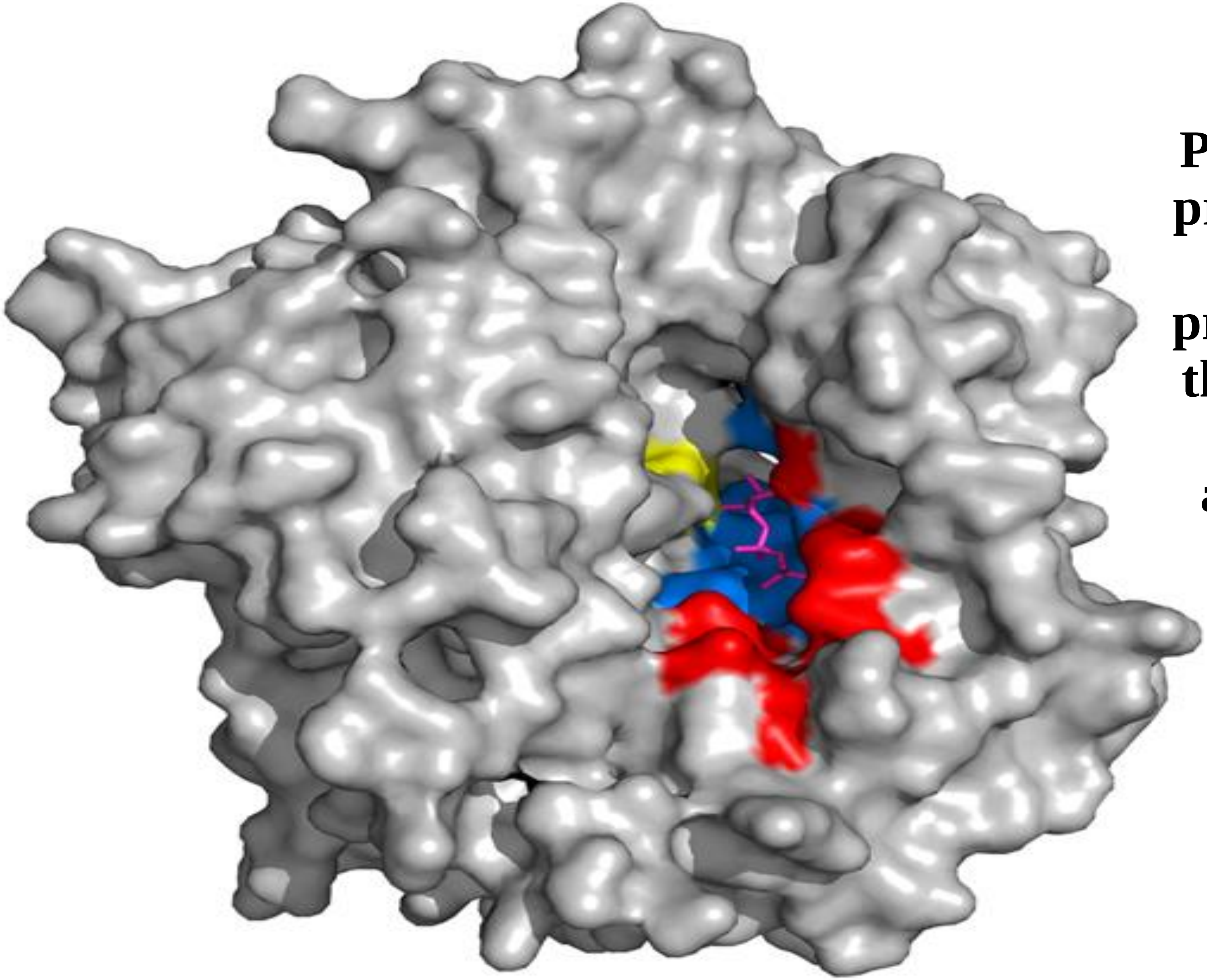


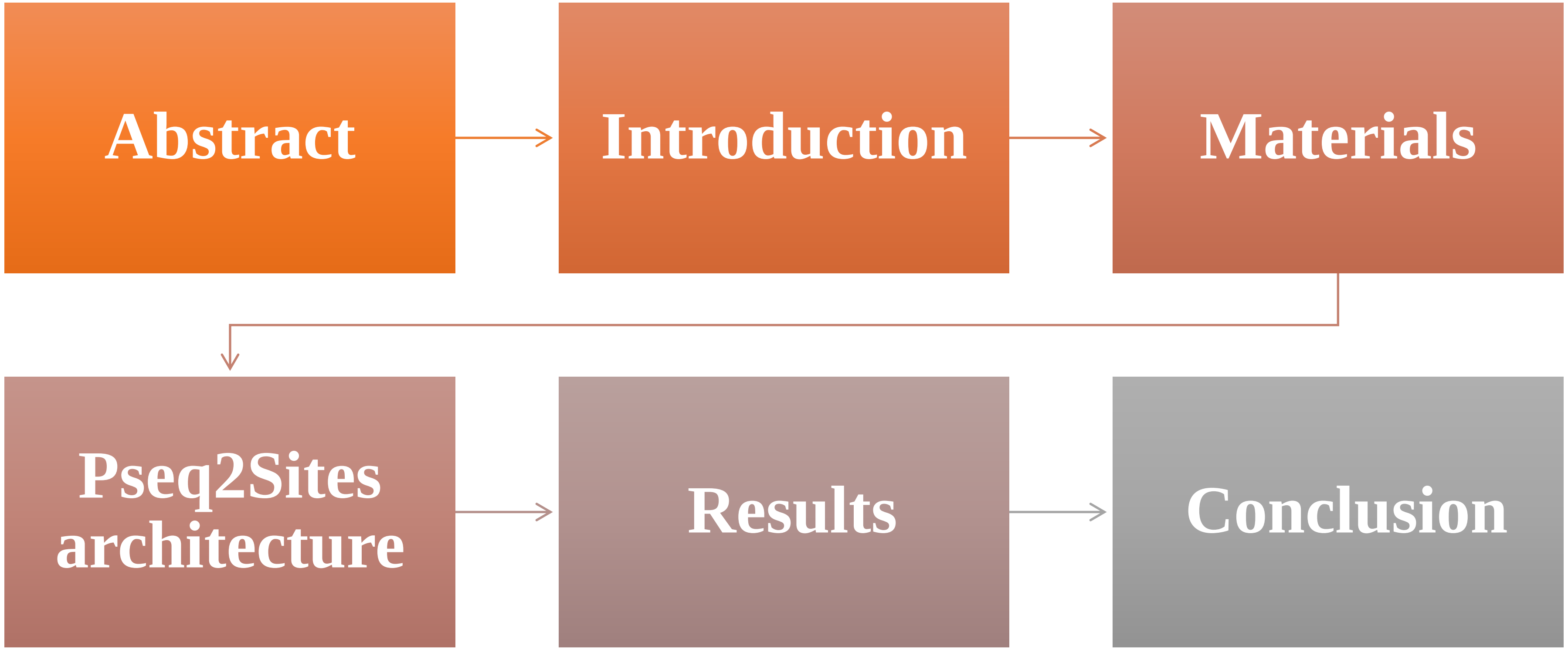
Pseq2Sites: Enhancing protein sequence-based ligand binding-site prediction accuracy via the deep convolutional network and attention mechanism

Asal Rabiee

610400053

School of Biology,
College of Science,
University of Tehran,
Tehran, Iran

Overview



Abstract

Introduction

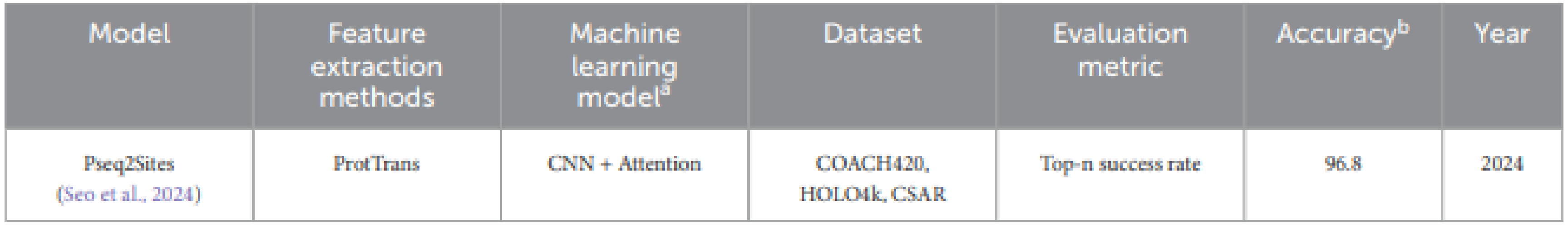
Materials

Pseq2Sites architecture

Results

Conclusion

Abstract



Model	Feature extraction methods	Machine learning model[a]	Dataset	Evaluation metric	Accuracy[b]	Year
Pseq2Sites (Seo et al., 2024)	ProtTrans	CNN + Attention	COACH420, HOLO4k, CSAR	Top-n success rate	96.8	2024

1: Understand the Protein Sequence

Uses ProtTrans (a protein language model) to convert each amino acid into a "smart code" (embedding) that captures its role in the protein.

2: Spot Local Patterns

A 1D-CNN (pattern detector) scans these codes to find small, important motifs (e.g., binding pockets) in short segments of the sequence.

3: Capture Long-Range Relationships

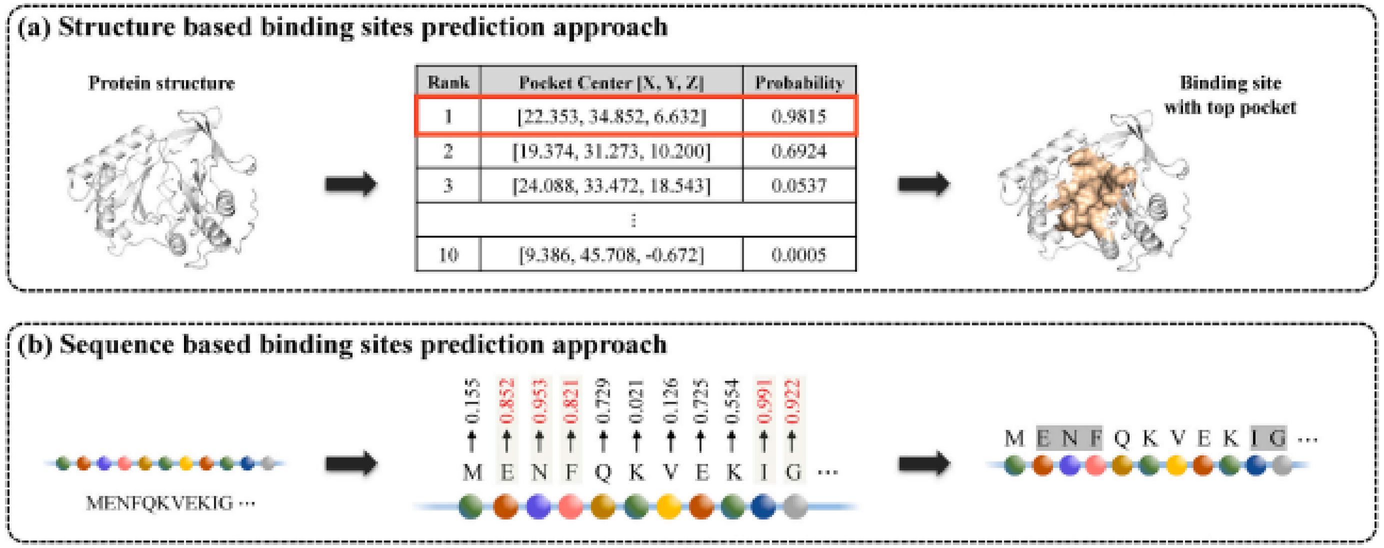
A position-based attention mechanism looks at the *entire sequence* to see how distant amino acids might work together to form a binding site.

4: Predict Binding Sites

Combines local (CNN) and global (attention) insights to predict which amino acids are part of binding sites.

