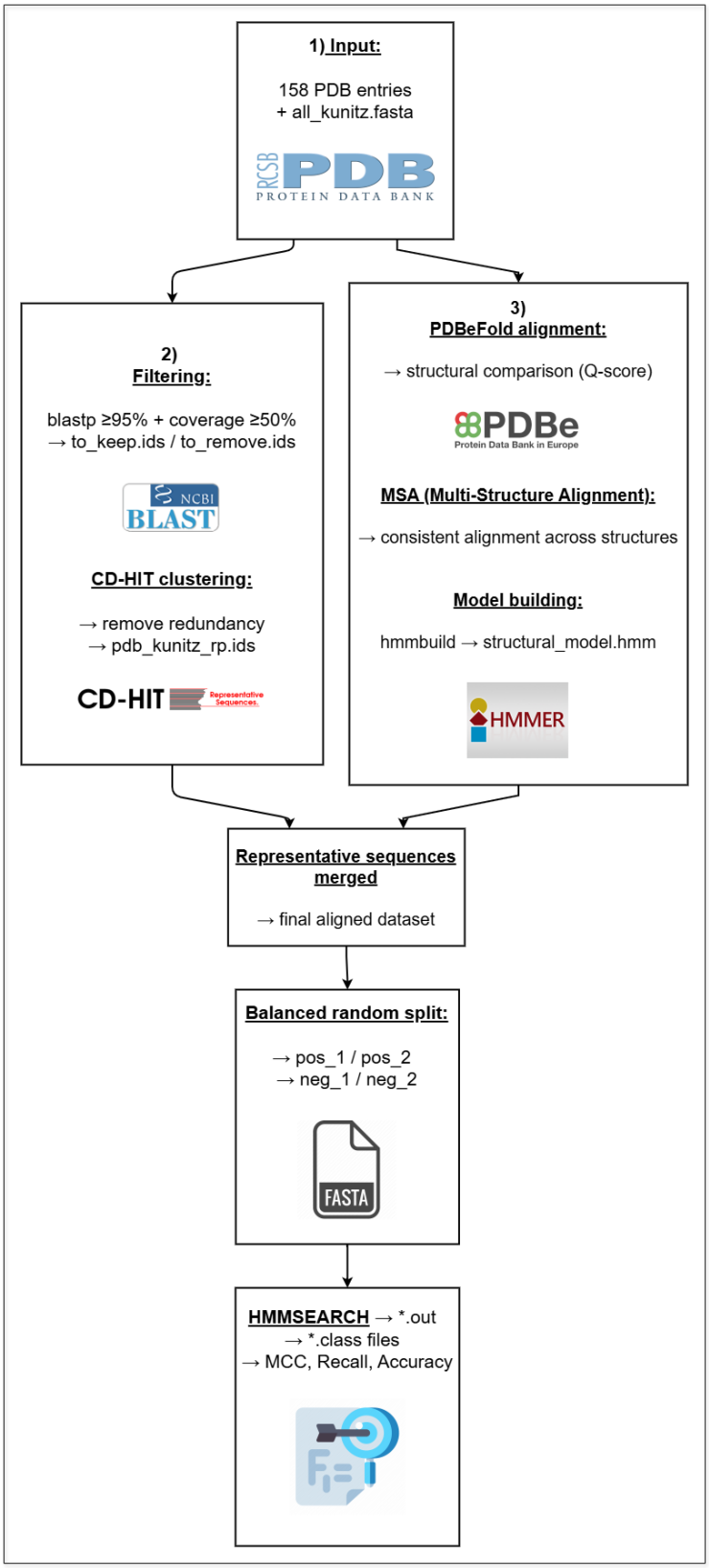


Figure 3. Structure-Guided HMM Construction and Evaluation Pipeline

Schematic representation of the computational workflow used for the structure-based construction of a profile Hidden Markov Model (HMM) for the detection of Kunitz-type domains.

2. Materials and Methods

2.1 Data Collection and Preprocessing

The complete Swiss-Prot protein dataset was downloaded in FASTA format from the UniProt database [5].

