

LABind: identifying protein binding ligand-aware sites via learning interactions between ligand and protein

Corresponding Author: Professor Lijun Quan

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

LABind introduces a novel ligand-aware approach for predicting protein-ligand binding sites using both protein and ligand representations. Notable results include LABind's superior performance across benchmark datasets (DS1, DS2, and DS3), outperforming existing methods like LigBind, DeepSurf, and GPSite. LABind achieves high accuracy in identifying binding sites even for proteins with predicted structures rather than native structures, maintaining robust predictive capabilities.

The work is significant to protein-ligand interaction modeling and drug discovery, addressing limitations in single- and multi-ligand approaches. It advances the state-of-the-art by incorporating ligand-specific information through cross-attention, allowing generalizability across unseen ligands. LABind's compatibility with predicted protein structures like those from ESMFold enhances its applicability. Also, its applicability to the molecular docking and the use case study, show the importance of LABind.

The data analysis is thorough, and the authors effectively interpret LABind's comparative performance. The methodology's design appears well-suited to the goals of the study, with a balanced use of graph representation and cross-attention for integrating ligand information.

The methodology in this work is solid and aligns with the high standards of computational biology research. The integration of graph-based spatial features with protein embeddings and cross-attention mechanisms is well-justified, establishing a robust framework for binding site prediction. Additionally, the use of large pre-trained models, such as Ankh for protein embeddings and MolFormer for ligand representations, reflects a modern, data-driven approach in bioinformatics. This combination of approaches is effective in capturing complex ligand-protein interactions and is consistent with the current best practices in the field.

Major Concerns

Concern #1: Improper Use of 5-Fold Cross-Validation on DS3

The methodology described in the LABind paper raises significant concerns regarding the use of 5-fold cross-validation, particularly in the context of the DS3 dataset. The authors state that "We performed 5-fold cross-validation on the training data" but do not clarify whether this process was limited to hyperparameter optimization or if it has influenced the final evaluation. While 5-fold cross-validation is acceptable for DS1 and DS2, as it aligns with prior approaches in the field, its application to DS3 is scientifically unacceptable.

DS3 is specifically constructed with strict train/validation/test splits to evaluate the model's ability to generalize to unseen ligands. Performing 5-fold cross-validation on DS3's training data violates this structure, introducing potential contamination of the test set and invalidating the claims about generalization to unseen ligands. Furthermore, the presence of five checkpoint files under the (<https://zenodo.org/records/13938443>) → model/Unseen/ directory in your shared weights repository confirms that 5-fold cross-validation was indeed applied to DS3, leaving no doubt that this practice undermines the dataset's intended purpose.

To address these issues, the authors need to:

- Clearly specify how 5-fold cross-validation was applied and justify its use on DS3, given the dataset's explicit design for generalization testing.
- Provide assurance that the predefined test set in DS3 was entirely untouched during training and hyperparameter tuning.

Without this clarification, the claims about LABind's ability to generalize to unseen ligands lack credibility. Applying 5-fold cross-validation to DS3 represents a critical methodological flaw that compromises the validity of the reported results and could mislead the field about LABind's actual performance on unseen ligands.

Concern #2: Unfair Comparison with Competing Methods

The methodology used for benchmarking LABind against competing methods raises serious concerns regarding fairness and validity. While LABind was specifically trained and evaluated on DS1, DS2, and DS3 (see major concern 1), it appears that the authors did not retrain comparative methods like LigBind, DeepSurf, DeepPocket, P2Rank etc. on the same dataset splits. Instead, they relied on pre-trained best weights on other datasets or webserver outputs for evaluation, as written in the Supplementary S2. This is a critical and severe issue for several reasons:

1. Unfair Comparison:

- LABind's results reflect models optimized for the specific dataset splits (e.g., DS3). In contrast, methods like DeepSurf were not retrained on DS1,2,3 but instead evaluated using pre-trained weights or the LigBind webserver, which likely uses models trained on other datasets. This creates a biased comparison, giving LABind an undue advantage.

- i. E.g DeepSurf was trained on scPDB and tested on COACH420, the HOLO4K datasets. But on the supplementary S2 you claim:

"We used its standalone program with pre-trained model weights from <https://github.com/stemylonas/DeepSurf> for inference and evaluation."

How is it fair to use pre-trained weights trained on other datasets to derive results DS1,2,3 test sets?

2. Problem with LigBind and DS3:

- The authors explicitly state that LigBind was evaluated on DS3 using its webserver. This is inconsistent with LABind's setup, where the model was specifically trained on DS3 splits. For a fair comparison, LigBind should have been retrained on the DS3 training set and evaluated on the DS3 test set, just as LABind was.

3. Standard Practice in Comparative Studies:

- To claim state-of-the-art performance, all methods must be retrained or evaluated under the same conditions and datasets. Using pre-trained weights or outputs from a webserver for some methods while training LABind on the same datasets violates this standard and invalidates the comparison.

Minor Issues and Suggestions:

- On Table 1 we see the results for 19 ligands derived from DS1 test set. But upon checking the GPSite paper: Ying Xia, Xiaoyong Pan, and Hong-Bin Shen. LigBind: Identifying Binding Residues for Over 1000 Ligands with Relation-Aware Graph Neural Networks. Journal of Molecular Biology, 435(13):168091, 2023.

on their Table 2 they say that this is the MCC comparison for 21 ligands not 19. Why did you omit Nucleic acids and peptides?

- Using the term "ensemble learning" is misleading, as it implies a set of models, whereas LABind is a unified model that dynamically adapts based on input ligand information. A more accurate description might be "a unified model with shared and ligand-specific representations" or simply a "multi-ligand capable model."

- Page 2: "But LMetalSite and GPSite are limited to specific ligands".

Reference without citation, nowhere in the paper!

- Page 3: "Firstly, LABind outperforms other advanced methods in various benchmark datasets (DS1, DS2, and DS3)." DS1, DS2, and DS3 are mentioned for the first time without any link or short reference to what they are. Somehow link them to 3.1 where you explain these datasets in detail

- Page 3: "Through ablation studies, we the importance of each protein feature source in LABind"

Verb is missing after the "We".

- Page 4: "The sequence and structure of the protein receptor are input into Ankh and DSSP, respectively, to obtain the protein embedding and DSSP feature, which are then combined to form the protein compound embedding."

Aknh and DSSP mentioned for the first time without any citation.

- On 3.1.1 line 2 you mention 16 ligands. I think the correct is 19

- On 3.1..3. based onligands -> based on ligands

- On 3.2.5 multi parallel crossattention -> multi parallel cross attention

(Remarks on code availability)

Comments #1

Instead of this manual download process you could have provided a short script for someone to download the pre-trained weights, like the one below

"download_pretrained_weights.py"

```
from transformers import AutoModel, AutoTokenizer
import torch
import os
```

```
user = os.getenv('USER')
```

```
# Define model names and paths
```

```
models = {
```

```
"facebook/esmfold_v1": f"/home/{user}/PycharmProjects/LABind/scripts/ESMFold_v1",
```

```
"ElnaggarLab/ankh-large": f"/home/{user}/PycharmProjects/LABind/scripts/Ankh-large",
"ibm/MolFormer-XL-both-10pct": f"/home/{user}/PycharmProjects/LABind/scripts/MolFormer-XL-both-10pct"
}
```

```
# Create directories and download models
for model_name, save_path in models.items():
# Create the directory if it does not exist
os.makedirs(save_path, exist_ok=True)
print(f"Downloading model '{model_name}' to '{save_path}'")

# Download and save model and tokenizer
model = AutoModel.from_pretrained(model_name, trust_remote_code=True)
tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)

# Save model and tokenizer locally
for param in model.parameters():
param.data = param.data.contiguous() # Make tensors contiguous

# Save in PyTorch format instead of safetensors
model.save_pretrained(save_path, safe_serialization=False)
tokenizer.save_pretrained(save_path)

print("All models downloaded successfully.")
```

Comments #2

Update the README file to provide a clear path as to how someone can reproduce the paper's results. Also it is written in the README file

"It is also necessary to install three pre-trained models: ESMFold_v1, Ankh-large, and MolFormer-XL-both-10pct. We use the pre-trained weights from HuggingFace for prediction. Please download them to your device and modify the corresponding paths in scripts/config.py."

When the weights are downloaded what script should i execute in order to reproduce the results?