Introduction

Ligand binding sites are specific regions on a protein where ligands interact through non-covalent forces, typically located within cavities or pockets of the protein's 3D structure. The amino acids forming the pocket define its chemical and functional properties. Understanding binding sites is crucial for molecular docking, ligand affinity prediction, and drug design. Binding sites are typically defined based on the distance between heavy atoms of the protein and ligand, though the specific distance thresholds vary across studies.


(a) Structure based binding sites prediction approach
Protein structure
Binding site with top pocket

Rank	Pocket Center [X, Y, Z]	Probability
1	[22.353, 34.852, 6.632]	0.9815
2	[19.374, 31.273, 10.200]	0.6924
3	[24.088, 33.472, 18.543]	0.0537
⋮		
10	[9.386, 45.708, -0.672]	0.0005

(b) Sequence based binding sites prediction approach
MENFQKVEKIG ···
0.155 0.852 0.953 0.821 0.729 0.021 0.126 0.725 0.554 0.991 0.922
M E N F Q K V E K I G ···
M E N F Q K V E K I G ···

- *Structure-Based Approaches:* These methods predict binding sites by first identifying and ranking candidate pocket regions (e.g., pocket centers) on the protein's 3D structure. Residues within a certain distance from the predicted pocket centers are then extracted and classified as binding sites. Tools such as Fpocket, P2Rank, DeepSurf, and DeepPocket follow this strategy and demonstrate strong performance. However, their success heavily depends on the availability of accurate and high-quality 3D protein structures.

- *Sequence-Based Approaches:* To overcome the limitation of missing 3D structures, sequence-based methods predict binding sites directly from the protein's amino acid sequence. They transform the raw sequence into structurally informative features—such as multiple sequence alignments (MSA), position-specific scoring matrices (PSSM), predicted secondary structures, or solvent accessibility. Deep learning models like DeepCSeqSite (CNN-based), BiRDS (ResNet-based), and HoTS (transformer-based) are then applied to determine whether each residue is a binding site. These approaches enable binding site prediction even when 3D structure data is unavailable.

Introduction

Protein-ligand interactions are central to biological functions and drug discovery, making the identification of ligand binding sites critically important. However, experimentally determining these sites is often time-consuming and costly. To address this, many computational approaches based on protein 3D structures—such as Fpocket, Q-SiteFinder, Cofactor, BioLip, P2Rank, DeepSite, and DeepPocket—have been developed. While these structure-based methods generally achieve strong binding site prediction performance, they rely heavily on the availability of accurate 3D protein structures, which limits their practical application.

To overcome this limitation, sequence-based prediction methods have gained increasing attention. Methods like LigandRFs, DeepCSeqSite, and HoTS have demonstrated significant improvements, but they still tend to underperform compared to 3D structure-based methods. Key challenges include noise introduced by predicted structural features and difficulties in capturing long-distance dependencies between amino acid residues.

Introduction

To address these challenges, the authors propose Pseq2Sites, a novel deep learning model that predicts ligand binding sites directly from protein sequences. Pseq2Sites leverages:

- Protein sequence embeddings from ProtTrans to represent each residue with rich contextual information,
- A 1D Convolutional Neural Network (CNN) to capture local patterns along the sequence, and
- A position-based attention mechanism to effectively model long-range interactions between residues that may be distant in sequence but close in 3D structure.

With this design, Pseq2Sites effectively models both local and global dependencies and achieves outstanding binding site prediction performance—even for proteins with low sequence similarity to the training dataset. As deep-learning methods like Pseq2Sites continue to evolve, the performance gap between sequence-based and structure-based approaches is rapidly narrowing.

Materials: Datasets and preprocessing

The study used binding-site data from the scPDB v.2017 and PDBbind v.2020 databases for model training. The scPDB database contains 17,594 druggable binding sites from the Protein Data Bank (PDB), while the PDBbind database provides 19,443 experimentally identified protein-ligand complexes. Preprocessing steps included filtering out complexes with protein sequences longer than 1500 amino acids and those with ligands that couldn't be parsed using tools like Openbabel and RDKit. After preprocessing, 16,818 complexes from scPDB and 17,187 from PDBbind were retained. In total, 28,728 protein-ligand complexes were used for training.

For model evaluation, four external test datasets were used: COACH420, HOLO4K, and two subsets of CSAR NRC-HiQ. Data preprocessing followed the same rules, excluding complexes with standard amino acids and ensuring no overlap between training and test data.

Each is divided into 2 test subsets:

1. Unseen-protein test dataset: Proteins that are structurally different from the training proteins (similarity < 40%) -> to evaluate performance on unseen proteins.
2. Unseen Complex: New protein-ligand pairs that are not in the training data.

Materials: Protein structure similarity

To rigorously evaluate the performance of binding-site prediction, unseen protein test datasets were created by assessing the structural similarity between proteins in the training and test sets. The Template Modeling (TM)-Score was used to quantify structural similarity, with scores ranging from 0 to 1, where a higher score indicates greater similarity. The highest TM-score between any chain structures was selected when proteins had multiple chains. The unseen protein test datasets consisted of proteins that had ≤ 40% structural similarity to those in the training set.

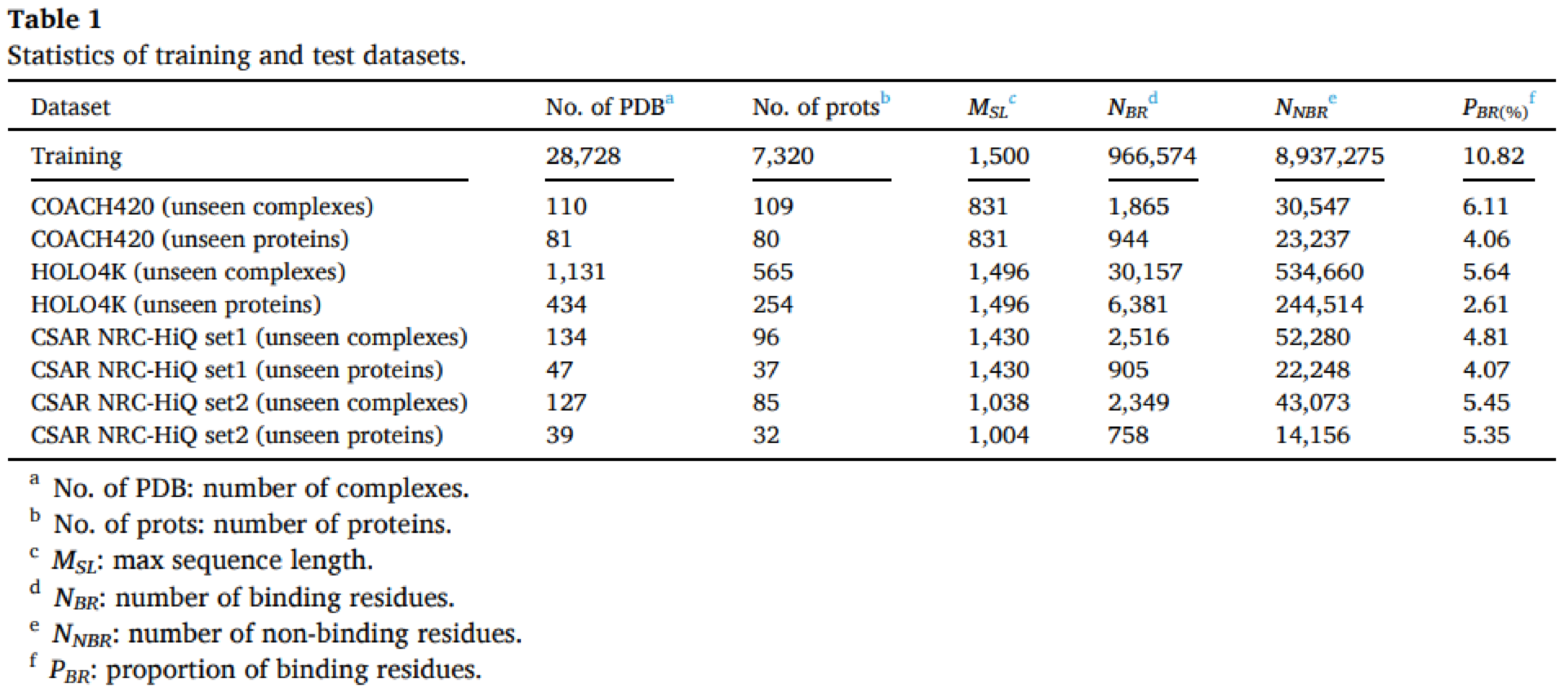


Table 1
Statistics of training and test datasets.

Dataset	No. of PDB[a]	No. of prots[b]	M_{SL}[c]	N_{BR}[d]	N_{NBR}[e]	$P_{BR(\%)}$[f]
Training	28,728	7,320	1,500	966,574	8,937,275	10.82
COACH420 (unseen complexes)	110	109	831	1,865	30,547	6.11
COACH420 (unseen proteins)	81	80	831	944	23,237	4.06
HOLO4K (unseen complexes)	1,131	565	1,496	30,157	534,660	5.64
HOLO4K (unseen proteins)	434	254	1,496	6,381	244,514	2.61
CSAR NRC-HiQ set1 (unseen complexes)	134	96	1,430	2,516	52,280	4.81
CSAR NRC-HiQ set1 (unseen proteins)	47	37	1,430	905	22,248	4.07
CSAR NRC-HiQ set2 (unseen complexes)	127	85	1,038	2,349	43,073	5.45
CSAR NRC-HiQ set2 (unseen proteins)	39	32	1,004	758	14,156	5.35

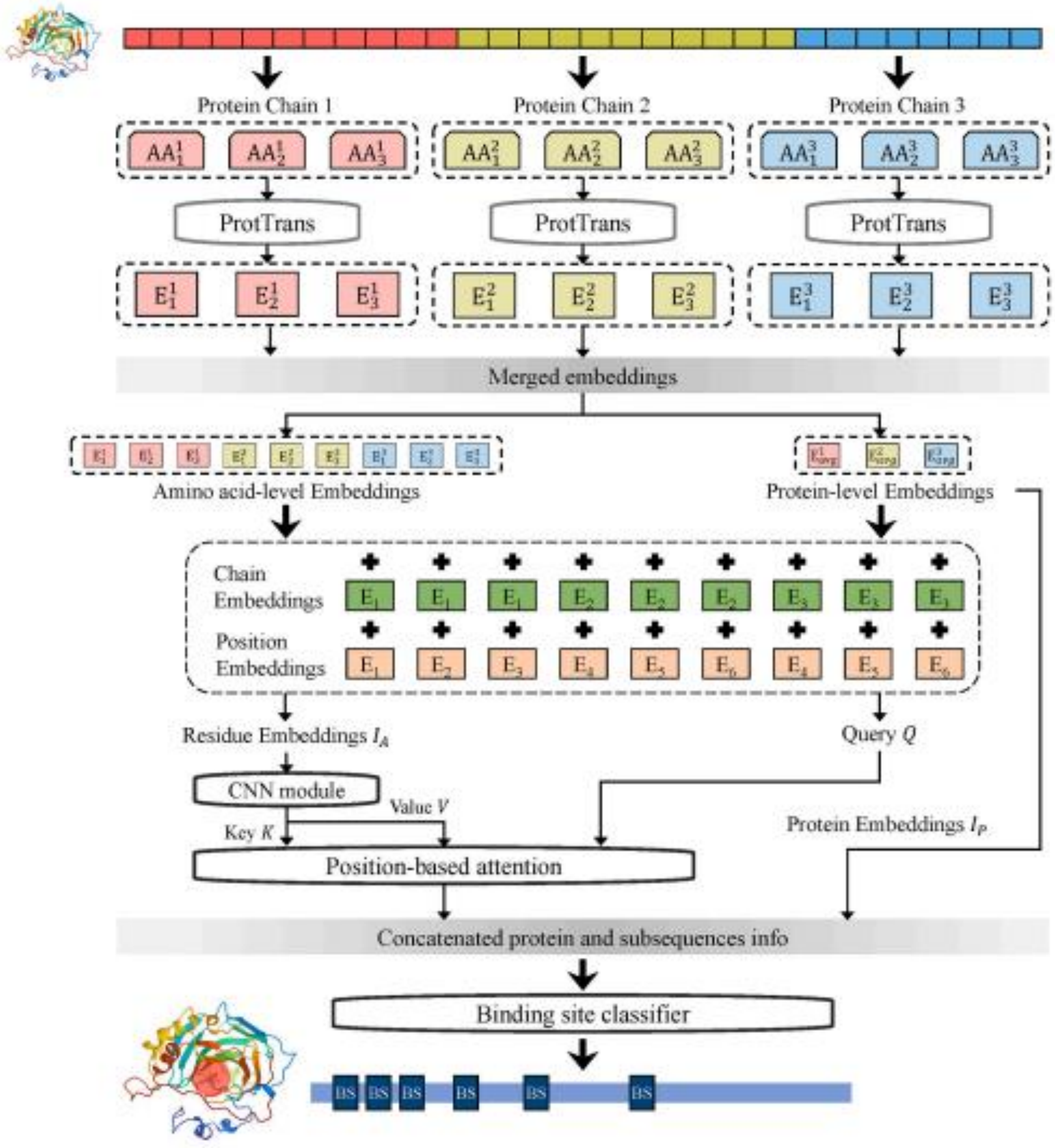
[a] No. of PDB: number of complexes.
[b] No. of prots: number of proteins.
[c] M_{SL}: max sequence length.
[d] N_{BR}: number of binding residues.
[e] N_{NBR}: number of non-binding residues.
[f] P_{BR}: proportion of binding residues.