To identify sequences containing the Kunitz-type protease inhibitor domain (Pfam ID: PF00014), annotations from InterPro [6] were used.

This initial dataset included 395 reviewed entries: 18 from human proteins and 377 from non-human sources. These were saved as human_kunitz.fasta, nothuman_kunitz.fasta, and combined into all_kunitz.fasta. In parallel, a list of 158 Protein Data Bank (PDB) entries annotated with PF00014 was retrieved by exporting a custom CSV report (rcsb_pdb_custom_report.csv) from the RCSB PDB [7]. Filters were applied during the query to include only entries resolved by X-ray crystallography with a resolution of 3.5 Å or better and a chain length between 45 and 80 amino acids. The resulting 158 high-quality chains were used directly for downstream analysis.

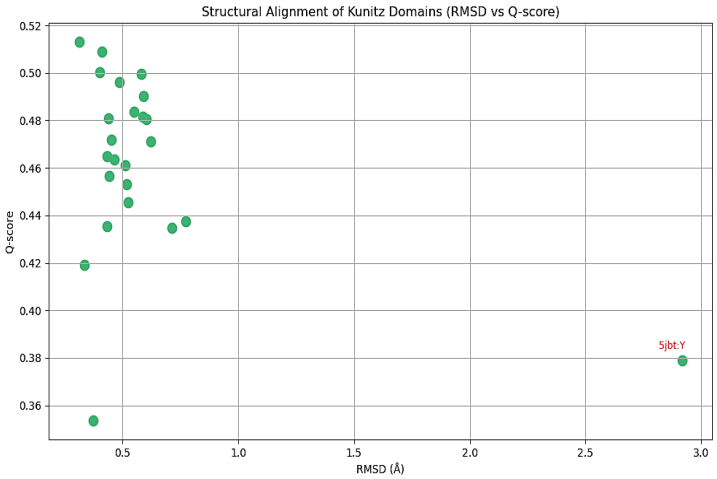
2.2 Structural Dataset Construction

To eliminate redundancy, the filtered PDB sequences were clustered using CD-HIT with a 90% identity threshold (-c 0.9), yielding 25 initial clusters. Representative sequences were selected using the clstr2txt.pl script and manually reviewed to assess structural completeness and domain integrity. After this quality control step, 23 non-redundant sequences were retained (pdb_kunitz_rp.fasta) for structural alignment and model construction.

2.3 Structural Alignment and Profile-HMM Construction

The 23 curated sequences were aligned using PDBeFold [9], a tool for pairwise and multiple structural alignments based on 3D similarity. The resulting alignment file (pdb_kunitz_rp.ali) was manually reviewed and reformatted (pdb_kunitz_rp_formatted.ali). To assess the quality of the structural alignment and identify potential outliers, RMSD and Q-score values were extracted from PDBeFold and visualized in a scatter plot (Figure 4).

An initial set of 25 non-redundant representative sequences was obtained after CD-HIT clustering. One entry (2ODY_E) was excluded prior to structural alignment due to excessive length, leaving 24 sequences aligned with PDBeFold. Among these, one structure (5jbt:Y) stood out with a low residue count (38 amino acids) and a high RMSD value (2.92 Å), indicating poor structural compatibility with the rest of the set.



$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} ||Xi - Yi||^2}$$

$$Q - score = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{1 + (\frac{di}{d0})^2}$$

Figure 4: Structural similarity among Kunitz-domain candidates evaluated via RMSD and Q-score.

The scatter plot (top) displays 24 PDB structures aligned using PDBeFold. One entry (5jbt:Y) is highlighted as an outlier with high RMSD (2.92 Å) and low Q-score; this structure was excluded from HMM training due to its unusually short length (Nres = 38). The bottom panel reports the mathematical definitions of RMSD and Q-score used for assessing structural similarity.

Based on these criteria, it was excluded from further analysis, resulting in a final selection of 23 structurally consistent sequences used for HMM training.

This alignment served as input for the HMMER suite [10], specifically hmmbuild, which generated a profile Hidden Markov Model (HMM) consisting of 58 match states. These states reflect the structurally conserved core of the Kunitz domain.