Please provide an end-to-end script or workflow to reproduce the main results.

Reviewer #2

(Remarks to the Author)

To date, discovery medicinal chemistry has been driven by iterative reduction, starting with the natural ligand, then its business end, to a peptide to small molecule mimetics. This represents opportunity cost in arriving at lead candidate libraries. The authors have identified LABind, a universal acceptor of sorts, that can prove disruptive in medicinal chemistry. They have rigorously tested the model using a validation set viz. the deadly SARS cov.

Technical Comments:

Elaborate on "Cross-attention" mechanism and "ablation". These are not clear.

"Due to the significant disparity between the number of binding and non-binding sites, MCC and AUPR are more reflective of the performance of model. In addition to AUC and AUPR, which are not affected by threshold selection, other metrics are calculated using a threshold to convert the predicted binding probabilities to binary predictions, which is determined by maximizing the MCC value. Therefore, we focused more on the AUPR performance of the model."

"We focused more"...is not objective. In fact, this is the basis of the findings.

"This result demonstrates that some protein-ligand binding samples in the DS1 training set may be insensitive to the ligand. LA" What is the physicochemical basis for this disparity

Throughout the MS there is subjectivity asserted and takes way from the findings (which in many cases are somewhat incremental from existing algorithms. The use of non-scientific terminology must also be resisted.

Examples

For proteins without accurate atomic positions, we combined protein structure prediction methods to achieve sequence-based ligand binding sites prediction. In pursuit of enhanced results, we compared the efficacy of ESMFold and OmegaFold in predicting binding sites.

we splited protein chains

we conducted an multitude of experiments

showed a significant improvement over other methods

LABind

still performed admirably in terms of MCC and AUPR metrics

exhibited outstanding differentiation ability in

The success of LABind is not a mere coincidence, it is the outcome of meticulous design and careful training of the model's computational components. This

(Remarks on code availability)

Unable to comment

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I appreciate the detailed and thoughtful rebuttal. The major concerns I raised—specifically regarding the 5-fold cross-validation on DS3 and the fairness of the benchmarking procedure—have been adequately addressed. The clarification on the training strategy and confirmation that the DS3 test set remained untouched resolves the methodological concern. Additionally, your efforts to retrain or justify the use of pre-trained baselines, along with the sequence identity analysis, strengthen the fairness of the comparisons.

The improvements to reproducibility and manuscript clarity are also noted and appreciated.

(Remarks on code availability)

No more comments

Reviewer #2

(Remarks to the Author)

these seem to be serious lapses

(Remarks on code availability)

i am not eligible to review it

Reviewer #3

(Remarks to the Author)

The paper presents a model for binding site detection that is both ligand and protein aware. The architecture is reasonable, the application relevant, and the ablation studies are informative. I am not aware of similar approaches and overall find the approach interesting. The reviewers were responsive to reviewer feedback, but there are a few major issues that weaken the paper.

Description of training. The response to reviewers is much more clear than the actual paper as to how training was done. I suggest the authors further expand 4.1. Fig S3 should probably be promoted to the main paper and it needs to be more clear that test sets are sequence split from training. The training procedure itself is rigorous and appropriate, but not clearly described.

Inadequate evaluation. A variety of classification metrics are described, but it is never stated what is being classified. The supplementary note S1 simply says "binding sites" without defining what this means. Looking at the GitHub I believe per-residue classification is being performed. This raises some concerns with the comparison with other methods as they may have different notions of what qualifies as a binding site residue. Some methods (e.g. DeepPocket) do not even produce per-residue scores as they are predicting the location of a whole pocket. The description of evaluation metrics needs to be expanded. A stronger paper would evaluate LABind using other tool's metrics (e.g. how well pockets are localized).

Nearly exclusive use of co-crystals. Most of the evaluations are done using holo structures that are perfect images for their cognate ligands. A stronger paper would evaluate DS3 using apo and/or predicted structures (it is not an interesting problem

to identify a binding site from a holo structure). Fig 3 is encouraging in this regard, but is only applicable to a small number of ligands. More data is needed to support the conclusion of 2.4.

More minor issues:

Page 3. "Given the relatively fixed structure of small molecules..." I don't know what is trying to be said here - small molecules have greater diversity of chemical structures than macromolecules formed from a limited number of building blocks.

Page 7. You use "picking set" without defining what it is.

2.4. "we filtered out predicted structures with a TM-score < 0.5 " Are you saying that 3b is only showing proteins where the TM-score is > 0.5 ? Why is 3d only showing TM-score < 0.5 proteins? I don't see how the figure supports the claim (AUPR distribution of ESMFold is superior to that of OmegaFold) - 3a shows most examples have TM-score > 0.5 so wouldn't those dominate the AUPR distribution (which is not shown)?

"For proteins with accurate native structures" - so they have inaccurate native structures? Do you mean without experimentally determined structures?

2.6. "we disregarded complexes identified by LABind as having no binding sites" All complexes in CASF-2016 have binding sites, so these examples are failures of LABind and should be acknowledged and quantified as such. Need to specify that you are doing whole protein docking with a box created around the whole protein. Unclear if multiple binding sites are docked to or only one. Would be interesting to know how "Vina with binding sites" compares to docking using the cognate ligand (e.g., autobox_ligand in smina/gnina).

2.8 "protein compound embeddings" This was confusing since "compound" can be synonymous with ligand. Suggest changing to protein-DSSP embedding

Figure 7b. This figure is inadequately labeled. Need a legend where it is clear what the different colors/styles are.

Page 16. "Sequence identity was reduced to 30%". I don't know what is being said here. Identity with what? How reduced. "Proteins with protein sequence identity below 30%..." identity with what? (presumably test set, but say it)

Page 18. "Gaussian radial basis function to compute the distances". RBFs don't compute distances, they encode them.

Page 19. W_O is defined but not used in any equation.

(Remarks on code availability)

I skimmed the repository, but did not run anything.

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

The authors were highly responsive to reviewer feedback and the result is a much stronger paper worthy of publication. Copy editing for English language usage would still be helpful.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to Reviewer #1

We sincerely thank the reviewer for the comments and suggestions. Before we respond to your comments, we need to clarify an issue that existed in the previous manuscript.

First of all, we are sorry that we misled you due to the misuse of the term “5-fold cross-validation”, which resulted in a series of related concerns. Instead, in this revised version, “5-fold average training” is the correct phrase to describe our training framework for DS3 model, which is briefly described here.

The DS3 dataset was divided into training/picking/test sets following diversity criteria, while the test set was kept aside and completely unseen during the training stage. A training set was partitioned into five folds. For a same set of hyperparameters, five models were trained on four folds and validated on the remaining fold. Notably here, the validation was only performed on the fold against those other four training folds based only on the training set and had not touched a bit of the aside test set. The average validation performance of the five convergent models was then used to determine the best hyperparameters via grid search of different sets of hyperparameters. The final trained models with the optimized hyperparameters were used in the prediction task on the test set. The picking set (“validation set” used in the previous manuscript might be another misleading term) of DS3 dataset is only used to determine the cutoff thresholds for binary classification on DS3 test set when certain evaluation metrics (such as MCC) are required. Again, we emphasize here that the picking set and the training set are entirely different from the test set.

For training DS1 and DS2 models, we strictly followed the same protocols as those of the comparing systems to ensure fair comparisons.

Based on the above brief, we think that “5-fold average training” is now the best expression for describing our training strategy. For predicting the test set, using the average score from the outputs of five trained models, each of which was trained on 4-fold among five folds of training set with the same set of hyperparameters, can be thought of as the simplest application of ensemble learning. We hope this brief will help to clarify the following detailed responses to Comment 1, 2, 4.

1. The Reviewer comments that

Concern #1: Improper Use of 5-Fold Cross-Validation on DS3

The methodology described in the LABind paper raises significant concerns regarding the use of 5-fold cross-validation, particularly in the context of the DS3 dataset. The authors state that “We performed 5-fold cross-validation on the training data” but do not clarify whether this process was limited to hyperparameter optimization or if it has influenced the final evaluation. While 5-fold cross-validation is acceptable for DS1 and DS2, as it aligns with prior approaches in the field, its application to DS3 is scientifically unacceptable.