Materials: Protein representation

A protein sequence comprises a series of amino acids in a specific order. The 20 unique amino acids (standard residues) are known, and protein sequences have different properties depending on the specific order of their amino acids. Some proteins contain non-standard (synthetic) residues. To encode the 21 residues in the protein sequences, we used ProtTrans, a pre-trained model based on the Bert Model for largescale protein sequences.

Pseq2Sites architecture

Overview of the proposed Pseq2Sites. Based on the (1) protein-level embeddings with global information about the input protein and (2) information about the adjacent (captured by CNN module) and long-distance dependencies (captured by position-based attention module) of binding residues, the classifier determines each position of the sequence as a binding or non-binding residue.



Pseq2Sites architecture

Goal:

Predict which amino acid residues in a protein sequence are binding sites (interact with ligands).

Key Steps:

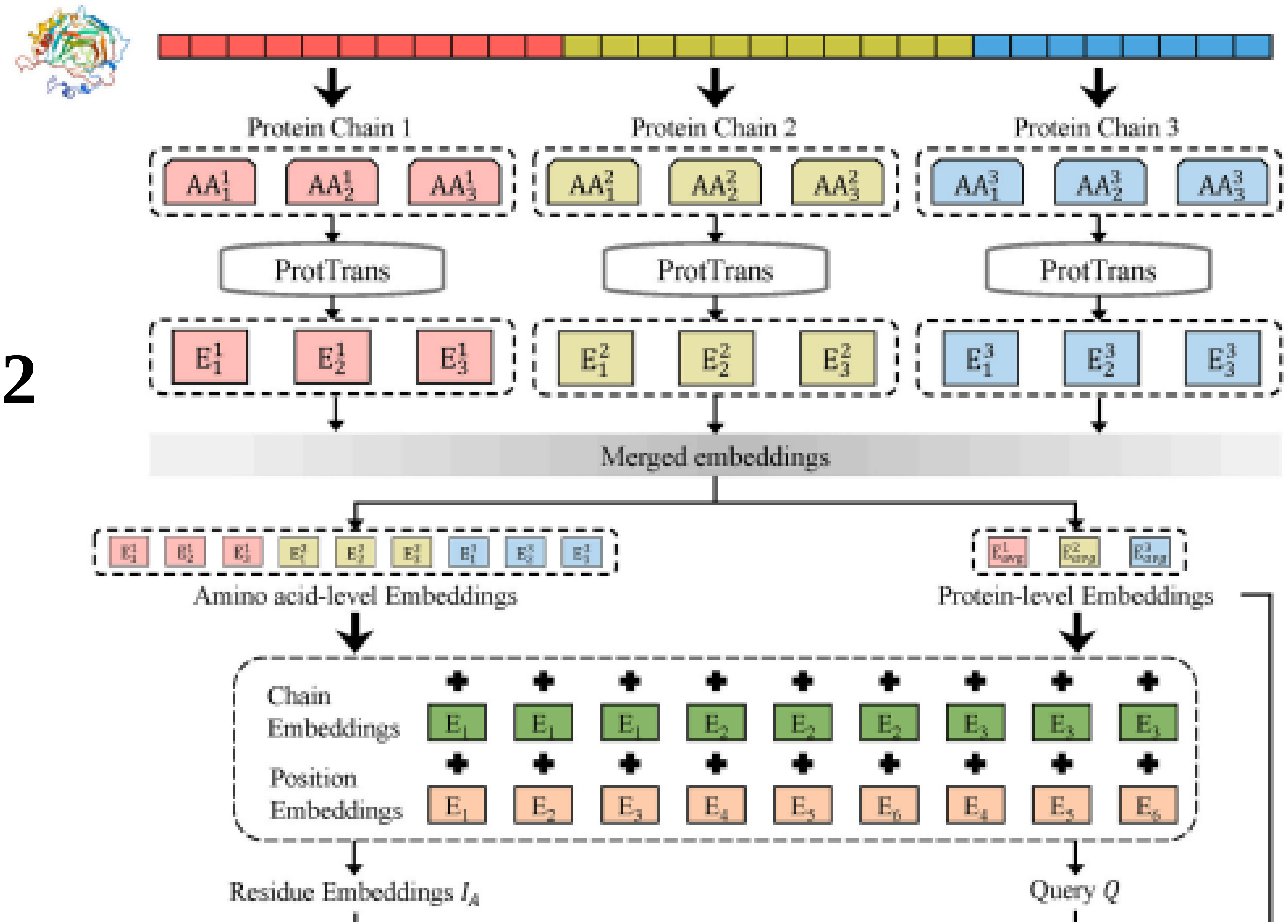
1. Input Protein Chains:

- A protein may have multiple chains (e.g., Chain 1: AA_1, AA_2, AA_3; Chain 2: AA_4, AA_5).
- Each amino acid (AA) is converted into a numerical embedding (vector representation) using ProtTrans (a protein language model).

2. Embedding Layers:

- Amino Acid-Level Embeddings (E_1, E_2, E_3, etc.): Numerical representations of each residue.
- Protein-Level Embedding (XP): Average of all amino acid embeddings in a chain.
- Position Embeddings (EP): Encodes the residue’s position in the sequence.
- Chain Embeddings (EC): Helps distinguish between different chains in multi-chain proteins.
- Final Residue Embeddings (IA): Combines all embeddings above for each amino acid.

Steps 1 & 2


Protein Chain 1
Protein Chain 2
Protein Chain 3
AA^1_1 AA^1_2 AA^1_3
AA^2_1 AA^2_2 AA^2_3
AA^3_1 AA^3_2 AA^3_3
ProtTrans
ProtTrans
ProtTrans
E^1_1 E^1_2 E^1_3
E^2_1 E^2_2 E^2_3
E^3_1 E^3_2 E^3_3
Merged embeddings
Amino acid-level Embeddings
E^1_{avg} E^2_{avg} E^3_{avg}
Protein-level Embeddings
Chain Embeddings
E_1 E_1 E_1 E_2 E_2 E_2 E_3 E_3 E_3
Position Embeddings
E_1 E_2 E_3 E_4 E_5 E_6 E_4 E_5 E_6
Residue Embeddings I_A
Query Q

Pseq2Sites architecture

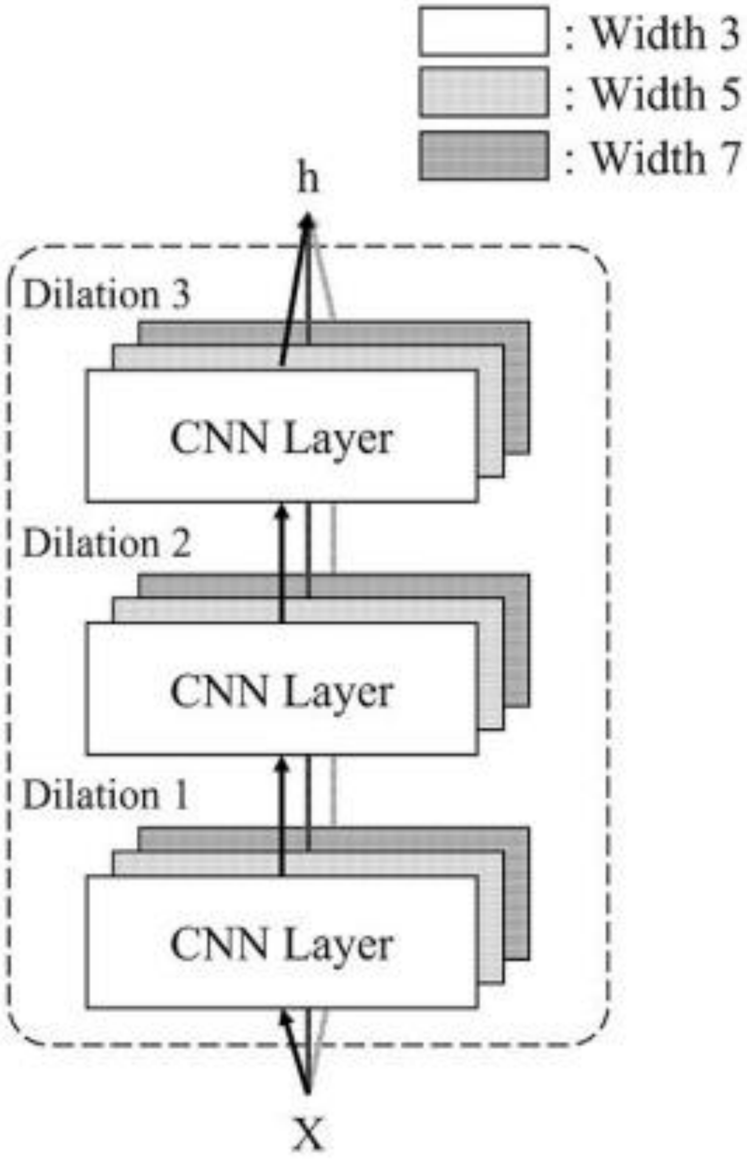
3. CNN Architecture:

- 3 stacked CNN blocks (hierarchical feature learning).
- Each block has:
  - Multiple kernel widths (3, 5, 7) → Detects patterns of different sizes.
  - Dilation rates (1, 2, 3) → Expands the "receptive field" (how much of the sequence the CNN "sees") without increasing computation.
- Output: A feature map (`O_C`) capturing local binding signatures.