To visualize the level of conservation at each aligned position, sequence logos were generated using WebLogo and Skylign [11][12].

2.4 Generation of Positive and Negative Sets

To avoid overlap between training and testing data, a BLASTp [13] search was performed between the 23 PDB sequences and the full Kunitz dataset. Sequences with 95% identity or more and an aligned region of at least 50 amino acids were excluded. The remaining 366 non-redundant Kunitz sequences were retained for evaluation and saved as ok_kunitz.fasta.

Negative sequences were selected from the full Swiss-Prot database (573,230 entries) by excluding all known Kunitz-domain proteins. This yielded a pool of 572,835 negative candidates. Two subsets of 286,417 (floor approximation) sequences each were randomly sampled and saved as neg_1.ids and neg_2.ids

2.5 Cross-Validation Setup

The 366 positive sequences were evenly split into two subsets (pos_1.ids and pos_2.ids). Each was paired with a large, randomly sampled set of negative sequences (neg_1.ids and neg_2.ids), resulting in two highly imbalanced folds for cross-validation. Stratified splitting of positives ensured class consistency across folds, while the large negative sets enabled robust evaluation of false positive rates. All FASTA files (pos_1.fasta, pos_2.fasta, neg_1.fasta, neg_2.fasta) were generated using custom scripts, with strict separation between training and test data.

2.6 Domain Search with HMMER

The trained HMM was applied to each test set using hmmsearch with the --max and -Z 1000 flags to ensure consistent statistical calibration. Output files were processed to classify sequences as positive or negative based on their full-sequence e-values, and saved in .class format for downstream performance analysis.

2.7 Performance Evaluation and Threshold Analysis

Model performance was assessed across a range of e-value thresholds (where the e-value represents the expected number of random matches with equal or better score) using a custom Python script (performance.py). The following six evaluation metrics were calculated based on the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN):

- **Accuracy:** the proportion of correctly classified sequences out of the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** the proportion of true positives among all sequences predicted as positive.

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

- **Recall (or True Positive Rate):** the proportion of true positives correctly identified.

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

- **Matthews Correlation Coefficient (MCC):** a balanced measure of binary classification quality.

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

- **False Positive Rate (FPR):** the proportion of actual negative sequences that were incorrectly predicted as positive by the model.

$$FPR = \frac{FP}{FP + TN}$$

- **False Negative Rate (FNR):** the proportion of actual positive sequences that were missed by the model (i.e., classified as negative).

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

In addition, **ROC curves** were generated to visualize the trade-off between sensitivity and specificity across different threshold values.

2.8 Comparison with Reference Methods

To benchmark model performance, the structure-guided HMM was compared to the Pfam reference model (PF00014.hmm) using the same datasets and hmmsearch settings (--max, -Z 1000). Tabular outputs were parsed to extract e-values and matched with class labels from the FASTA files. The resulting .class files were processed with the same performance.py script to compute classification metrics across multiple thresholds.

3. Results and Discussion

3.1 Model Architecture: Structural HMM vs Pfam Model

The custom profile HMM developed in this project was built from 23 structurally aligned, non-redundant PDB entries containing the Kunitz domain, resulting in 58 match states that capture the conserved structural core of the domain. In contrast, the Pfam reference model (PF00014.hmm) includes 53 match states derived from a multiple sequence alignment of 99 manually curated seed sequences from diverse organisms. While the Pfam model emphasizes sensitivity and phylogenetic breadth—capturing both conserved and variable positions to detect distant homologs—the structural model focuses on specificity, incorporating only residues conserved across high-resolution structures, including key disulfide-bonding cysteines and aromatic or glycine residues critical for domain stability. When evaluated on the same datasets and using identical hmmsearch parameters at an e-value threshold of 1e-6, the structural HMM achieved robust performance with high precision and MCC in both folds, despite a small number of false positives in one of them. The Pfam model, whose results are shown in Table 1, maintained perfect precision and consistent MCC across folds. This comparison highlights the complementarity of the two approaches: Pfam offers wide coverage and stable generalization, while the structural HMM delivers interpretable and high-specificity annotations.