DS3 is specifically constructed with strict train/validation/test splits to evaluate the model’s ability to generalize to unseen ligands. Performing 5-fold cross-validation on DS3’s training data violates this structure, introducing potential contamination of the test set and invalidating the claims about generalization to unseen ligands. Furthermore, the presence of five checkpoint files under the (<https://zenodo.org/records/13938443>) → model/Unseen/ directory in your shared weights repository confirms that 5-fold cross-validation was indeed applied to DS3, leaving no doubt that this practice undermines the dataset’s intended purpose.

To address these issues, the authors need to:

- *Clearly specify how 5-fold cross-validation was applied and justify its use on DS3, given the dataset's explicit design for generalization testing.*
- *Provide assurance that the predefined test set in DS3 was entirely untouched during training and hyperparameter tuning.*

Without this clarification, the claims about LABind's ability to generalize to unseen ligands lack credibility. Applying 5-fold cross-validation to DS3 represents a critical methodological flaw that compromises the validity of the reported results and could mislead the field about LABind's actual performance on unseen ligands.

Response:

We sincerely appreciate the reviewer's insightful comments and acknowledge the need for corrections in our description. We apologize for the misuse of the term "5-fold cross-validation", which cause some misunderstandings. In our methodology, we only partitioned the training dataset, while the test set remained entirely independent.

To clarify, the proteins in the DS3 test set share less than 30% sequence identity with those in the training and picking sets, and none of the ligands in the test set appear in either the training or picking sets. The DS3 training set was partitioned into five equal parts, with four parts used for training and the remaining one for validation. This process was repeated five times to ensure robust training and validation. Importantly, only the picking set of DS3 was used to determine the threshold for binary classification. The predefined DS3 test set remained entirely untouched during both training and hyperparameter tuning and was reserved exclusively for the final evaluation. We have added a description to provide a more detailed introduction to the role of each set in the section "DS3: new assembled dataset for ligand-aware model" on page 16:

... We use the training set to train the model's weights, the picking set to determine the threshold for binary classification, and the test set to evaluate the model's generalization performance on unseen ligands.

Additionally, to enhance clarity and accuracy, we have revised the relevant description in the section "Training Procedures" on page 20:

The proteins in the DS1, DS2, and DS3 test sets share less than 30% sequence identity with those in the training sets. Additionally, in DS3 dataset, none of the ligands in the test set appear in the training set. Consistent with previous work [16, 17], we adopted a '5-fold average training' strategy to randomly partition the training set into five folds. For each round, the model was trained on four folds and validated on the remaining one, resulting in five complete training and validation cycles. Notably, during training, the benchmark test sets remained completely unseen by the model. During testing, the final LABind predictions were obtained by averaging the outputs from all five trained models.

We specify that the DS3 test set is intended solely for independent testing in the section "Prediction accuracy of LABind on benchmark" on page 8:

The hyperparameter settings for the DS3 training set were identical to those used for the DS1 and DS2 training sets. The DS3 picking set was used solely to determine the binary classification threshold based on the model's outputs, ensuring that no information from the DS3 test set influenced the training process.

By explicitly stating these points, we confirm that the DS3 test set was never used in training, or hyperparameter tuning, preserving its role as an independent benchmark for evaluating LABind's generalization to unseen ligands.

2. The Reviewer comments that

Concern #2: Unfair Comparison with Competing Methods

The methodology used for benchmarking LABind against competing methods raises serious concerns regarding fairness and validity. While LABind was specifically trained and evaluated on DS1, DS2, and DS3 (see major concern 1), it appears that the authors did not retrain comparative methods like LigBind, DeepSurf, DeepPocket, P2Rank etc. on the same dataset splits. Instead, they relied on pre-trained best weights on other datasets or webserver outputs for evaluation, as written in the Supplementary S2. This is a critical and severe issue for several reasons:

1. Unfair Comparison:

○ LABind's results reflect models optimized for the specific dataset splits (e.g., DS3). In contrast, methods like DeepSurf were not retrained on DS1,2,3 but instead evaluated using pre-trained weights or the LigBind webserver, which likely uses models trained on other datasets. This creates a biased comparison, giving LABind an undue advantage.

i. E.g DeepSurf was trained on scPDB and tested on COACH420, the HOLO4K datasets. But on the supplementary S2 you claim:

“We used its standalone program with pre-trained model weights from <https://github.com/stemylonas/DeepSurf> for inference and evaluation.”

How is it fair to use pre-trained weights trained on other datasets to derive results DS1,2,3 test sets?

2. Problem with LigBind and DS3:

○ The authors explicitly state that LigBind was evaluated on DS3 using its webserver. This is inconsistent with LABind's setup, where the model was specifically trained on DS3 splits. For a fair comparison, LigBind should have been retrained on the DS3 training set and evaluated on the DS3 test set, just as LABind was.

3. Standard Practice in Comparative Studies:

○ To claim state-of-the-art performance, all methods must be retrained or evaluated under the same conditions and datasets. Using pre-trained weights or outputs from a webserver for some methods while training LABind on the same datasets violates this standard and invalidates the comparison.

Response:

Thank you for your valuable feedback. We appreciate your concern regarding the fairness and validity of the comparison between LABind and other competing methods. Due to the misuse of 5-fold cross-validation, we declare that the DS3 test set was kept entirely independent during training. To ensure a fair comparison, we retrained DeepPocket on the DS3 dataset and removed protein sequences from both the training and picking sets that shared more than 30% sequence identity with those in the test set. We also ensured that the ligands in the test set do not appear in the training set and the picking set. We acknowledge the points you raised and have made the following clarifications and adjustments to address the issues:

(1) Unfair comparison with pre-trained weights (for comments 2.1 and 2.2)

As pointed out, methods such as LigBind, DeepSurf, and P2Rank were not retrained on the same dataset splits as LABind, and instead, pre-trained weights or webserver outputs were used. We recognize this discrepancy, and we would like to clarify the reasoning behind it.

- (a) LigBind and DeepSurf: Unfortunately, neither LigBind nor DeepSurf provide retraining scripts for their models; they only offer pre-trained models for inference. Consequently, for these methods, we had no choice but to use the pre-trained weights provided by their respective authors or webserver.
- (b) DeepPocket and P2Rank: Unlike the previous methods, both DeepPocket and P2Rank provided retraining scripts in their public repositories. We made an attempt to retrain P2Rank and DeepPocket using the DS3 training set.
 - b.1) However, during the retraining of P2Rank, we encountered non-convergence issues, a problem also observed in related work (e.g., in EquiPocket: E(3)-Equivariant Graph Neural Networks for Ligand Binding Site Prediction, DOI: <https://dl.acm.org/doi/10.5555/3692070.3694552>, Page:16). As a result, we were forced to use the pre-trained weights from the official P2Rank repository for the evaluation.
 - b.1) For DeepPocket, we followed the default training procedure from the public repository, but since DeepPocket requires the 3D structure of protein-ligand complexes, and the BioLiP2 database did not provide ligand positions, we downloaded the required PDB files from the RCSB database to complete the process.

(2) Ensuring Fairness in Evaluation (for comment 2.3)

As DS1 and DS2 are well-established classic datasets, our processing methodology aligns with prior approaches in the field. Therefore, the comparison results for these datasets are directly taken from previously published papers. Specifically, the DS1 results come from LigBind, and the DS2 results come from GPSite. In contrast, DS3 is a newly constructed dataset designed to test results on unseen ligands. Thus, the comparison in this work involves the four methods mentioned earlier: LigBind, DeepSurf, P2Rank, and DeepPocket. Due to the aforementioned limitations, LigBind, DeepSurf, and P2Rank could only be evaluated using the pre-trained weights provided by their official websites, while only DeepPocket was retrained under the same conditions and datasets as LABind.

To mitigate potential bias and ensure a fairer comparison of LABind with the other methods, we took several steps to address the issues in the DS3 dataset:

- (a) Revised DS3 Training and Test Sets for LABind: We re-evaluated and revised the DS3 training and test sets to remove proteins from the training and picking sets that shared more than 30% sequence identity with any protein in the DS3 test set. This ensures that the proteins in the DS3 test set represent unseen ligands, making the test more challenging and rigorous for LABind, and reflecting the true generalization capability of the model.
- (b) Sequence Identity Analysis: we also analyzed the sequence identity between the DS3 test set and the training datasets of the competing methods:
 - b.1) DeepSurf: Trained on the scPDB. Upon conducting sequence alignment with MMseqs2, we found that 481 proteins in the DS3 test set had >30% sequence identity with proteins in scPDB, and 294 proteins had >80% sequence identity.
 - b.1) P2Rank: Trained on the CHEN11 and JOINED datasets, P2Rank's sequences

showed >30% sequence identity in 207 cases and >80% sequence identity in 79 cases with the DS3 test set.

b.1) LigBind: LigBind's training set showed >30% sequence identity in 808 proteins and >80% in 483 proteins when compared with the DS3 test set.