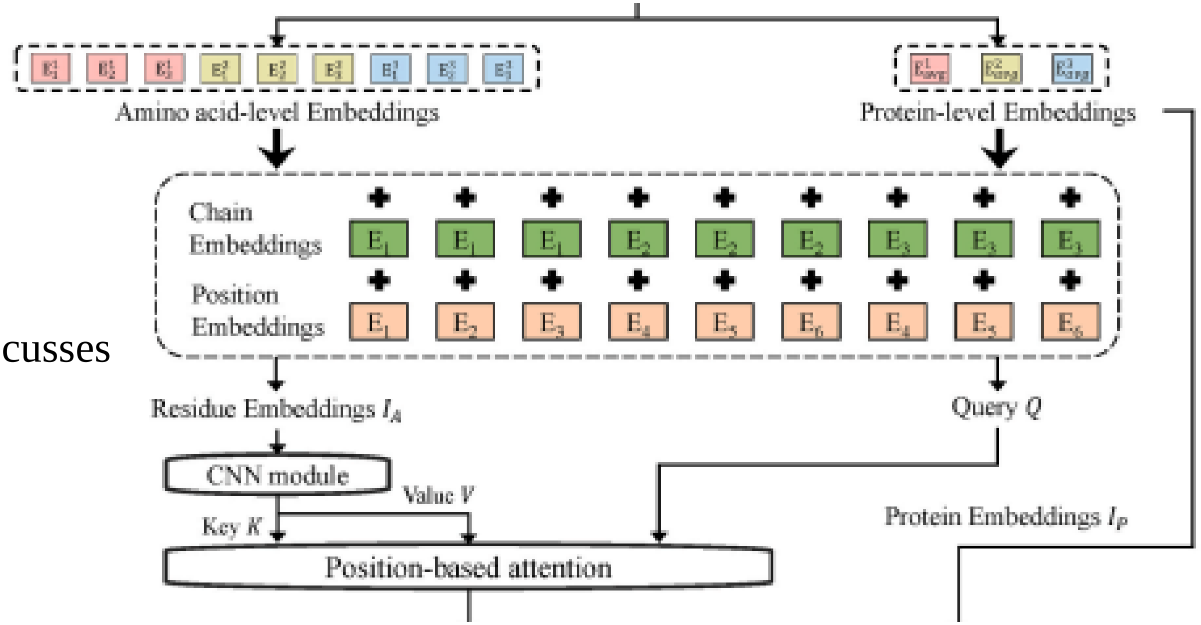
Mathematical Operation:

$$y_c[i] = \sum_{j=1}^{k} x[i + r \cdot j] w_c[j]$$

- `x`: Input residue embeddings (e.g., from ProtTrans).
- `w_c`: Learnable weights (kernel filters).
- `r`: Dilation rate (spacing between kernel taps).
- `k`: Kernel width (e.g., 3, 5, 7).



Pseq2Sites architecture



4. Attention Architecture :
The model uses three components to "pay attention":
1. Query (Q): The "question" you’re asking:
   - "Which residues look like binding sites?"
2. Key (K): A "tag" for each residue, summarizing its role.
   - Like sticky notes on book pages ("this page discusses binding").
3. Value (V): The actual content of each residue.
   - The detailed text on each page.

Step-by-Step Process

1. Compare Q & K:

   - For each residue, the model checks: "How well does this Key (K) match my Query (Q)?"

   - This gives an attention score (how much to focus on each residue).

   - Multi-head attention: Repeats this with different "perspectives" (heads) to capture diverse relationships.

   - Each "head" might focus on different patterns (e.g., one for charges, another for shapes).

2. Weighted Sum of Values (V):

   - Residues with high attention scores influence the prediction more.

Pseq2Sites architecture

5. Classifier (The Decision Maker):

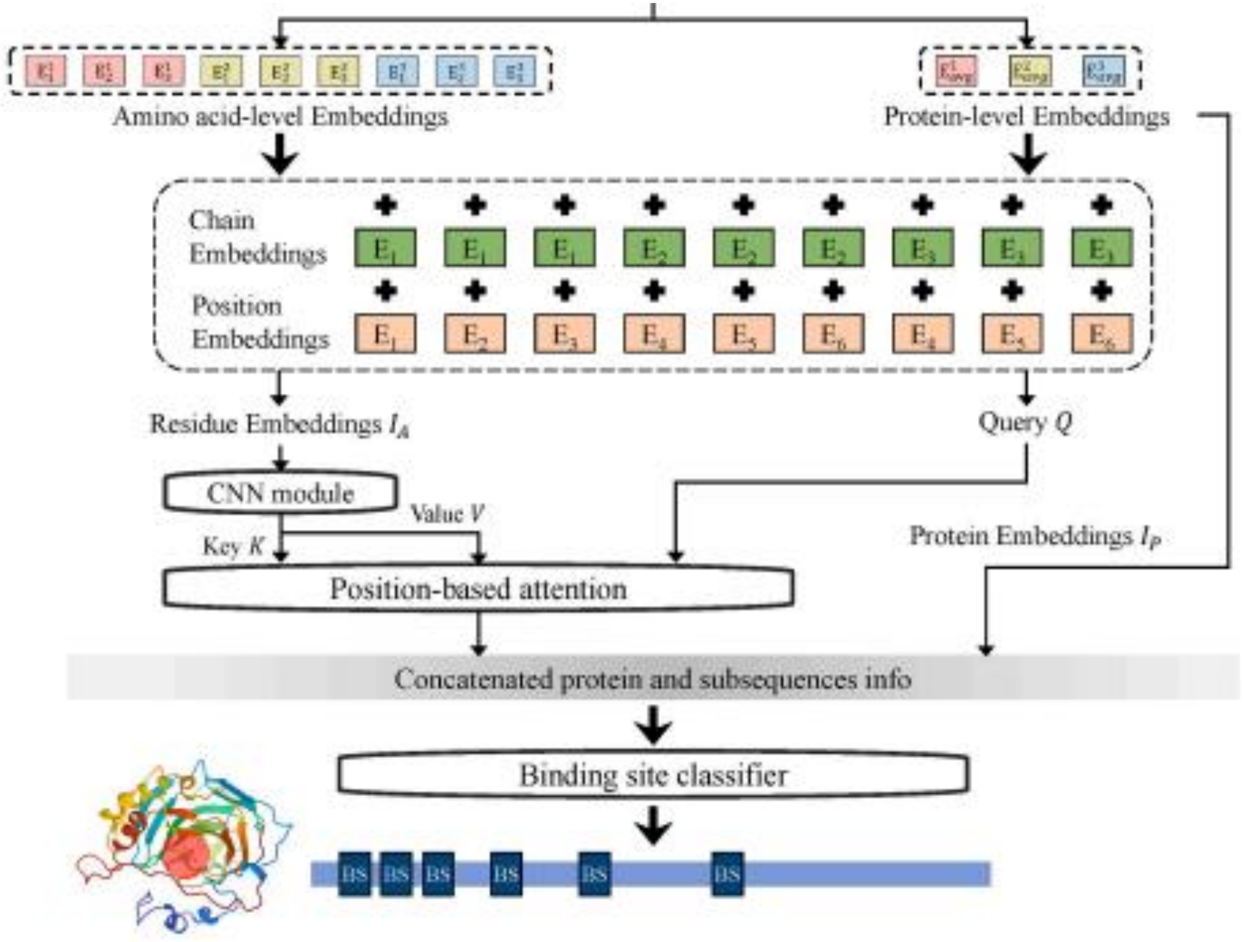
Purpose: Combines local (CNN) and global (attention) features to predict binding sites.

Step-by-Step Process

1. Concatenation: Merges protein embeddings (`$I_p$`) and attention output (`O_A`) into a single feature set (`X`).

- this is as combining a microscope view (CNN) with a satellite view (attention).

2. Fully Connected Layers(FC): Processes `X` through 4 neural network layers to make final predictions.



Results: Evaluation metrics

1) Residue-Based Success Rate (Main Metric)

Goal: See if the model's predicted binding sites match the known (real) binding sites.

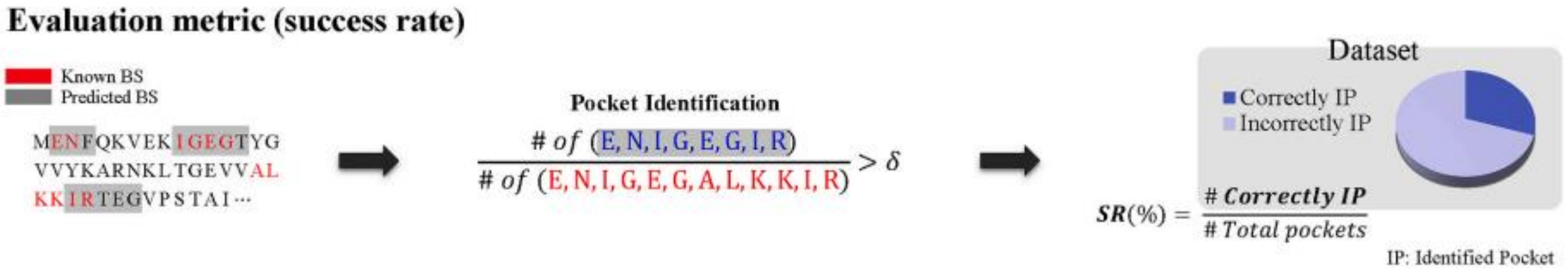
Step-by-Step Process

1. Compare predicted binding residues (`Rpred`) vs. known binding residues (`Rknown`).
2. If enough predicted residues overlap with the real ones (more than a threshold `δ`), it’s a correct prediction.

$$\frac{|R_{known} \cap R_{pred}|}{|R_{known}|} > \delta$$

3. Success Rate (SR):

$$SR\ (\%) = \frac{no.\ of\ correctly\ identified\ pockets}{total\ number\ of\ pockets}$$


Evaluation metric (success rate)
Known BS
Predicted BS
MENFQKVEKIGEGTYG
VVYKARNKLTGEVVAL
KKIRTEGVPSTAI···
Pocket Identification
$\frac{\#\ of\ (E,N,I,G,E,G,I,R)}{\#\ of\ (E,N,I,G,E,G,A,L,K,K,I,R)} > \delta$
Dataset
Correctly IP
Incorrectly IP
$SR(\%) = \frac{\#\ Correctly\ IP}{\#\ Total\ pockets}$
IP: Identified Pocket

Results: Evaluation metrics

2) Top-(n + 2) Prediction Rule:

- Some methods predict multiple possible binding regions (e.g., top 3 guesses).

- The study uses Top-(n + 2) predictions (e.g., top 5 instead of top 3) for fairer comparison, since more guesses improve accuracy.

- Reason: Binding sites can be tricky, so extra guesses help avoid missing the true site.

3) Additional metrics for evaluating binding site predictive performance include the precision, recall (sensitivity), specificity, G-mean, Fβ score, and accuracy
TP: number of correctly predicted binding residues
FP: number of non-binding residues predicted as binding sites
TN: number of correctly predicted non-binding residues
FN: number of binding residues predicted as non-binding sites

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

$$recall\ (sensitivity) = \frac{TP}{TP + FN}$$

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

$$G - mean = \sqrt{sentivitiy \times specificity}$$

$$F_{\beta} = \frac{(\beta^2 + 1) precision \times recall}{\beta^2 precision + recall}$$

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

Results: Hyperparameter Tuning

The threshold (τ) in Pseq2Sites that decides if a residue is a binding site (e.g., τ=0.8 means "call it a binding site if the model is ≥80% confident").

Step-by-Step Process

1. *Parameter Tested:* Binding site threshold (τ).

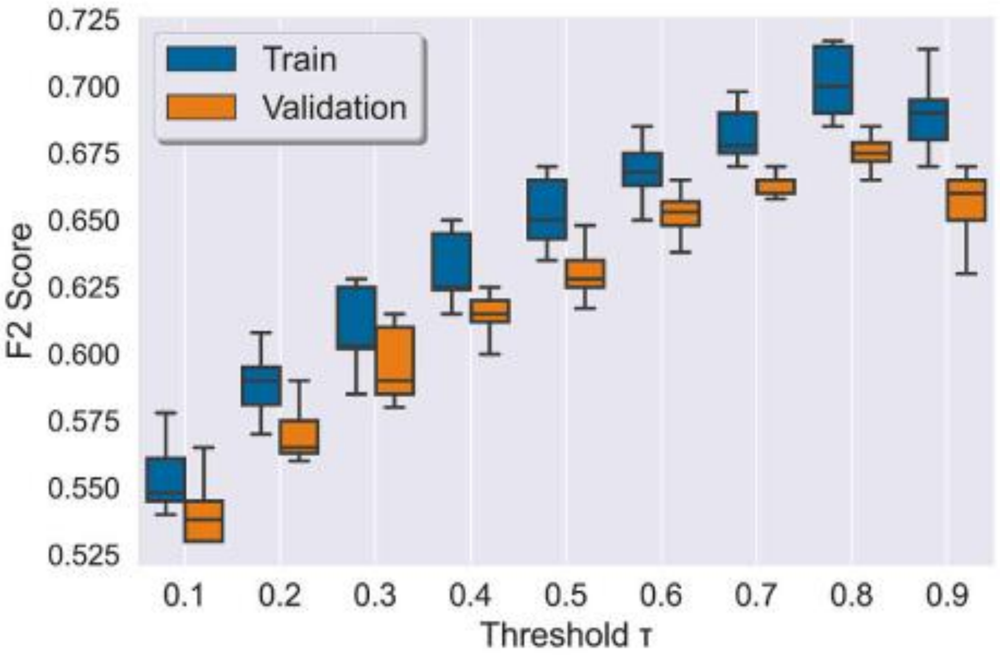
- Tested values: 0.1, 0.2, ..., 0.9 (in 0.1 increments).