Table 1: Performance metrics for the PFAM HMM on Set 1 and Set 2 at e-value threshold 10^{-6}.

Metric	Set 1	Set 2
True Positives (TP)	181	181
False Negatives (FN)	2	2
False Positives (FP)	0	0
True Negatives (TN)	286417	286417
Matthews Correlation Coefficient (MCC)	0.9945	0.9945
Precision (PPV)	1.000	1.000
Recall (TPR)	0.9891	0.9891
Accuracy (Q2)	1.000	1.000

3.2 Sequence and Structural Conservation Analysis

To confirm that the model captures biologically relevant features, two complementary analyses were performed: sequence conservation and 3D structural alignment. Sequence conservation analysis was performed using both Skylign (Figure 5.1) and WebLogo (Figure 5.2) to visualize the information content of the 58 match states in the structural HMM.

Both representations highlighted key features of the Kunitz domain, notably the six conserved cysteines involved in disulfide bonding (C1–C6, C2–C4, C3–C5), along with several aromatic (F, Y) and glycine residues critical for structural stability.

The WebLogo output provides a compact view of residue frequency across the alignment, emphasizing highly conserved positions along the domain sequence. In contrast, Skylign integrates position-specific scoring data, offering a probabilistic perspective that enhances interpretation of subtle conservation patterns and variation probabilities. While the two logos differ in style, they consistently confirm that the model captures the structurally and functionally constrained core of the Kunitz domain.

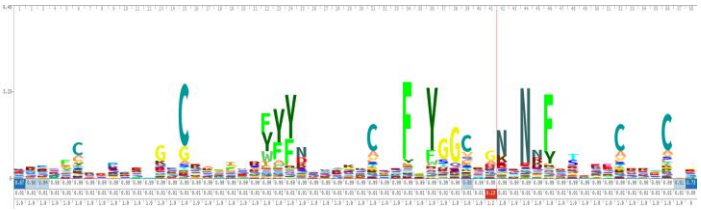


Figure 5.1: Sequence logo of the structural HMM generated with Skylign.

This logo displays the 58 match states of the profile HMM, with conserved cysteine residues and other structurally relevant positions clearly visible. The height and color of each amino acid represent the information content and posterior probability at each position.

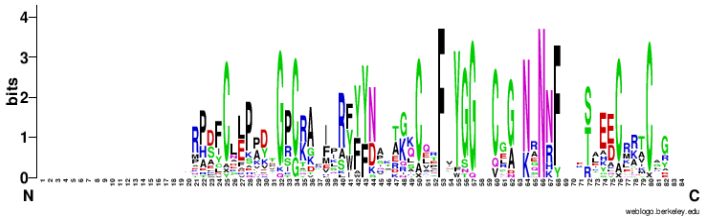


Figure 5.2: Sequence logo of the same alignment generated with WebLogo.

This representation emphasizes residue conservation based on relative amino acid frequency. The most conserved columns correspond to the structurally and functionally essential core of the Kunitz domain.

At the optimal threshold of 1e-6—selected based on maximal MCC—the model failed to classify four sequences (two per fold) as positives, despite having a true label of 1. Their e-values were slightly above the decision threshold, leading to their exclusion. These represent borderline false negatives:

sequences annotated as Kunitz (label = 1) but not detected due to their slightly weaker statistical signal. Their exclusion reflects the model's strict specificity, though their annotation and structure suggest they may contain valid Kunitz domains.

To verify their structural validity, 3D superposition was performed using ChimeraX [14] and AlphaFold-predicted models, comparing each sequence to the Kunitz domain as represented in the BPTI–trypsin complex (PDB: 3TGI).