Given this information, using pre-trained models based on datasets with overlapping sequence identities could introduce biases and potentially give unfair advantages to the competing methods. However, LABind ensures fairness by making sure that no proteins in the DS3 test set share more than 30% sequence identity with those in the DS3 training set, and none of the ligands in the test set appear in the training set. This approach minimizes potential biases and ensures a fairer comparison.

(3) Updating manuscript

Given the aforementioned issues, we have updated the descriptions in both the manuscript and the supplementary information.

(a) We modified the construction process of the DS3 dataset to ensure that all proteins in the DS3 test set are also sufficiently different. we have revised the relevant description in the section "DS3: new assembled dataset for ligand-aware model" on page 16:

As shown in Supplementary Fig. S3, we removed protein chains with resolutions of > 3.0 Å and lengths of > 1500. Similar to previous studies [14, 17], we used MMseqs2 [41] to remove redundancy from receptor proteins. Proteins with protein sequence identity below 30% and alignment coverage greater than 30% were retained. Next, we used protein data released before January 1, 2021, for the training set. Data released after this date, where ligands appeared in the training set, were used as the picking set, while data released after this date with ligands not seen in the training set were used as the test set. Finally, we removed protein sequences from the training and picking sets that shared more than 30% sequence identity with the protein sequences in the test set. This approach ensures that the ligands in the DS3 test set were not seen in the training set and that the protein structures are sufficiently distinct, thereby enabling a better assessment of the model's generalization capability on unseen data.

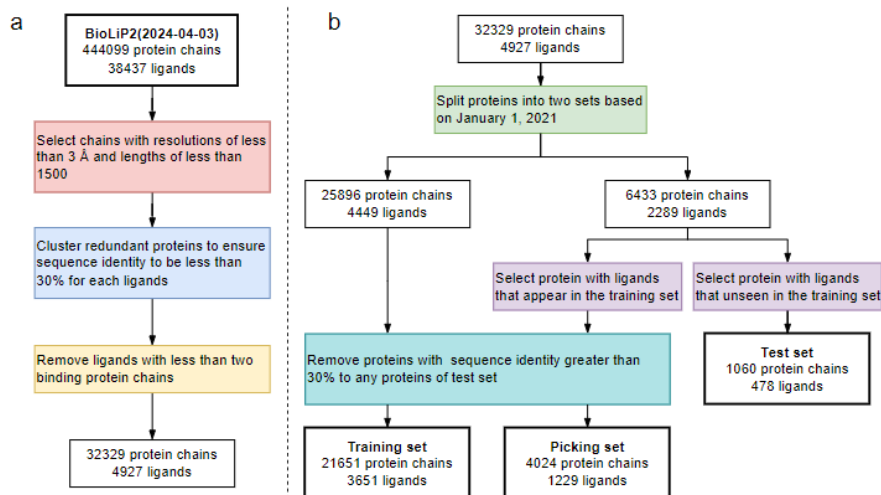


Fig. S3 The construction of training set, picking set and test set for DS3 dataset. (a) The screening process from the BioLiP2 database. (b) The process of splitting DS3 dataset.

(b) We have revised the description of the comparative methods to provide explanations for why we did not retrain them. In the supplementary information section “Note S2 Brief introductions to the competitive methods” on page 9-10:

P2Rank. ...To ensure a fair comparison, we attempted to retrain P2Rank on the DS3 training set. However, we found that P2Rank encountered convergence issues on the DS3 training set. This issue also occurred in [9]. Therefore, we used its standalone program from <https://github.com/rdk/p2rank> for inference and evaluation.

DeepSurf. ...Due to the lack of retraining code in the official DeepSurf repository, we could not retrain DeepSurf. We used its standalone program with pre-trained model weights from <https://github.com/stemylonas/DeepSurf> for inference and evaluation.

DeepPocket. ...We retrained DeepPocket using the default settings from <https://github.com/devalab/DeepPocket>. Since DeepPocket requires the 3D structures of protein-ligand complexes, and the BioLiP2 database did not provide ligand positions, we downloaded the required PDB files from the RCSB database to complete the process.

LigBind. ...Similar to DeepSurf, LigBind also does not provide the code for training. We used its webserver (<http://www.csbio.sjtu.edu.cn/bioinf/LigBind/index.html>) for evaluation on the DS3 dataset and standalone program with pre-trained model weights from <http://www.csbio.sjtu.edu.cn/bioinf/LigBind/sourcecode.html> for evaluation on the DS2 dataset.

(c) We have revised the description of evaluating LABind on the DS3 dataset and the description regarding the sequence identity between the datasets used by the comparative methods and the DS3 test set. In the section “Prediction accuracy of LABind on benchmark” on page 8:

As shown in Supplementary Note S3, P2Rank, DeepSurf, DeepPocket, and LigBind use pre-trained models based on their training datasets, which have high sequence identity compared to the DS3 test set. This could introduce biases and potentially give unfair advantages to the competing methods. However, LABind still outperforms these methods on the DS3 test set, demonstrating the better generalization capability of the LABind model on unseen data.

(d) Revisions in the supplementary information: We have added a description of the sequence identity between the different training sets and the DS3 test set in the supplementary information section “Note S3 The sequence identity between the DS3 test set and other datasets” on page 10:

We used mmseqs easy-search [15] to align all the training sets used by P2Rank, DeepSurf, DeepPocket, and LigBind with the DS3 test set. We used the number of protein sequences in the DS3 test set that have a sequence identity exceeding the threshold with sequences in the other training sets to represent the degree of similarity. The sequence identity between the DS3 test set and the training sets of P2Rank, DeepSurf, DeepPocket, and LigBind is shown in Table S6. Among all 1,049 proteins in the DS3 test set, there are 481, 44, 163, and 808 protein sequences that share more than 30% sequence identity with the scPDB, CHEN11, JOINED, and LigBind datasets, respectively. Additionally, there are 294, 20, 59, and 483 protein sequences that share more than 80% sequence identity with these datasets, respectively. By removing sequences from the DS3 training and picking sets that share more than 30% sequence identity with those in the DS3 test set, and ensuring that none of the ligands in the test set appear in the training set, this approach ensures that the comparison is not biased towards our method LABind.

Table S6 The sequence identity between the DS3 test set and other datasets.

Training set	$N_{>30\%}$	$N_{>80\%}$
scPDB	481	294
CHEN11	44	20
JOINED	163	59
LigBind	808	483

Note: $N_{>30\%}$ and $N_{>80\%}$ denote the number of sequences in the DS3 test set with sequence identity greater than 30% and 80%, respectively.

3. The Reviewer comments that

On Table 1 we see the results for 19 ligands derived from DS1 test set. But upon checking the GPSite paper:

Ying Xia, Xiaoyong Pan, and Hong-Bin Shen. LigBind: Identifying Binding Residues for Over 1000 Ligands with Relation-Aware Graph Neural Networks. Journal of Molecular Biology, 435(13):168091, 2023.

on their Table 2 they say that this is the MCC comparison for 21 ligands not 19. Why did you omit Nucleic acids and peptides?

Response:

Thank you for raising this question. When the ligand is a macromolecule (nucleic acids or peptides), the distinct properties of the ligand itself can also affect the binding sites. This can introduce ambiguity when predicting protein binding sites for different structural types of macromolecular ligands. For example, the protein (PDB ID: 1a2c, chain: H) can form binding sites with two different peptides (sequence: tLPR and sequence: DFEEIPEEY). However, due to the differences in peptide sequences, the binding sites are not consistent. Our goal is to identify all binding sites within a protein for a given ligand. If each macromolecule is characterized using an individual vector representation, the significant differences in binding mechanisms between macromolecules and small molecules may lead to suboptimal results. In LABind, such an approach can only treat macromolecules as identical features, which we believe this is unreasonable. Therefore, LABind excludes the recognition of binding sites for proteins with macromolecules.