2. *Evaluation Metric:* Used the F2 score (prioritizes finding real binding sites over avoiding false alarms).

3. *Method:*

- 5-fold cross-validation: Split training data into 5 parts, trained on 4 and tested on 1 (repeated 5x for reliability).

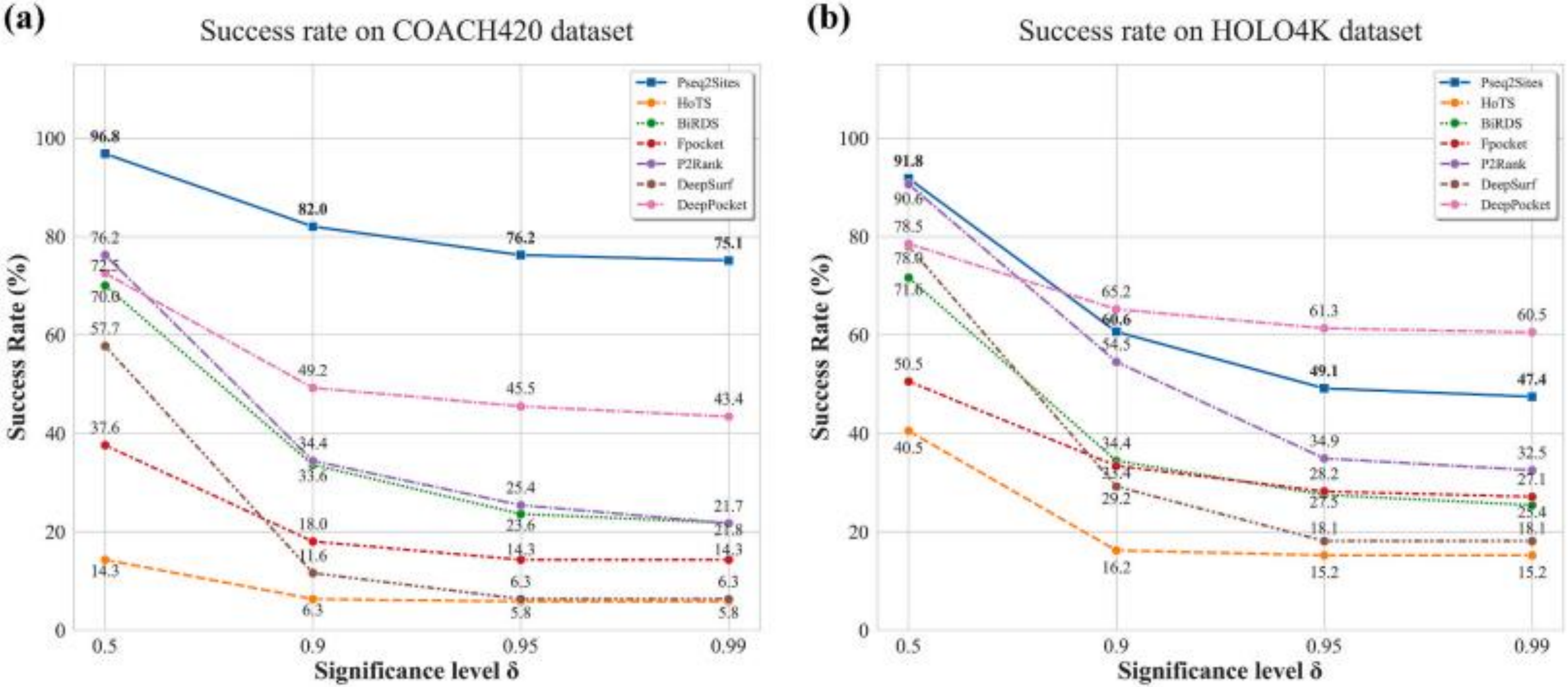
- Result: Best F2 score at τ=0.8 (see graph where validation score peaks at 0.8).



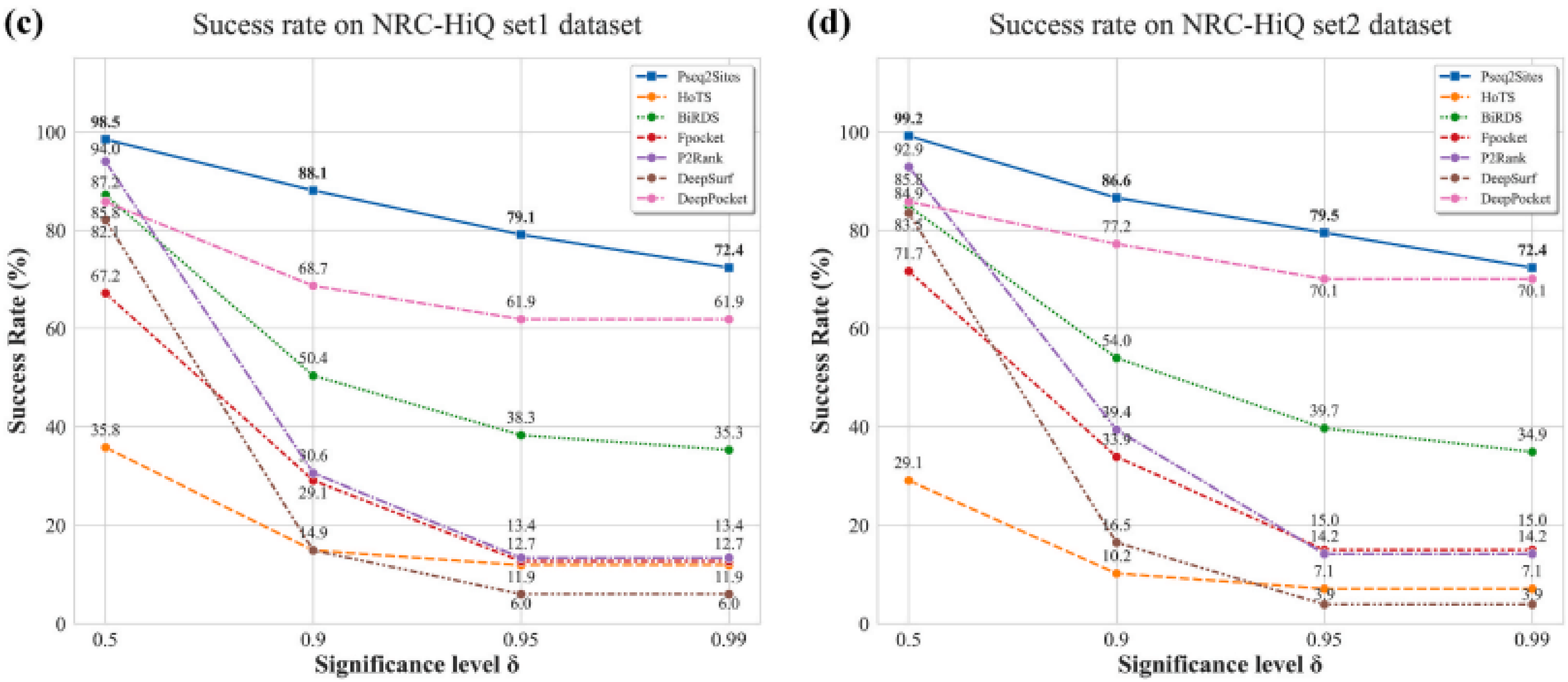
Results: Performance comparison based on the success rate

1. Pseq2Sites is the best at predicting protein binding sites, beating all other methods in most tests.
2. Success Rate (Accuracy):
   - COACH420 dataset ($\delta=0.5$):
     - Pseq2Sites: 96.8% (extremely accurate)
     - Others: Much worse (e.g., Fpocket: ~48%, HoTS: ~16%).
   - HOLO4K dataset ($\delta=0.5$):
     - Pseq2Sites: 91.8%
     - DeepPocket (2nd best): 78.5%
   - NRC-HiQ datasets: Pseq2Sites wins at all difficulty levels.
3. Stricter Tests (Higher δ = Harder to Pass):
   - All methods perform worse when the rules get stricter ($\delta=0.99$), but Pseq2Sites declines the least (-22% vs. -40% to -89% for others).

Results: Performance comparison based on the success rate



Results: Performance comparison based on the success rate



Results: Performance comparison for unseen proteins

Pseq2Sites was tested on completely new proteins (not seen during training) to check if it works in real-world scenarios.

1. COACH420 Dataset (Hardest Test)
   - Pseq2Sites: #1 at all difficulty levels (δ=0.5 to 0.99).
   - Even with ≤40% similarity to training data, it predicted binding sites better than all other methods.
2. HOLO4K Dataset
   - P2Rank (3D-based) won at δ=0.5, but Pseq2Sites was #2 overall.
   - DeepPocket (another 3D method) performed slightly better at stricter levels.
3. NRC-HiQ Datasets (Best Performance)
   - Pseq2Sites: ~98% success at δ=0.5 (almost perfect!).
   - Even at δ=0.99, it was ~10% better than the runner-up.
4. Comparison to Other Sequence-Based Models
   - Pseq2Sites crushed HoTS (another sequence-based tool), which came last in every test.
   - Remarkably, Pseq2Sites even beat most 3D-based methods—despite using only sequence data.

Results: Performance Comparison: Pseq2Sites vs. Other Models

- Trade-off: Precision & Specificity:

  - DeepSurf (3D-based) had the best precision (fewer false positives).

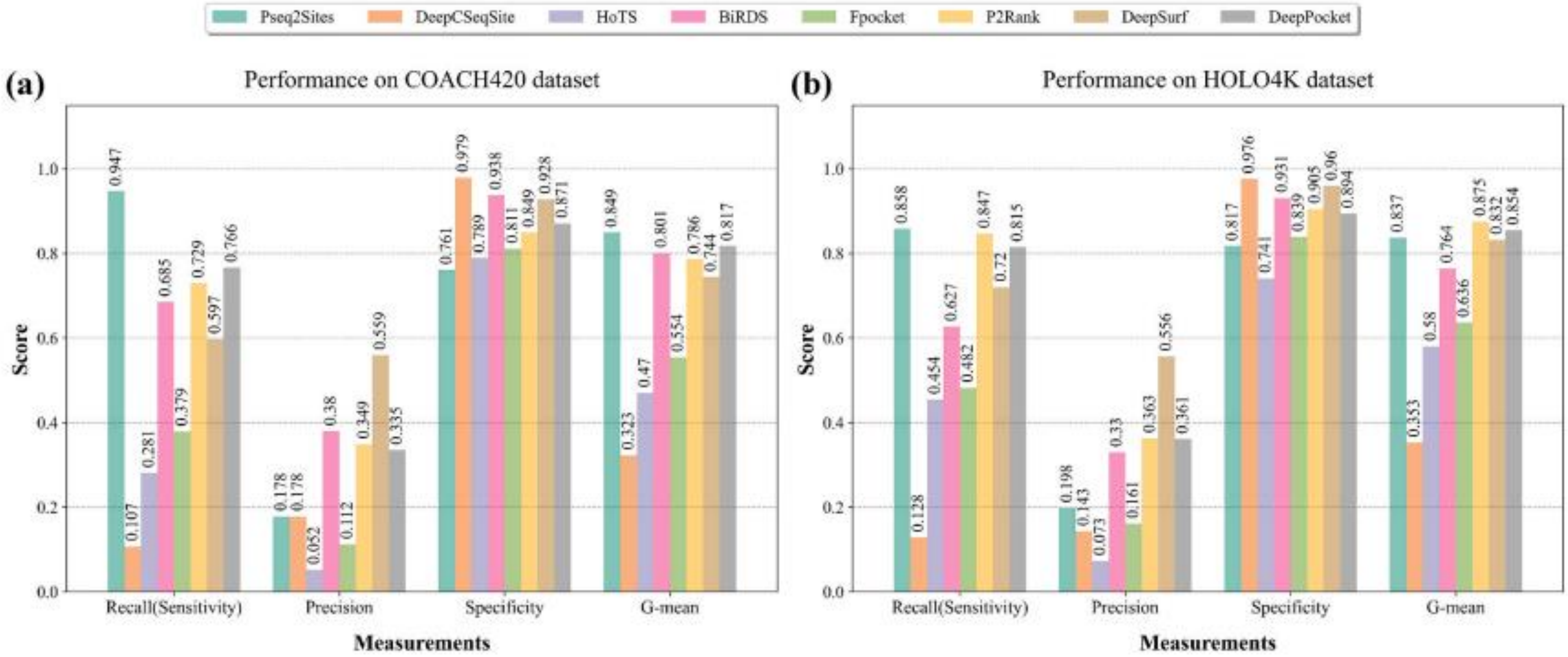
  - Pseq2Sites predicted more binding residues, leading to slightly lower precision but higher overall accuracy (G-mean).