The structural superpositions are shown in Figure 6.
The two sequences with available models, D3GGZ8 and Q8WPG5, showed strong structural similarity:

• D3GGZ8: RMSD = 0.819 Å, alignment score = 76.9

• Q8WPG5: RMSD = 0.889 Å, alignment score = 101.2

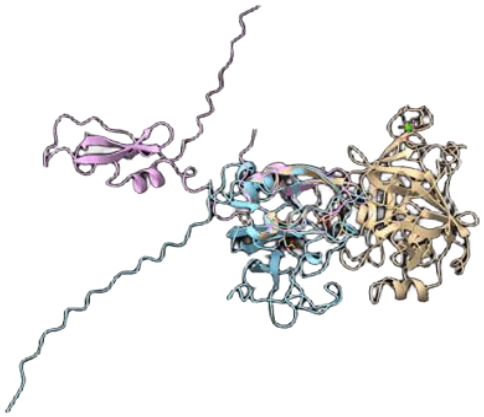


Figure 6: Structural superposition of false negative predictions with the canonical Kunitz fold

Superposition of D3GGZ8 (Light Blue) and Q8WPG5 (Lilac) onto the canonical Kunitz inhibitor BPTI (PDB: 3TGI, Beige), showing strong structural similarity with RMSD < 0.9 Å.

Both maintained the hallmark features of the Kunitz fold, including the α/β topology and conserved disulfide bond pattern. These results indicate that their exclusion was due not to a lack of structural relevance, but rather to the model's stringent design.
This finding supports the interpretation that the HMM prioritizes specificity over recall, intentionally limiting the risk of false positives.
Sequences near the decision boundary—such as these—should be interpreted cautiously, and where necessary, confirmed through structural validation in high-sensitivity annotation pipelines.

3.3 Cross-Validation Performance and Confusion Matrices

To rigorously evaluate the classification performance of the structural HMM, we adopted a two-fold cross-validation strategy. The dataset was split into two non-overlapping subsets, which alternated as training and test sets. This approach ensured that each fold was evaluated on truly unseen sequences, allowing us to assess the model's generalization capacity while minimizing the risk of overfitting.
The classification task was framed as a binary problem: each sequence was labeled as either positive (containing a Kunitz domain) or negative (lacking the domain). At the optimal e-value threshold of 1e-6, selected based on maximum Matthews Correlation Coefficient (MCC), the model achieved highly consistent and robust results across folds. In Fold 1, the model correctly classified 286417 negatives and 181 positives, with only 2 false negatives and no false positives, as shown in Figure 7. In Fold 2, performance remained strong with 286414 true negatives, 181 true positives, 2 false negatives, and only 3 false positives, illustrated in Figure 8.
These confusion matrices summarize the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in each test set. The corresponding evaluation metrics—reported in Table 2—confirm excellent model performance, with MCC values of 0.9945 and 0.9864, and precision and recall exceeding 98% in both folds.

Table 2: Performance metrics at e-value threshold 10^{-6} for both validation folds.

Metric	Fold 1	Fold 2
True Positives (TP)	181	181
True Negatives (TN)	286417	286414
False Positives (FP)	0	3
False Negatives (FN)	2	2
Accuracy (Q2)	1.000	1.000
Precision (PPV)	1.000	0.9837
Recall (TPR)	0.9891	0.9891
Matthews Correlation Coefficient (MCC)	0.9945	0.9864
False Positive Rate (FPR)	0.000000	0.0000105
False Negative Rate (FNR)	0.0110	0.0110

Despite the strong class imbalance (over 286,000 negatives versus 183 positives per fold), the structural HMM retained a high level of discrimination. The minimal number of misclassifications demonstrates the model's ability to maintain high specificity and sensitivity, making it suitable for reliable domain identification even under stringent conditions.