We have added descriptions to eliminate this doubt in the section “Introduction” on page 3:

Considering that the physical structure of macromolecules affects protein binding sites, previous studies (LMetalSite, LigBind, and GPSite) combine binding sites of the same protein with structurally different macromolecules of the same type into a single sample. Given the relatively fixed structure of small molecules and the wide structural diversity of macromolecules, macromolecules exhibit much different properties comparing with protein-small molecule binding [21]. In this study, we focused on predicting protein binding sites for small molecules and ions.

Furthermore, we have addressed the limitations of LABind in the “Discussion” section on page 15.

... However, LABind currently lacks the capability to identify binding sites of protein-

macromolecule binding sites. In summary, we believe that LABind provides certain assistance in the work related to protein binding sites, and it can offer help in understanding protein functions and in the field of drug design.

4. The Reviewer comments that

Using the term "ensemble learning" is misleading, as it implies a set of models, whereas LABind is a unified model that dynamically adapts based on input ligand information. A more accurate description might be "a unified model with shared and ligand-specific representations" or simply a "multi-ligand capable model."

Response:

Thank you for your suggestions for better improvement. We have updated all occurrences of "ensemble learning" in the manuscript to "multi-ligand capable model."

In the section "Introduction" on page 3:

This multi-ligand capable model enables the model to learn representations shared between different ligand binding sites, while also learning representations specific to each ligand binding sites.

In the section "Discussion" on page 15:

This multi-ligand capable model captures both shared representations across different ligand binding sites and distinct representations specific to each binding site.

5. The Reviewer comments that

*Page 2: "But LMetalSite and GPSite are limited to specific ligands".
Reference without citation, nowhere in the paper!*

Response:

Thank you for pointing out the error. We have added references and thoroughly reviewed the manuscript to ensure that such issues are avoided. The revisions are as follows:

In the section "Introduction" on page 2:

LMetalSite[16] and GPSite[17] utilize multi-task learning to combine multiple ligand datasets and train a single model for multiple specific ligands.

6. The Reviewer comments that

Page 3: . "Firstly, LABind outperforms other advanced methods in various benchmark datasets (DS1, DS2, and DS3)."

DS1, DS2, and DS3 are mentioned for the first time without any link or short reference to what they are. Somehow link them to 3.1 where you explain these datasets in detail

Response:

Following the reviewer's suggestion, we have added a link to section "Benchmark datasets" in the section "Introduction" on page 3:

Firstly, LABind outperforms other advanced methods in various benchmark datasets (DS1, DS2, and DS3, see "Benchmark datasets" for dataset details).

7. The Reviewer comments that

*Page 3: "Through ablation studies, we the importance of each protein feature source in LABind"
Verb is missing after the "We".*

Response:

Thank you for pointing out the error. We have added the verb after "We" in the section "Introduction" on page 3:

Through ablation studies, we demonstrated the importance of each protein feature source in LABind, with the protein representation proving to be crucial and the positive impact of ligand features also being evident.

8. The Reviewer comments that

*Page 4: "The sequence and structure of the protein receptor are input into Ankh and DSSP, respectively, to obtain the protein embedding and DSSP feature, which are then combined to form the protein compound embedding."
Ankh and DSSP mentioned for the first time without any citation.*

Response:

Thank you for pointing out the error. We have meticulously reviewed the manuscript to prevent such issues and have added references in the section "Overview of LABind" on page 4:

The sequence and structure of the protein receptor are input into Ankh and DSSP [27], respectively, to obtain the protein embedding and DSSP feature, which are then combined to form the protein compound embedding.

9. The Reviewer comments that

On 3.1.1 line 2 you mention 16 ligands. I think the correct is 19

Response:

Regarding the issue you raised, LigBind created ligand-specific sets, pre-training sets,

unseen sets of 16 ligands. The DS1 dataset was constructed using 19 ligands from the ligand-specific set.

To eliminate any potential misunderstandings arising from the descriptions, we have revised the descriptions of DS1 and DS2 dataset. In the section “DS1: dataset for literature single-ligand-oriented methods” on page 15:

The dataset created by LigBind contains ligand-specific sets of 1159 ligands, pre-training sets consisting of 1301 ligands and unseen sets of 16 ligands. We selected 19 ligands from LigBind’s ligand-specific dataset to construct the DS1 benchmark dataset. The details of DS1 dataset are described in Supplementary Note S5.

In the section “DS2: dataset for multi-ligand-oriented methods” on Page 16:

The dataset is composed of 10 ligands including biomacromolecules (DNA, RNA, peptide, protein), small molecules (ATP, HEM) and metal ions (Zn^{2+} , Ca^{2+} , Mg^{2+} , Mn^{2+}). We focus on the identification of binding sites for proteins with small molecules and ions. Therefore, we excluded biomacromolecules to construct the DS2 dataset, which ultimately includes six ligands (ATP, HEM, Zn^{2+} , Ca^{2+} , Mg^{2+} , and Mn^{2+}).

10. The Reviewer comments that

On 3.1..3. based onligands -> based on ligands

Response:

Thank you for pointing out the error. And following another reviewer's suggestion, we have revised the original sentence, primarily by removing the ambiguous term “we splited protein chains into multiple sets based onligands”. We have revised this description in the section “DS3: new assembled dataset for ligand-aware model” on page 16:

We removed ligands with fewer than two binding protein chains in the training, picking, and test sets to reduce potential bias introduced by individual samples.

11. The Reviewer comments that

On 3.2.5 multi parallel crossattention -> multi parallel cross attention

Response:

Thank you for pointing out the error. We have revised this description in the section “Attention-based learning interaction” on page 19:

In this work, we employ the multi-head cross-attention mechanism with multi attention heads, so the outputs of multi parallel cross-attention layers are concatenated together.

12. The Reviewer comments that

*Instead of this manual download process you could have provided a short script for someone to download the pre-trained weights, like the one below
“download_pretrained_weights.py”*

Response:

Thank you very much for this suggestion. This is a good point. We have added a script for downloading pre-trained weights to the GitHub repository (scripts/download_weights.py). Users can run the script to download the pre-trained models required for LABind, including Ankh, MolFormer, and ESMFold. We hope that this file will meet users' needs and help save time.

We have updated the README file description as follows:

You can also use python ./scripts/download_weights.py -o <path> to download model weights or to automatically download to the default path during prediction.

13. The Reviewer comments that

Update the README file to provide a clear path as to how someone can reproduce the paper's results. Also it is written in the README file

“It is also necessary to install three pre-trained models: ESMFold_v1, Ankh-large, and MolFormer-XL-both-10pct. We use the pre-trained weights from HuggingFace for prediction. Please download them to your device and modify the corresponding paths in scripts/config.py.”

When the weights are downloaded what script should i execute in order to reproduce the results?

Please provide an end-to-end script or workflow to reproduce the main results.

Response:

Thank you very much for this suggestion. We have uploaded a jupyter notebook file for reproducing the results of the paper (scripts/reproduction.ipynb). Additionally, we have added a Python file utilizing the wandb library for hyperparameter optimization, facilitating the retraining of LABind by others (scripts/wandb.py). We have also provided a script for creating the dataset based on the BioLiP2 database (scripts/create_ds.py). We updated the README file. The updated description is as follows:

Retrain

We have provided a detailed workflow to reproduce the results presented in the paper, using training and validation based on DS3 as an illustrative example. (./scripts/reproduction.ipynb) If you need to retrain the model, you can try modifying ./scripts/wandb.py for training, which requires installing the wandb package. pip install wandb

If you need to create a dataset based on the latest version of the [BioLiP2](#) database, you can use ./scripts/create_ds.py to generate a new dataset.
python create_ds.py -o <out_path> -m <max_length>

Response to Reviewer #2

We thank the reviewer for the comments and suggestions. The following changes have been made to address the issues.

1. The Reviewer comments that

Elaborate on "Cross-attention" mechanism and "ablation". These are not clear.