- Best Balanced Performance (G-mean):

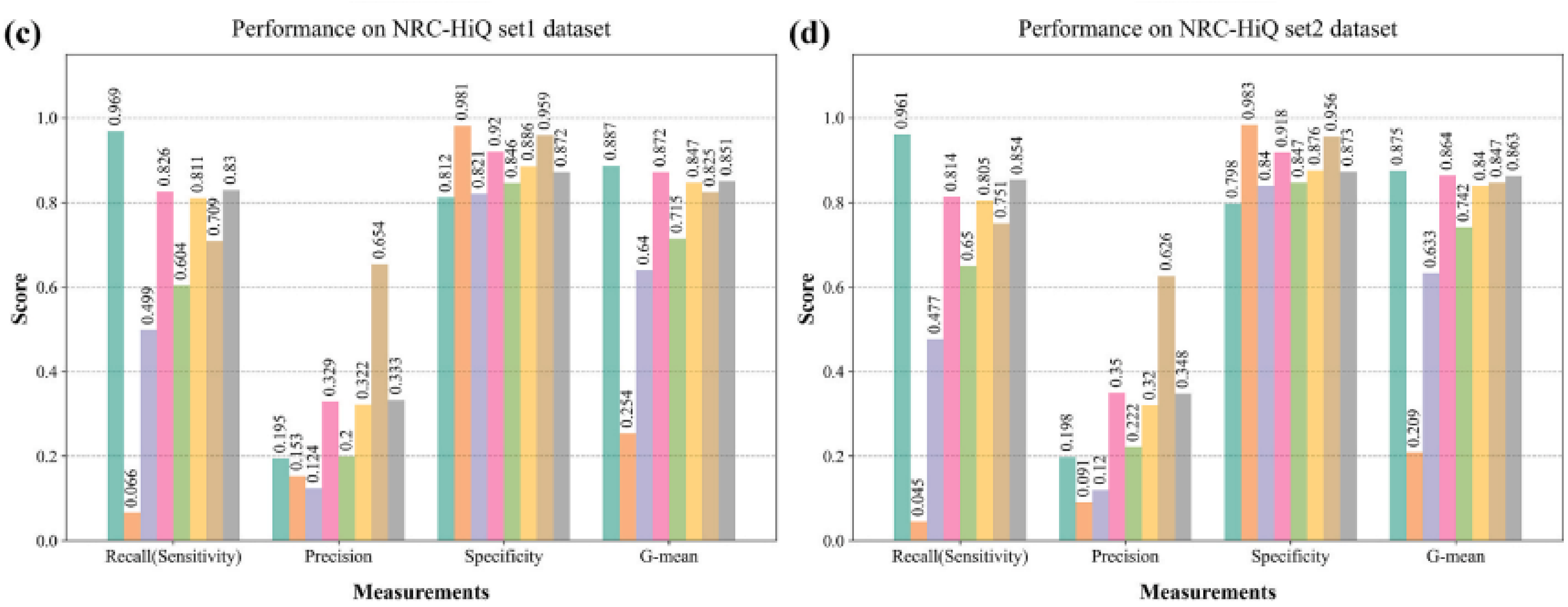
  - Pseq2Sites ranked #1 in G-mean (balanced accuracy) for almost all datasets.

  - Statistically significant in 23/26 tests (proven by t-tests and Cohen’s d) (with strong math proof, like a "p-value < 0.05").

Results: Performance Comparison: Pseq2Sites vs. Other Models



Results: Performance Comparison: Pseq2Sites vs. Other Models



Results: Ablation test

Researchers removed parts of Pseq2Sites to see how each piece affects performance.

The model has 3 key parts:

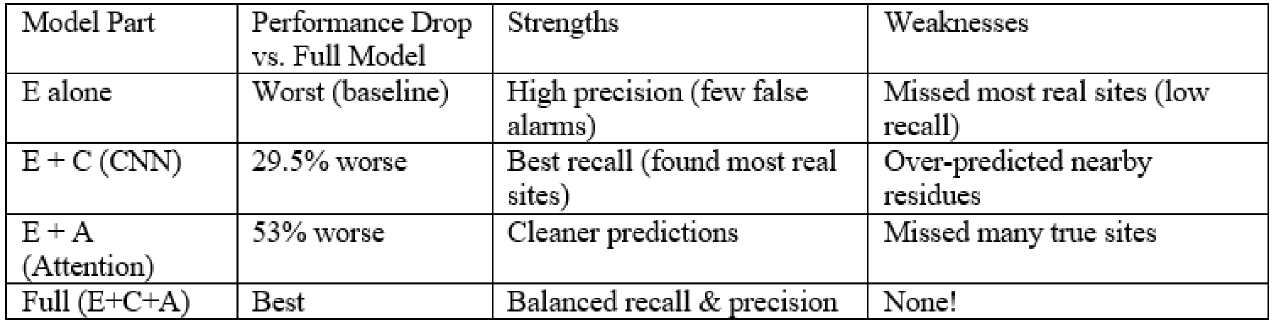
1. Embeddings (E): Basic amino acid info + position/chain data.

2. CNN Module (C): Detects local binding patterns.

3. Attention Module (A): Finds long-distance relationships.

Tests:

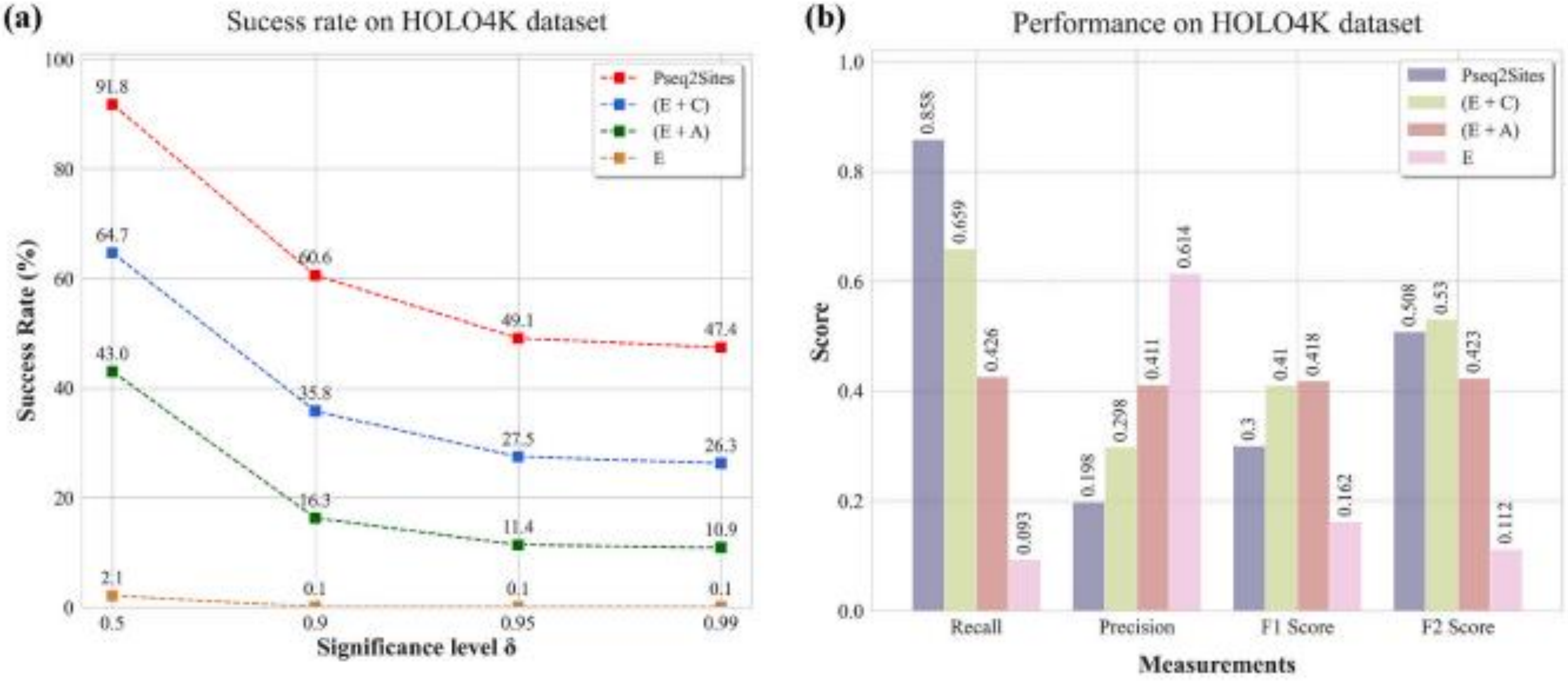
- E alone: Just raw data, no fancy processing.
- E + C: CNN added (local patterns only).
- E + A: Attention added (long-range only).
- Full model (E + C + A): Everything combined.



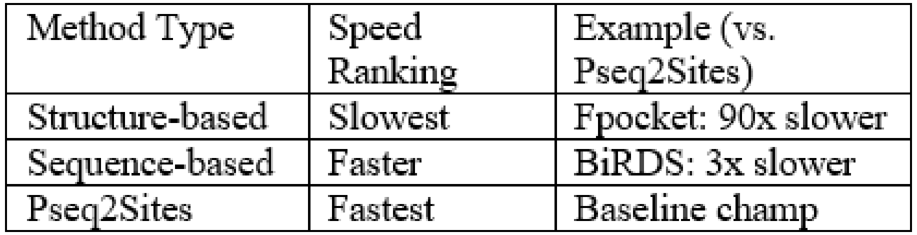
Model Part	Performance Drop vs. Full Model	Strengths	Weaknesses
E alone	Worst (baseline)	High precision (few false alarms)	Missed most real sites (low recall)
E + C (CNN)	29.5% worse	Best recall (found most real sites)	Over-predicted nearby residues
E + A (Attention)	53% worse	Cleaner predictions	Missed many true sites
Full (E+C+A)	Best	Balanced recall & precision	None!

1. CNN (C) is crucial: It catches most binding sites but gets "overexcited" (labels neighbors as binding too).
2. Attention (A) helps focus: Reduces false positives but can’t work alone.
3. Together they shine:
   - CNN finds candidates, attention refines them.