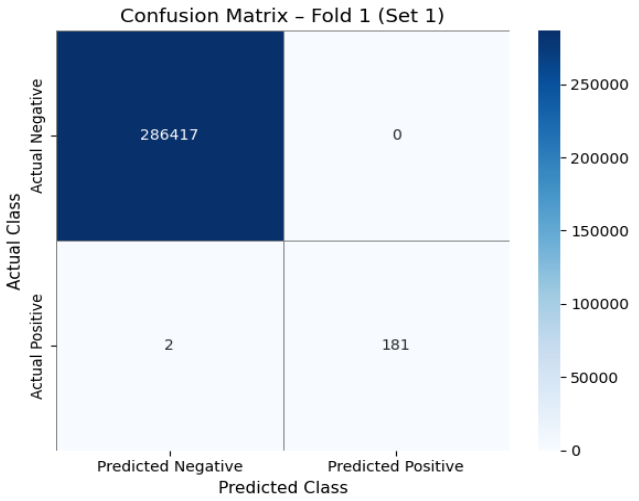


Figure 7. Confusion matrix for Fold 1 model trained on subset B and tested on subset A.

No false positives, 2 false negatives, 286417 true negatives, and 181 true positives (at the e-value threshold of 1e-6).

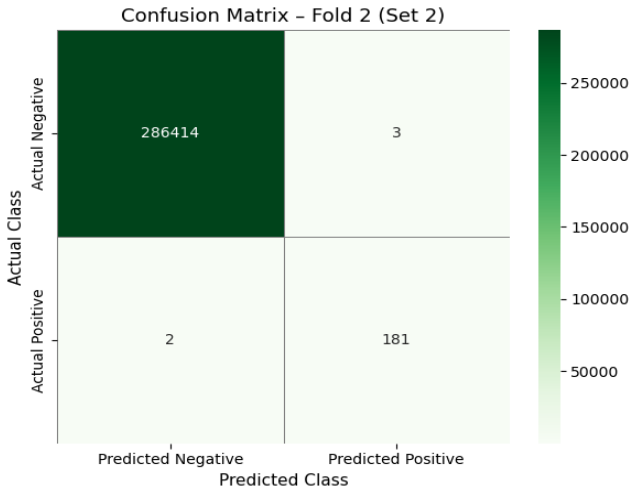


Figure 8. Confusion matrix for Fold 2 model trained on subset A and tested on subset B .

2 false negatives, 3 false positives, 286414 true negatives, and 181 true positives (at the e-value threshold of 1e-6).

As further illustrated in the confusion matrices (Figures 7 and 8) and the summary metrics in Table 2, the model demonstrated strong and consistent classification performance across both folds. While a small number of misclassifications occurred—specifically, two false negatives per fold and three false positives in Fold 2—the overall results confirm the model's robustness and its ability to generalize effectively, even in the presence of severe class imbalance.

3.4 Threshold Optimization and MCC Trends

To investigate how model performance varies across decision thresholds, the structural HMM was evaluated over a wide range of e-values, from 1e-10 (highly stringent) to 1e-1 (highly permissive). The Matthews Correlation Coefficient (MCC) was selected as the principal metric, given its ability to provide a balanced assessment even in the presence of strong class imbalance.

As shown in Figure 9, MCC values remained consistently high across a broad range of thresholds. In Set 1, MCC plateaued around 0.9918 from 1e-10 up to 1e-6, while Set 2 exhibited a slight increase, reaching its peak at 1e-6. Beyond this point, more permissive thresholds such as 1e-4 and above led to a gradual decline in performance, primarily due to the increase in false positives. At 1e-2, the average MCC dropped below 0.98, and performance degraded further at 1e-1.

The best-performing threshold, marked in red in Figure 9, was therefore 1e-6, where both validation folds achieved their highest combined MCC scores (0.9945 and 0.9864, respectively). This threshold reflects the optimal trade-off between sensitivity and specificity and was used for all subsequent performance analyses.