Response:

We appreciate the reviewer's suggestion. We have revised the manuscript to provide a clearer explanation of cross-attention and ablation. The revisions are as follows:

In the section "Attention-based learning interaction" on page 19:

(ii) Multi-head cross-attention in protein-ligand interactions. In order to achieve information interaction between ligand representations and protein representations, we utilize cross-attention for information exchange. The cross-attention mechanism is implemented in Transformer [42] to allow each position in the decoder to attend to all positions within the input sequence. Previous work has demonstrated that the cross-attention mechanism can effectively combine different types of information [48]. The architecture and calculation of multi-head cross-attention is shown in Fig. 1d. Firstly, we use the residue representations as queries and multiply them by the matrix W_Q . The ligand representation is used as both keys and values, which are multiplied by matrices W_K and W_V , respectively. We then compute the dot product of the queries and keys to get the attention coefficients, which are transformed into a probability distribution using the softmax function. Lastly, the attention coefficients are applied to the ligand representation to focus on the relevant information. Formally, the definition of the cross-attention mechanism between proteins and ligands is as follows:

$$h''^{(l)} = h'^{(l)} + \alpha_x^{(l)} W_v^{(l)} x^{(l)}$$
$$\alpha_x^{(l)} = \text{softmax} \left(\frac{\left(W_q^{(l)} h'^{(l)} \right) \left(W_k^{(l)} x^{(l)} \right)^T}{\sqrt{d}} \right) \quad (5)$$

Here, x represents the ligand representation, $h''^{(l)}$ represents the updated residue representations, with the rest of the meanings being similar to those in the neighbor-attention mechanism. In this work, we employ the multi-head cross-attention mechanism with multi attention heads, so the outputs of multi parallel cross-attention layers are concatenated together.

In the section "Ablation study to verify the contribution of each protein feature source of LABind" on page 13-14:

We conducted ablation studies to verify the contribution of each protein feature source of LABind. As illustrated in Fig. 7a, we compared the impact of removing four main features—protein embeddings (Ankh features), residue-residue interactions of the protein representations, protein compound embeddings (Ankh features + DSSP features), and protein graphs on the performance of binding sites identification on the DS1 test sets of six different ligands for demonstration. We selected metal ions (Zn^{2+} and Ca^{2+}), anions (PO_4^{3-}), and small molecules (ATP, HEM, and GDP) and retained the average AUPR on the DS1 test set. Removing feature means that regardless of the input, we replaced the features with

a matrix of all ones. When removing the protein embeddings from LABind, the average AUPR decreased by 0.043 across all 19 ligands (average AUPR of 0.573). This indicates that the absence of protein sequence information leads to a decrease in LABind's performance. The removal of residue-residue interactions (retention of solely the positional residue-residue indices) results in an average decrease of 0.132 in AUPR of the predictive outcomes. The LABind results in inadequate performance (average AUPR of 0.484) by neglecting the relative positions of residues. The removal of protein compound features within LABind leads to a decrease of 0.048 in average AUPR. When protein graphs (node spatial features and edge spatial features) are removed, randomizing residue-residue interactions as well as the spatial information of residue atoms, LABind experiences the most significant decrease in AUPR. Removing protein graphs underperformed LABind by 0.132, 0.284, 0.135, 0.200, 0.200, 0.224 in AUPR on Zn^{2+} , Ca^{2+} , PO_4^{3-} , ATP, GDP and HEM test sets. Given the complete lack of protein graph information, the model leads to a significant decline in performance (average AUPR from 0.616 to 0.433). Hence, it is evident that protein graphs are of paramount importance within the LABind architecture.

2. The Reviewer comments that

"Due to the significant disparity between the number of binding and non-binding sites, MCC and AUPR are more reflective of the performance of model. In addition to AUC and AUPR, which are not affected by threshold selection, other metrics are calculated using a threshold to convert the predicted binding probabilities to binary predictions, which is determined by maximizing the MCC value. Therefore, we focused more on the AUPR performance of the model."

"We focused more" ...is not objective. In fact, this is the basis of the findings.

Response:

Thank you for your suggestion. We acknowledge that the previous description had issues, despite being based on earlier work. We revised it accordingly in the section "Evaluation metrics" on page 4.

Due to the highly imbalanced distribution and number of binding sites and non-binding sites, MCC and AUPR are more reflective of the performance of model in imbalanced two-class classification tasks [28, 29]. In addition to AUC and AUPR, which are not affected by threshold selection, other metrics are calculated using a threshold to convert the predicted binding probabilities to binary predictions, which is determined by maximizing the MCC value. Therefore, we adopt AUPR as the criterion for hyperparameter optimization and ablation study.

3. The Reviewer comments that

"This result demonstrates that some protein-ligand binding samples in the DS1 training set may be insensitive to the ligand. LA" What is the physicochemical basis for this disparity

Response:

We thank the reviewer for the valuable suggestion. The description led to a lack of scientific rigor in the manuscript. And we have checked the manuscript for similar issues and made corresponding revisions.

In the section “Comparing with multi-ligand-oriented methods” on page 6:

This result demonstrates that some protein-ligand binding samples in the DS1 training set are insensitive to the ligand. This is attributed to the fact that protein-ligand binding sites are generally located in pockets (clefts, grooves) on the surface of proteins [30], which have the capacity to interact with multiple ligands. Consequently, similar proteins with similar ligands can exhibit analogous binding sites [31].

We revised the phrase “This discrepancy may be due to the smaller size of metal ions and their binding tendencies, which differ from those of small molecule ligands.” to a more evidence-based description in the section “Prediction accuracy of LABind on benchmark” on page 7:

Due to the smaller size of metal ions, the interactions of the metal ions with proteins are often found significantly more versatile and flexible compared with larger size ligands [32].

4. The Reviewer comments that

Throughout the MS there is subjectivity asserted and takes way from the findings (which in many cases are somewhat incremental from existing algorithms. The use of non-scientific terminology must also be resisted.

Examples

For proteins without accurate atomic positions, we combined protein structure prediction methods to achieve sequence-based ligand binding sites prediction. In pursuit of enhanced results, we compared the efficacy of ESMFold and OmegaFold in predicting binding sites.

we splited protein chains

we conducted an multitude of experiments

showed a significant improvement over other methods

LABind still performed admirably in terms of MCC and AUPR metrics

exhibited outstanding differentiation ability in

The success of LABind is not a mere coincidence, it is the outcome of meticulous design and careful training of the model’s computational components. This

Response:

Thank you very much for the reviewer's suggestions, which will help us present the results analysis in a more rigorous and scientific manner. We recognize that our writing included numerous instances where objectivity was compromised. We have conducted a thorough review of the manuscript and have made revisions to address these issues throughout the document. The specific modifications are as follows:

(1) *“For proteins without accurate atomic positions, we combined protein structure prediction methods to achieve sequence-based ligand binding sites prediction. In pursuit of enhanced results, we compared the efficacy of ESMFold and OmegaFold in predicting binding sites”*:

We revised the phrase “For proteins without accurate atomic positions, we combined protein structure prediction methods to achieve sequence-based ligand binding sites prediction. In pursuit of enhanced results, we compared the efficacy of ESMFold and OmegaFold in predicting binding sites.” to more accurate description on page 8.

For proteins without accurate atomic positions, protein structure prediction methods were combined to achieve sequence-based ligand binding site predictions. Specifically, we tested the effects of combining ESMFold and OmegaFold with LABind separately.

(2) *“we splited protein chains”*:

We have removed the ambiguous sentence “we splited protein chains into multiple sets based on ligands” and revised the original sentence on page 17.

We removed ligands with fewer than two binding protein chains in the training, picking, and test sets to reduce potential bias introduced by individual binding samples.

(3) *“we conducted an multitude of experiments”*:

We revised the phrase “To thoroughly analyze the performance of LABind, we conducted an multitude of experiments” to a formal description on page 15.

To thoroughly analyze the performance of LABind, we conducted comparative experiments on three benchmark datasets, demonstrations of the same protein binding with different ligands, auxiliary molecular docking experiments, predictions of binding sites for the SARS-CoV-2 NSP3 macrodomain protein with unseen ligands, ablation studies, and visualizations of hidden features.

(4) *“showed a significant improvement over other methods”*:

We revised the phrase “LABind showed a significant improvement over other methods.” to a formal description on page 15.

Initially, we compared LABind with other multi-ligand-oriented and single-ligand-oriented methods using the DS1 dataset, and LABind demonstrated an improvement of 0.109 in the average MCC compared to LigBind.

(5) *“LABind still performed admirably in terms of MCC and AUPR metrics”*:

We revised the phrase “LABind still performed admirably in terms of MCC and AUPR metrics,” to a more accurate description on page 15.

Despite the increased difficulty, compared to similar methods, LABind demonstrated an average improvement of at least 0.131 in MCC and 0.125 in F1, showed high precision and robustness.