Results: Ablation test



Results: Comparison of computational efficiency



Method Type	Speed Ranking	Example (vs. Pseq2Sites)
Structure-based	Slowest	Fpocket: 90x slower
Sequence-based	Faster	BiRDS: 3x slower
Pseq2Sites	Fastest	Baseline champ

Pseq2Sites is Lightning Fast:

- 90x faster than Fpocket (a structure-based method).
- 3x faster than BiRDS (another sequence-based tool).
- Practical research needs quick results (e.g., screening thousands of proteins for drug discovery).

Why?:

1. Sequence-Based Advantage: Avoids slow 3D structure calculations.
2. Optimized Architecture: CNN + attention balance speed and accuracy.

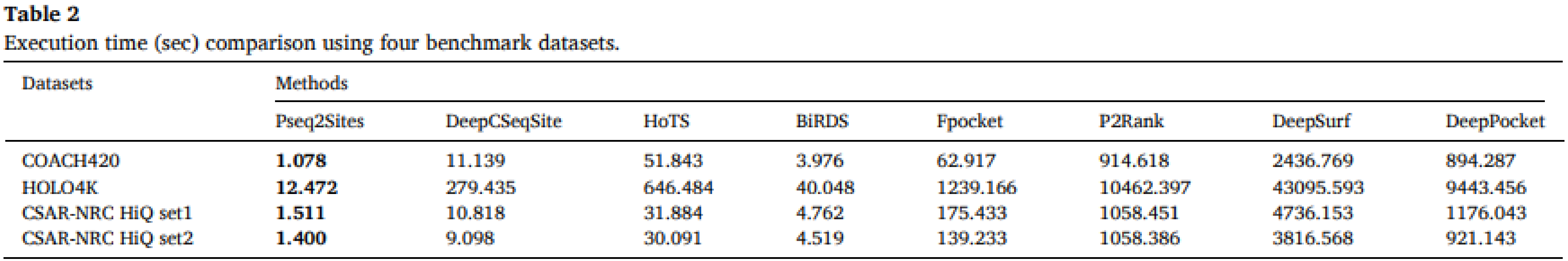


Table 2
Execution time (sec) comparison using four benchmark datasets.

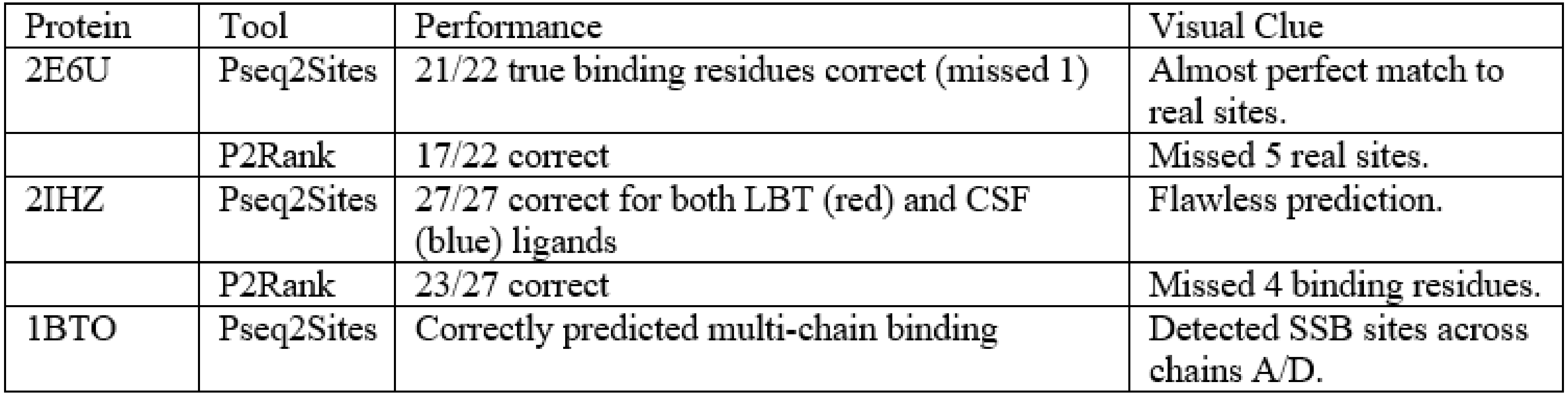
Datasets	Methods							
	Pseq2Sites	DeepCSeqSite	HoTS	BiRDS	Fpocket	P2Rank	DeepSurf	DeepPocket
COACH420	**1.078**	11.139	51.843	3.976	62.917	914.618	2436.769	894.287
HOLO4K	**12.472**	279.435	646.484	40.048	1239.166	10462.397	43095.593	9443.456
CSAR-NRC HiQ set1	**1.511**	10.818	31.884	4.762	175.433	1058.451	4736.153	1176.043
CSAR-NRC HiQ set2	**1.400**	9.098	30.091	4.519	139.233	1058.386	3816.568	921.143

Results: Visualization of Pseq2Sites prediction

The Test Cases:

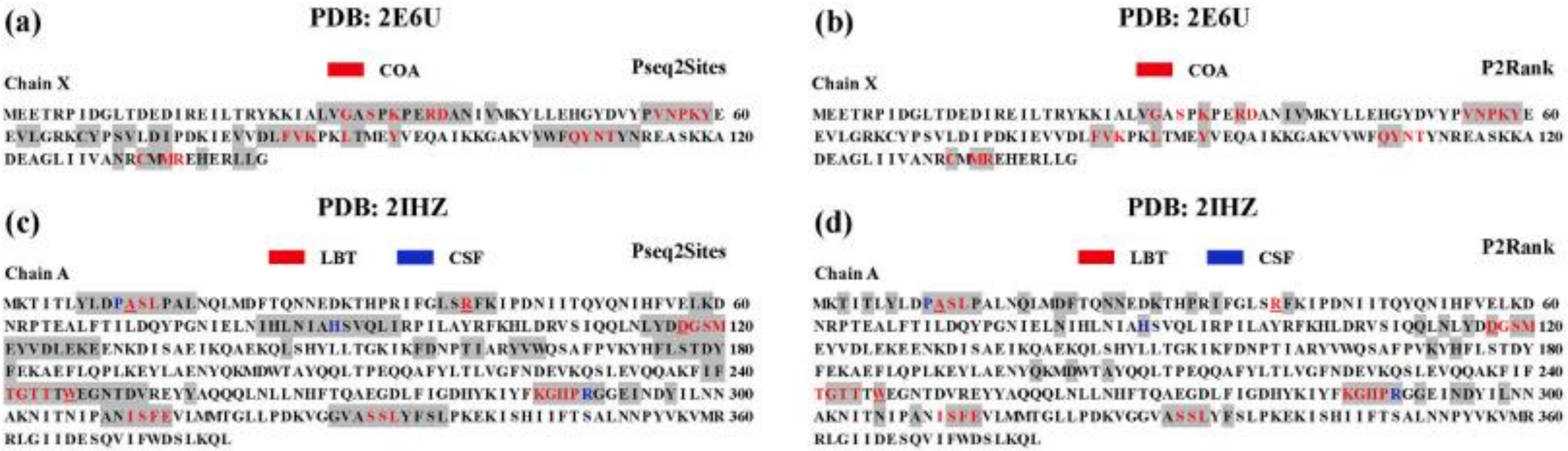
Researchers picked 3 real protein-ligand complexes (from unseen test data) to compare Pseq2Sites vs. P2Rank:

- 2E6U: Single-chain protein (242 residues) binding to COA ligand.
- 2IHZ: Larger single-chain protein (+319 residues) binding to LBT and CSF ligands.
- 1BTO: Complex 4-chain protein binding to NAD and SSB ligands.

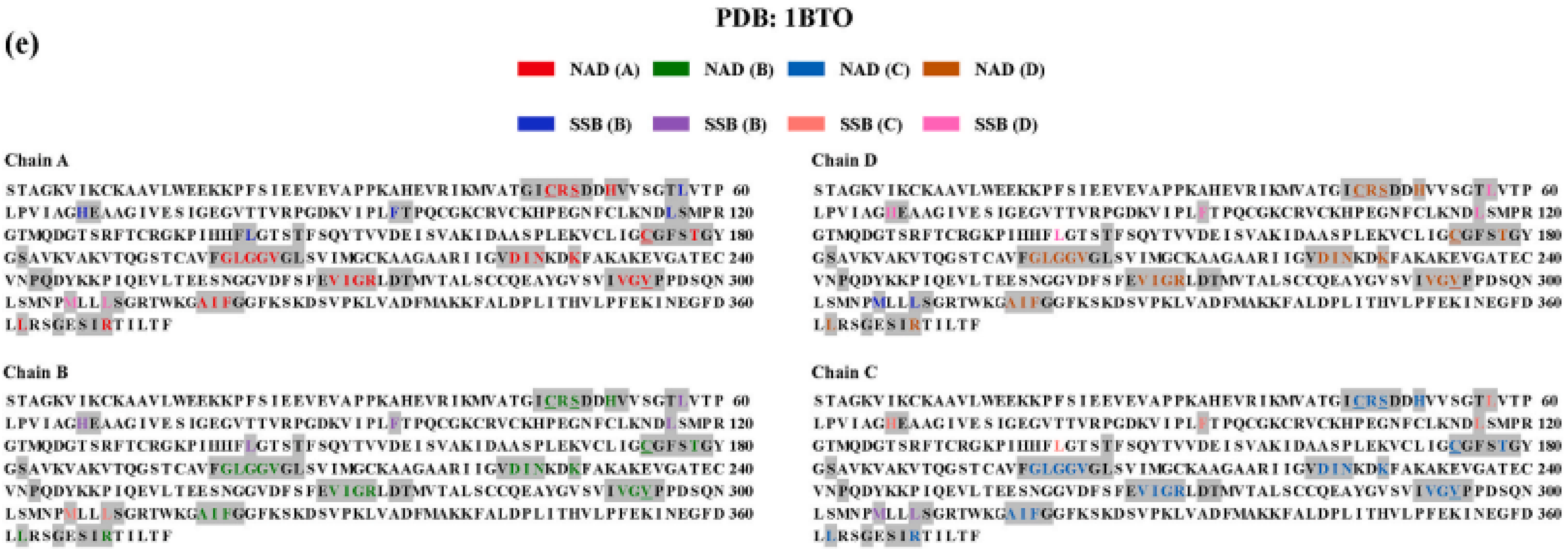


Protein	Tool	Performance	Visual Clue
2E6U	Pseq2Sites	21/22 true binding residues correct (missed 1)	Almost perfect match to real sites.
	P2Rank	17/22 correct	Missed 5 real sites.
2IHZ	Pseq2Sites	27/27 correct for both LBT (red) and CSF (blue) ligands	Flawless prediction.
	P2Rank	23/27 correct	Missed 4 binding residues.
1BTO	Pseq2Sites	Correctly predicted multi-chain binding	Detected SSB sites across chains A/D.