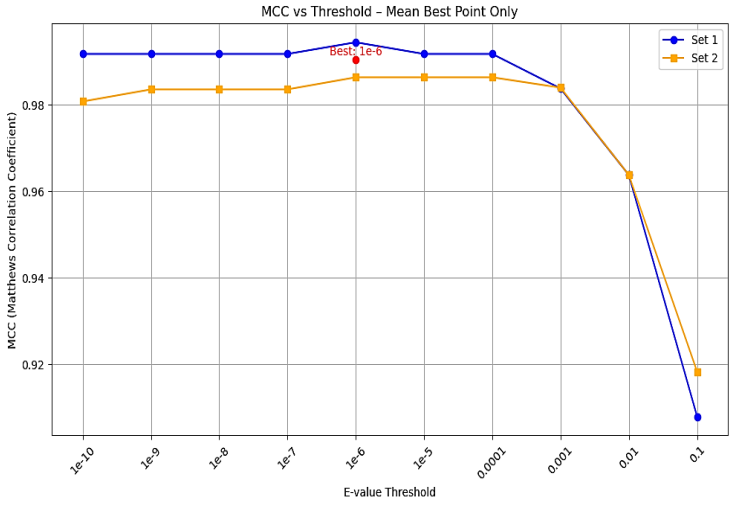


Figure 9: Matthews Correlation Coefficient (MCC) across different e-value thresholds for Fold 1 and Fold 2

The best threshold (1e-6) is highlighted in red. MCC remains stable across stringent thresholds and declines at more permissive values due to increased false positives.

3.5 ROC Curve Analysis

To further assess the model's discriminative ability, Receiver Operating Characteristic (ROC) curves were generated using the full-sequence e-value scores as continuous predictors. The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate (FPR) across all possible thresholds, offering a comprehensive view of classifier performance.

As shown in Figure 10, both validation folds produced ideal ROC curves that closely follow the upper-left boundary of the plot, reflecting excellent separation between classes. The corresponding Area Under the Curve (AUC) is 1.000 for both folds, confirming that the model consistently assigns highly discriminative scores to positive and negative sequences.

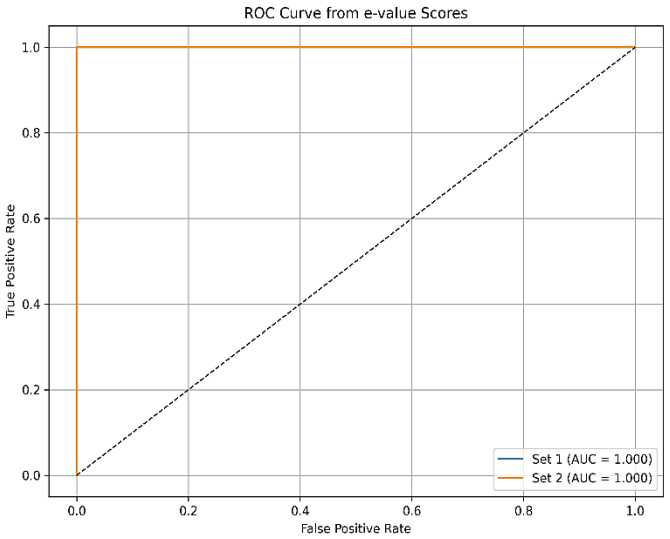


Figure 10: ROC curves for Fold 1 and Fold 2, derived from full-sequence e-value scores

Both curves yield an AUC of 1.000, indicating perfect separability between true positives and true negatives across thresholds.

4. Conclusion

This work presents a structure-guided profile Hidden Markov Model (HMM) specifically developed for the precise identification of Kunitz-type serine protease inhibitor domains. By incorporating only structurally conserved residues extracted from 23 non-redundant PDB entries, the model captures the functional and evolutionary core of the domain while minimizing the influence of sequence variability. This redundancy-aware design, combined with a deterministic thresholding strategy, ensures high interpretability and robustness, especially in the context of compact, cysteine-rich domains. Compared to traditional sequence-based models such as PFAM.hmm, the structural HMM achieved similar overall performance at an e-value threshold of 10^{-6}, while offering greater specificity. Although Pfam maintained perfect precision across validation folds, our model showed strong generalization, especially on borderline cases. Structural superposition with AlphaFold-predicted models confirmed Kunitz-like folds in sequences misclassified at stricter thresholds (e.g., 1e-10), highlighting the limitations of overly conservative filters that ignore structural evidence. Two-fold cross-validation proved well-suited for this task, balancing computational efficiency with robust evaluation on independent data.