(6) *“exhibited outstanding differentiation ability in”*:

We revised the phrase “exhibited outstanding differentiation ability” to a more accurate description on page 15.

Through the application of LABind in predicting binding sites for the same protein with various ligands, we demonstrate that LABind can accurately differentiate the binding sites among distinct ligands.

(7) “*The success of LABind is not a mere coincidence, it is the outcome of meticulous design and careful training of the model’s computational components. This*”:

We revised the phrase “The success of LABind is not a mere coincidence, it is the outcome of meticulous design and careful training of the model’s computational components. This” to a neutral description on page 15.

The features, design, and training of LABind have resulted in consistent performance improvements across three benchmark datasets

Additionally, we revised other parts of the manuscript where there were instances of subjectivity.

(8) We revised the phrase “We believe that when binding protein, macromolecules may exhibit much different properties comparing with protein-small molecule binding.” to more a more valid description in the section “Introduction” on page 3.

Considering that the physical structure of macromolecules affects protein binding sites, previous studies (LMetalSite, LigBind, and GPSite) combine binding sites of the same protein with structurally different macromolecules of the same type into a single sample. Given the relatively fixed structure of small molecules and the wide structural diversity of macromolecules, macromolecules exhibit much different properties comparing with protein-small molecule binding [21]. In this study, we focused on predicting protein binding sites for small molecules and ions.

(9) We revised the phrase “Overall, the results demonstrate that LABind achieves the best performance on unseen ligands.” to a more scientifically objective description on page 8.

As shown in Fig. 2b, on the whole binding sample, LABind exceeds P2rank, DeepSurf, DeepPocket and LigBind by 0.213, 0.151, 0.184 and 0.131 in MCC, and 0.250, 0.156, 0.200 and 0.125 in F1, respectively. Overall, the results demonstrate that LABind achieves performance improvement on unseen ligands compared to the comparative methods.

(10) We revised the phrase "To uncover the reasons behind the superior performance of LABind," to a more scientifically objective description on page 15.

To uncover the roles of input features and computational components in LABind, we conducted ablation studies and visualized residue representations derived from attention-based learning interactions.

5. The Reviewer comments that

(Remarks on code availability):

Unable to comment

Response:

Thank you for your comment. On Github, users can post new issues and engage in detailed discussions under the corresponding issues. In contrast, the Zenodo repository does not currently support open comments for interaction.

Response to Reviewer #1

We sincerely thank the reviewer for the valuable comments and helpful suggestions.

1. The Reviewer comments that

I appreciate the detailed and thoughtful rebuttal. The major concerns I raised—specifically regarding the 5-fold cross-validation on DS3 and the fairness of the benchmarking procedure—have been adequately addressed. The clarification on the training strategy and confirmation that the DS3 test set remained untouched resolves the methodological concern. Additionally, your efforts to retrain or justify the use of pre-trained baselines, along with the sequence identity analysis, strengthen the fairness of the comparisons.

The improvements to reproducibility and manuscript clarity are also noted and appreciated.

Response:

We sincerely thank the reviewer for their positive feedback and are glad that our revisions have addressed the concerns appropriately.

Response to Reviewer #2

We sincerely thank the reviewer for the valuable comments and helpful suggestions. The following changes have been made to address the issues raised.

1. The Reviewer comments that

these seem to be serious lapses

Response:

Thank you for your comments. We apologize for any confusion that may have arisen. However, we are not entirely sure which specific sections the phrase “these seem to be serious lapses” refers to. To address this concern thoroughly, we have carefully reviewed the manuscript in its entirety — including the main text, data, and source code — to identify and rectify any potential issues.

2. The Reviewer comments that

*(Remarks on code availability):
i am not eligible to review it*

Response:

Thank you for your comments. We try to understand this comments. We released our code and data to GitHub and Zenodo, as stated in the 1st version of our submitted manuscript, on Oct 16, 2024. And so far, the dataset published on Zenodo has been downloaded 144 times. We have promptly engaged an independent researcher to conduct a fresh installation of the GitHub repository in a new computational environment (Ubuntu 20.04 via WSL2). We are pleased to confirm that:

- (1) The environment was successfully configured following the README instructions.
- (2) All test cases produced expected outputs.

Response to Reviewer #3

We sincerely thank the reviewer for the valuable comments and helpful suggestions. The following changes have been made to address the issues raised.

1. The Reviewer comments that

Description of training. The response to reviewers is much more clear than the actual paper as to how training was done. I suggest the authors further expand 4.1. Fig S3 should probably be promoted to the main paper and it needs to be more clear that test sets are sequence split from training. The training procedure itself is rigorous and appropriate, but not clearly described.

Response:

We sincerely appreciate the reviewer's valuable suggestions, which have helped us to present the experimental design in a more rigorous manner. In response to the reviewer's comments, we have enhanced the methodological clarity by adding a new subsection (Section 4.1.4 "Data splitting during training") in the "4.1 Benchmark datasets" section page 18 to thoroughly describe the data partitioning protocol.

4.1.4 Data splitting during training

Consistent with previous works [16, 17], DS1 and DS2 dataset was divided into training/test sets according to stringent sequence diversity thresholds (maximum 30% sequence identity between test and training sequences). A training set was partitioned into five folds. For a same set of hyperparameters, five models were trained on four folds and validated on the remaining fold. Notably here, the validation was only performed on the fold against those other four training folds based only on the training set and had not touched a bit of the aside test set. The final trained models with the optimized hyperparameters were used in the prediction task on the test set. The final predictions on the test set were obtained by averaging the outputs of the five converged models. DS3 dataset was also divided into training/picking/test sets following sequence diversity thresholds. Additionally, none of the ligands in the test set appear in the training set. The picking set of DS3 dataset is only used to determine the cutoff thresholds for binary classification on DS3 test set when some evaluation metrics (such as MCC) need. We emphasize here that the picking set, as well as the training set, is entirely different enough to the test set.

Additionally, we have revised the section 4.1.3 to clarify the dataset processing pipeline and moved Fig S3 to Fig 8 on page 17:

... Fig. 8 illustrates the construction process of the DS3 dataset. Firstly, we removed protein chains with resolutions of $> 3.0 \text{ \AA}$ and lengths of > 1500

2. The Reviewer comments that

Inadequate evaluation. A variety of classification metrics are described, but it is never stated _what_ is being classified. The supplementary note S1 simply says "binding sites" without defining what this means. Looking at the GitHub I believe per-residue classification is being performed. This raises some concerns with the comparison with other methods as they may have different notions of what qualifies as a binding site

residue. Some methods (e.g. DeepPocket) do not even produce per-residue scores as they are predicting the location of a whole pocket. The description of evaluation metrics needs to be expanded. A stronger paper would evaluate LABind using other tool's metrics (e.g. how well pockets are localized).

Response:

Thank you for your thoughtful and rigorous comments. In this study, a binding site is defined as a set of residues that interact with a ligand. To clarify the scope of our research, we now explicitly state that the identification of binding residues is the primary objective of our method when first introducing the task. Accordingly, we have revised the section introduction on page 2 as follows:

In this study, we propose a method capable of predicting protein binding sites for ligands that are not present in the training set. A binding site is defined here as a set of protein residues located within a specific distance from the ligand [21]. Consequently, the classification task in LABind is per-residue prediction, where the goal is to determine whether each residue in a protein is part of a binding site. ...

We also acknowledge your concern regarding the comparability of LABind with existing methods such as P2Rank, DeepPocket, and DeepSurf, which primarily focus on predicting binding site centers rather than identifying individual binding residues. To address this, we have added new section “Baseline methods” on Page 6 to highlight the conceptual differences among these approaches:

... Among the compared methods, P2Rank, DeepSurf, and DeepPocket are structure-based approaches that predict the spatial center of a binding site rather than individual binding residues. The introductions of these methods are provided in Supplementary Note S2.

Furthermore, we fully agree with the reviewer’s comment that “A stronger paper would evaluate LABind using other tools’ metrics.” To gain a more comprehensive comparison, we have conducted a supplementary evaluation focusing on binding pocket localization, following practices adopted by pocket-centric methods. We have also excluded the unfair comparisons made with baseline methods (P2Rank, DeepSurf, DeepPocket) in evaluation on the benchmark datasets. Specifically, we incorporated two widely used metrics (DCC and DCA) into Section 2.2 (page 4) to evaluate binding site center localization:

... In addition, we used DCC (distance between the predicted binding site center and the true binding site center) and DCA (distance between the predicted binding site center and the closest ligand atom) to evaluate the performance in localizing binding site centers. Their calculations are reported in Supplementary Note S1.