Results: Visualization of Pseq2Sites prediction


(a)
PDB: 2E6U
COA
Pseq2Sites
Chain X
MEETRPIDGLTDEDIREILTRYKKIALVGASPKPERDANIVMKYLLEHGYDVYPVNPKYE 60
EVLGRKCYPSVLDIPDKIEVVDLFVKPKLTMEYVEQAIKKGAKVVWFQYNTYNREASKKA 120
DEAGLIIVANRCMMREHERLLG
(b)
PDB: 2E6U
COA
P2Rank
Chain X
MEETRPIDGLTDEDIREILTRYKKIALVGASPKPERDANIVMKYLLEHGYDVYPVNPKYE 60
EVLGRKCYPSVLDIPDKIEVVDLFVKPKLTMEYVEQAIKKGAKVVWFQYNTYNREASKKA 120
DEAGLIIVANRCMMREHERLLG
(c)
PDB: 2IHZ
LBT
CSF
Pseq2Sites
Chain A
MKTITLYLDPASLPALNQLMDFTQNNEDKTHPRIFGLSRFKIPDNIITQYQNIHFVELKD 60
NRPTEALFTILDQYPGNIELNIHLNIAHSVQLIRPILAYRFKHLDRVSIQQLNLYDDGSM 120
EYVDLEKEENKDISAEIKQAEKQLSHYLLTGKIKFDNPTIARYVWQSAFPVKYHFLSTDY 180
FEKAEFLQPLKEYLAENYQKMDWTAYQQLTPEQQAFYLTLVGFNDEVKQSLEVQQAKFIF 240
TGTITWEGNTDVREYYAQQQLNLLNHFTQAEGDLFIGDHYKIYFKGHPRGGEINDYILNN 300
AKNITNIPANISFEVLMMTGLLPDKVGGVASSLYFSLPKEKISHIIFTSALNNPYVKVMR 360
RLGIIDESQVIFWDSLKQL
(d)
PDB: 2IHZ
LBT
CSF
P2Rank
Chain A
MKTITLYLDPASLPALNQLMDFTQNNEDKTHPRIFGLSRFKIPDNIITQYQNIHFVELKD 60
NRPTEALFTILDQYPGNIELNIHLNIAHSVQLIRPILAYRFKHLDRVSIQQLNLYDDGSM 120
EYVDLEKEENKDISAEIKQAEKQLSHYLLTGKIKFDNPTIARYVWQSAFPVKYHFLSTDY 180
FEKAEFLQPLKEYLAENYQKMDWTAYQQLTPEQQAFYLTLVGFNDEVKQSLEVQQAKFIF 240
TGTITWEGNTDVREYYAQQQLNLLNHFTQAEGDLFIGDHYKIYFKGHPRGGEINDYILNN 300
AKNITNIPANISFEVLMMTGLLPDKVGGVASSLYFSLPKEKISHIIFTSALNNPYVKVMR 360
RLGIIDESQVIFWDSLKQL

Results: Visualization of Pseq2Sites prediction



(e)

PDB: 1BTO

🟥 NAD (A) 🟩 NAD (B) 🟦 NAD (C) 🟧 NAD (D)

🟦 SSB (B) 🟪 SSB (B) 🟥 SSB (C) 🟪 SSB (D)

Chain A

```
STAGKVIKCKAAVLWEEKKPFSIEEVEVAPPKAHEVRIKMVATGICRSDDHVVSGTLVTP  60
LPVIAGHEAAGIVESIGEGVTTVRPGDKVIPLFTPQCGKCRVCKHPEGNFCLKNDLSMPR 120
GTMQDGTSRFTCRGKPIHHFLGTSTFSQYTVVDEISVAKIDAASPLEKVCLIGCGFSTGY 180
GSAVKVAKVTQGSTCAVFGLGGVGLSVIMGCKAAGAARIIGVDINKDKFAKAKEVGATEC 240
VNPQDYKKPIQEVLTEESNGGVDFSFEVIGRLDTMVTALSCCQEAYGVSVIVGVPPDSQN 300
LSMNPMLLLSGRTWKGAIFGGFKSKDSVPKLVADFMAKKFALDPLITHVLPFEKINEGFD 360
LLRSGESIRTILTF
```

Chain D

```
STAGKVIKCKAAVLWEEKKPFSIEEVEVAPPKAHEVRIKMVATGICRSDDHVVSGTLVTP  60
LPVIAGHEAAGIVESIGEGVTTVRPGDKVIPLFTPQCGKCRVCKHPEGNFCLKNDLSMPR 120
GTMQDGTSRFTCRGKPIHHFLGTSTFSQYTVVDEISVAKIDAASPLEKVCLIGCGFSTGY 180
GSAVKVAKVTQGSTCAVFGLGGVGLSVIMGCKAAGAARIIGVDINKDKFAKAKEVGATEC 240
VNPQDYKKPIQEVLTEESNGGVDFSFEVIGRLDTMVTALSCCQEAYGVSVIVGVPPDSQN 300
LSMNPMLLLSGRTWKGAIFGGFKSKDSVPKLVADFMAKKFALDPLITHVLPFEKINEGFD 360
LLRSGESIRTILTF
```

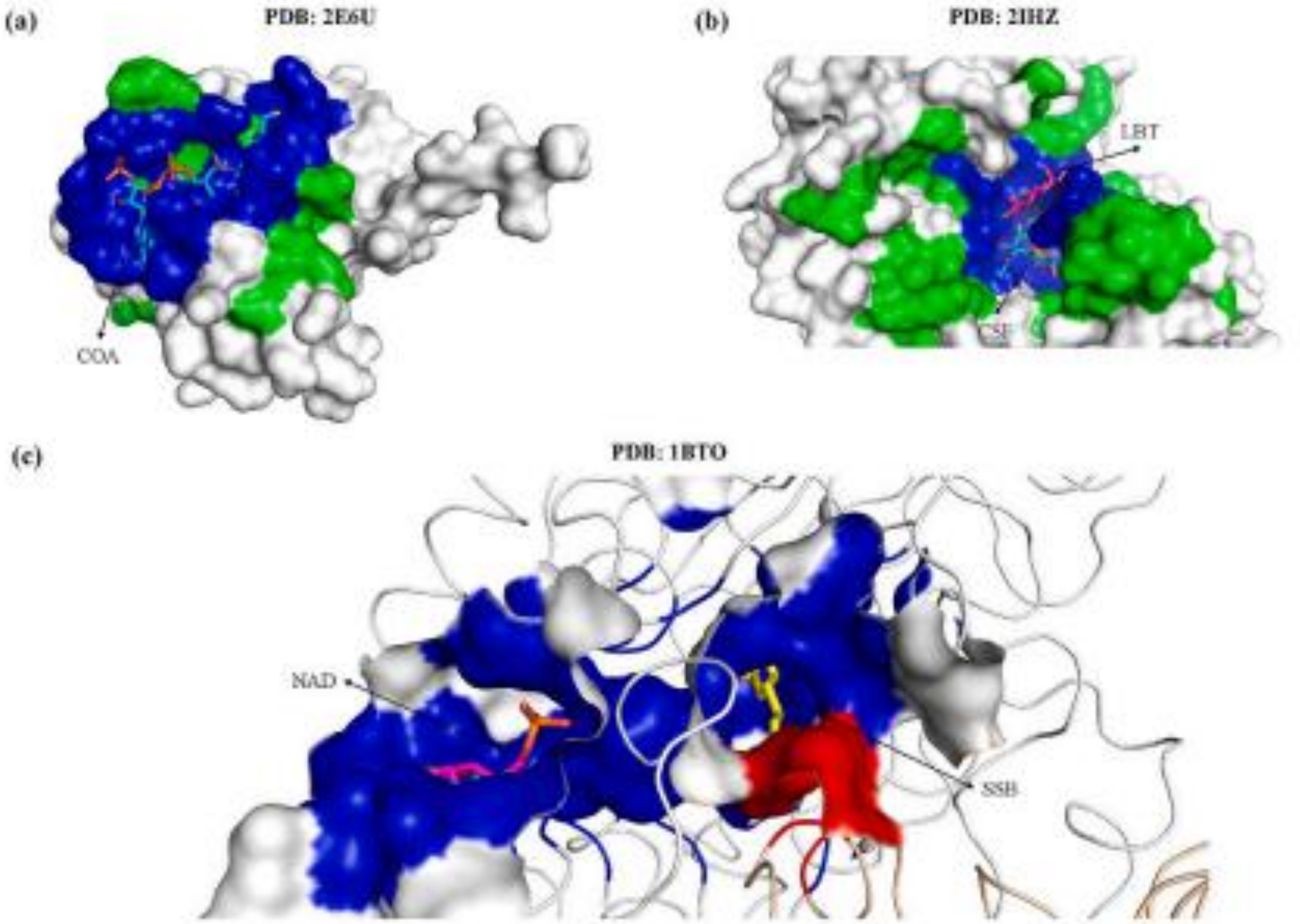
Chain B

```
STAGKVIKCKAAVLWEEKKPFSIEEVEVAPPKAHEVRIKMVATGICRSDDHVVSGTLVTP  60
LPVIAGHEAAGIVESIGEGVTTVRPGDKVIPLFTPQCGKCRVCKHPEGNFCLKNDLSMPR 120
GTMQDGTSRFTCRGKPIHHFLGTSTFSQYTVVDEISVAKIDAASPLEKVCLIGCGFSTGY 180
GSAVKVAKVTQGSTCAVFGLGGVGLSVIMGCKAAGAARIIGVDINKDKFAKAKEVGATEC 240
VNPQDYKKPIQEVLTEESNGGVDFSFEVIGRLDTMVTALSCCQEAYGVSVIVGVPPDSQN 300
LSMNPMLLLSGRTWKGAIFGGFKSKDSVPKLVADFMAKKFALDPLITHVLPFEKINEGFD 360
LLRSGESIRTILTF
```

Chain C

```
STAGKVIKCKAAVLWEEKKPFSIEEVEVAPPKAHEVRIKMVATGICRSDDHVVSGTLVTP  60
LPVIAGHEAAGIVESIGEGVTTVRPGDKVIPLFTPQCGKCRVCKHPEGNFCLKNDLSMPR 120
GTMQDGTSRFTCRGKPIHHFLGTSTFSQYTVVDEISVAKIDAASPLEKVCLIGCGFSTGY 180
GSAVKVAKVTQGSTCAVFGLGGVGLSVIMGCKAAGAARIIGVDINKDKFAKAKEVGATEC 240
VNPQDYKKPIQEVLTEESNGGVDFSFEVIGRLDTMVTALSCCQEAYGVSVIVGVPPDSQN 300
LSMNPMLLLSGRTWKGAIFGGFKSKDSVPKLVADFMAKKFALDPLITHVLPFEKINEGFD 360
LLRSGESIRTILTF
```

Results: Visualization of Pseq2Sites prediction



Visualization of the 3D structure of binding sites predicted by Pseq2Site. In the 2E6U (a) is about the 2E6U complex and (b) is about the 2IHA complex. (c) Is about Chain A (marked in white) and chain B (marked in ivory) of 1BTO complex.

Results: Case study: analysis of binding-site detection results

Pseq2Sites predicts more binding residues than other tools (lower precision), but:

- ~80% of its predictions are within 12Å of real binding sites (vs. <40% for HoTS).
- These "extra" residues likely participate in long-range interactions (e.g., electrostatic forces) critical for ligand binding.

Binding Sites Aren’t Just 4Å Contacts

- Traditional methods focus on residues directly touching ligands (≤4Å).
- Pseq2Sites also flags residues involved in:
  - Long-range interactions (e.g., stabilizing ligand orientation via charge).
  - Allosteric effects (indirectly influencing binding).

Pseq2Sites sacrifices some precision to capture critical long-range interactions, making it uniquely valuable for:

- Drug development (better target identification).
- Protein engineering (designing binding interfaces).

Conclusion

Pseq2Sites is a deep-learning tool that predicts where small molecules (ligands) bind to proteins—using only protein sequences (no 3D structure needed!).

Why It’s Special?

1. CNN + Attention:

- CNN: Finds local binding motifs (e.g., "HIS-ASP-SER").

- Attention: Links distant residues that work together (e.g., residue 10 + residue 100).

2. Outperforms 3D Methods:

- More accurate than tools requiring protein structures.

- Works even on never-before-seen proteins.

3. Fast & Practical:

- Runs 90x faster than some 3D-based tools.

- Predicts biologically meaningful sites (even if not direct contacts).

Limitations & Future Work:

- Sequence Length Limit: Can’t process extremely long proteins (yet).

- Next Steps:

  1. Remove length restrictions.

  2. Predict binding strength (affinity) using the same tech.

Why It Matters?

- Drug Discovery: Accelerates finding new drug targets.

- Accessibility: Works without expensive 3D data—democratizing research.

Reference

Seo, S., Choi, J., Choi, S., Lee, J., Park, C., and Park, S. (2024). Pseq2Sites: enhancing protein sequence-based ligand binding-site prediction accuracy via the deep convolutional network and attention mechanism. Eng. Appl. Artif. Intell. 127, 107257. doi:10.1016/j.engappai. 2023.107257