However, a key limitation in the comparison with Pfam is that it includes all known Kunitz sequences, including the 23 PDB entries used for training our model. This creates an overlap between training and test sets. To avoid artificially inflated performance, future evaluations should split the Pfam dataset in half—using one part for model construction and the other for testing—ensuring that each model is evaluated on fully independent data. This approach demonstrates the value of integrating three-dimensional information into probabilistic models for domain detection. The use of profile HMMs in this context proves especially effective, enabling biologically meaningful predictions with high reproducibility. The alignment strategy—based on structural constraints—could be further refined by exploring tools such as MUSTANG [15], and by incorporating more accurate multiple sequence alignment (MSA) algorithms such as MAFFT [16], T-Coffee [17], and ProbCons [18], particularly in regions of high sequence divergence, despite the increased computational cost.

Importantly, the model focuses on identifying conserved domains—structurally stable and functionally autonomous units—rather than short local motifs, reinforcing the biological relevance of the predictions. Beyond performance metrics, the approach is generalizable and could be extended to other conserved yet variable domain families, particularly in poorly annotated or taxonomically diverse datasets. When integrated with curated domain databases and functional prediction tools, this strategy lays the foundation for hybrid annotation pipelines that combine structural information with probabilistic modeling to improve accuracy and interpretability. In conclusion, this study highlights the benefits of embedding structural knowledge into computational domain models. Structure-guided HMMs provide a scalable, interpretable, and biologically sound method for protein function annotation, with wide-ranging applications in functional genomics, evolutionary biology, and

biomedical research.

5. References

[1] Mishra, M (2020). “Evolutionary aspects of the structural convergence and functional diversification of kunitz- domain inhibitors.” In: J Mol Evol. 88. url: https://doi.org/10.1007/s00239-020-09959-9

[2] de Magalhães, M.T.Q. et al. (2018). *Serine protease inhibitors containing a Kunitz domain: their role in modulation of host inflammatory responses and parasite survival*. Microbes and Infection, 20(9–10):606–609. https://doi.org/10.1016/j.micinf.2018.01.003

[3] Yoon, B.-J. (2009). *Hidden Markov Models and Their Applications in Biological Sequence Analysis*. Current Genomics, 10(6):402–415. https://doi.org/10.2174/138920209789177575

[4] ScienceDirect Topics. *Kunitz Domain – Neuroscience*. https://www.sciencedirect.com/topics/neuroscience/kunitz-domain

[5] "UniProt: The Universal Protein Knowledgebase in 2023." UniProt, 2023, https://www.uniprot.org/

[6] Blum, Markus, et al. "InterPro." Version 87.0, EMBL-EBI, 2021, https://www.ebi.ac.uk/interpro

[7] Burley, Stephen K., et al. "Protein Data Bank (PDB)." RCSB, 2021, https://www.rcsb.org/

[8] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658–9. doi:10.1093/bioinformatics/btl158

[9] Krissinel, Evgeny, and Kim Henrick. "PDBeFold." Version 2.56, European Bioinformatics Institute, 2004, http://www.ebi.ac.uk/msd-srv/ssm/

[10] Sean R Eddy. Hmmer3: a new generation of sequence homology search software. Bioinformatics, 27(17):2957–2958, 2011.

[11] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res*. 2004;14(6):1188–1190. doi:10.1101/gr.849004

[12] O'Donnell, T. J., Rubinsteyn, A., & Laserson, U. (2015). *Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile models*. Retrieved May 2024, from https://skylign.org/

[13] Altschul, Stephen F., et al. "BLAST." Version 2.11.0, National Center for Biotechnology Information, 1990, https://blast.ncbi.nlm.nih.gov/Blast.cgi

[14] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2021). *UCSF ChimeraX: Structure visualization for researchers, educators, and developers*. Protein Science, 30(1), 70–82. https://doi.org/10.1002/pro.3943

[15] Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., & Lesk, A. M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*, 64(3), 559–574. https://doi.org/10.1002/prot.20921

[16] Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

[17] Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042

[18] Do, C. B., Mahabhashyam, M. S., Brudno, M., & Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15(2), 330–340. https://doi.org/10.1101/gr.2821705