Since LABind outputs per-residue predictions rather than explicit pocket centers, we used the mean-shift clustering algorithm to identify pocket centers from the predicted binding residues. The experimental setup is described in Supplementary Note S4. To strengthen the robustness of this evaluation, we performed experiments not only on our existing datasets (DS1, DS2, DS3), but also on COACH420, a benchmark dataset commonly used in previous studies for pocket center prediction. We further curated a subset named

COACH64 by removing redundancy, following the stringent data leakage prevention procedure used in DeepPocket. Specifically, we excluded any proteins with >50% sequence identity to the DS3 training set or with both ligand similarity >0.9 and sequence identity >30%. The details are described in Supplementary Note S5. A new subsection “Performance of LABind in predicting binding site centers” has been added to Section “Results” (Page 8) summarizing the results of these two metrics (DCC&DCA) on DS1, DS2, DS3 and a subset of COACH420:

2.5 Performance of LABind in predicting binding site centers

We further evaluated and compared the performance of LABind and three baseline methods (P2Rank, DeepSurf, and DeepPocket) in predicting binding site centers on the benchmark datasets. Since LABind do not directly output predicted binding site centers, we applied the mean-shift clustering algorithm [34] to the predicted binding residues from LABind to determine the binding site centers. This mean-shift-based identification of binding site centers and evaluation protocol is detailed in Supplementary Note S4. Table 3 summarizes the performance of different methods in localizing binding site centers for small-molecule ligands. For DS1 and DS2, we retained only the data involving small-molecule ligands, as the compared methods were specifically trained for small-molecule binding prediction. LABind outperforms other methods in predicting binding site centers on the DS1 and DS2 test sets; however, these datasets have limited ligand diversity. In contrast, DS3 includes a broader range of ligands. On the DS3 test set, LABind achieves a DCC success rate only 0.8% lower than P2Rank but surpasses all other methods in DCA performance. Because the DS1, DS2, and DS3 datasets are not specifically curated for evaluating binding site center localization, we further selected 64 binding sites from the COACH420 benchmark as an additional test set. Following DeepPocket’s stringent data filtering criteria, we excluded protein-ligand complexes from COACH420 that share more than 50% sequence identity with proteins in the DS3 training set or exhibit both greater than 0.9 ligand similarity and more than 30% sequence identity. The resulting dataset, termed COACH64, is described in detail in Supplementary Note S5. On the COACH64 test set, LABind achieves the best overall performance, with DCC and DCA success rates 4.7% and 3.1% higher, respectively, than those of the second-best method. Although LABind is not specifically designed to predict binding site centers, these results demonstrate its strong potential in this regard.

Table 3 Performance comparison of LABind and similar methods in binding site localization on the four test sets.

Method	DS1		DS2		DS3		COACH64	
	DCC	DCA	DCC	DCA	DCC	DCA	DCC	DCA
P2Rank	0.337	0.574	<u>0.448</u>	0.600	13.8	<u>28.8</u>	46.9	<u>82.8</u>
DeepSurf	0.391	<u>0.628</u>	0.441	<u>0.655</u>	12.6	26.9	<u>48.4</u>	<u>82.8</u>
DeepPocket	<u>0.395</u>	0.600	0.421	0.628	12.3	28.0	39.1	67.2
LABind	0.533	0.735	0.566	0.745	<u>13.0</u>	30.2	53.1	85.9

Note: Boldface and underline indicate the best performance and the second best performance among the compared methods, respectively. The success rates of DCC and DCA are both based on the top- n predicted binding site centers, where n denotes the number of identical cognate ligands.

The supplementary information related to the above discussion is as follows:

Note S1 The calculations of metrics

...To holistically evaluate LABind, we employed DCC (distance between the predicted binding site center and the true binding site center) and DCA (distance between the predicted binding site center and the closest ligand atom) to assess its ability in localizing binding sites. Performance was evaluated by calculating the success rate, defined as the proportion of complexes where DCC (DCA) values fell below the predetermined threshold.

Consistent with previous works [4, 5], the threshold is set to 4 Å. Therefore, the success rate is calculated as follows:

$$\text{Success Rate(DCC/DCA)} = \frac{\text{number of predicted binding sites with DCC/DCA} < 4\text{\AA}}{\text{total number of binding sites}} \times 100\%. \quad (\text{S7})$$

Note S4 Mean-shift-based identification of binding site centers and evaluation protocol

We adopted the mean-shift algorithm [17] as the clustering method because it can automatically determine the number of clusters. However, the mean-shift algorithm still requires the specification of a bandwidth parameter. In this study, the bandwidth was set to 12.0. The cluster centers were defined as the average positions of atoms within each cluster. The binding sites were ranked by cluster size from high to low. For P2Rank, DeepSurf, and DeepPocket, we directly used the predicted binding site centers provided by each method. We selected the top- n binding sites to calculate the evaluation metrics, where n represents the number of identical ligands in the corresponding native protein-ligand complex. For example, in identifying the binding site center for a protein with ligand A , we set n to the number of protein-ligand complexes containing ligand A , rather than the total number of all ligands.

Note S5 Construction and evaluation settings of the COACH64 dataset

COACH420 was derived from the COACH test set, and consisted of 420 single chain structures containing a mix of drug targets and naturally occurring ligands. COACH420 is a benchmark dataset commonly used by P2Rank, DeepSurf, and DeepPocket. To avoid data leakage arising from discrepancies among training sets, we applied a redundancy filtering procedure. Following DeepPocket, we removed complexes from the COACH420 that share more than 50% sequence identity with proteins in the DS3 training set, or have both greater than 0.9 ligand similarity and more than 30% sequence identity. Here, ligand similarity was calculated using the Tanimoto coefficient between molecular fingerprints of the two ligands. After filtering, 59 protein chains were retained, containing a total of 64 binding sites. To distinguish it from COACH420, we named the dataset COACH64. To compare performance across methods, we used the pre-trained models of P2Rank, DeepSurf, and DeepPocket for binding site center prediction on COACH64. We used LABind trained on DS3 to predict binding sites on COACH64.

These revisions enhance the rigor of the paper and address the reviewer's constructive feedback. To facilitate further research, we have updated the GitHub repository by adding functionality to output binding site centers using the mean-shift algorithm, and the corresponding dataset has been uploaded to Zenodo.

3. The Reviewer comments that

Nearly exclusive use of co-crystals. Most of the evaluations are done using holo structures that are perfect images for their cognate ligands. A stronger paper would evaluate DS3 using apo and/or predicted structures (it is not an interesting problem to identify a binding site from a holo structure). Fig 3 is encouraging in this regard, but is only applicable to a small number of ligands. More data is needed to support the conclusion of 2.4.

Response:

Thank you very much for your valuable comment. We agree with the reviewer's point that holo structures tend to facilitate binding site identification, accurate prediction for apo

and/or predicted structures holds greater methodological value. Therefore, we have supplemented the evaluation with results on the DS3 test set using predicted structures generated by ESMFold, employing model weights pre-trained exclusively on holo structures. We have revised Figure 2b to include the performance metrics on predicted structures and revised the section 2.6 (previously section 2.4) on page 9-10:

2.6 Validation of LABind on binding sites of proteins without experimentally determined structures

Previous results are based on holo structures (ligand-bound protein structures). However, physicochemical interactions between binding sites and ligands can induce conformational changes in proteins, leading to transitions between apo structures (unbound protein structures) and holo structures [35]. Accurately predicting binding sites on apo or predicted structures poses a greater challenge and holds higher practical significance compared to holo structures. Furthermore, due to the limitations of some experimental techniques, there are currently many proteins without a clear tertiary structure. ...

... To further validate the robustness of LABind on ESMFold-predicted structures, we used the predicted structures from ESMFold as structural input to predict binding sites on the DS3 test set. As illustrated in Fig. 2b, the performance of LABind using ESMFold-predicted structures still outperforms LigBind evaluated on holo structures by 0.082 in MCC and 0.094 in AUPR. This demonstrates the robustness of LABind when using predicted structural inputs and its potential generalizability to sequence-based binding site prediction. ...

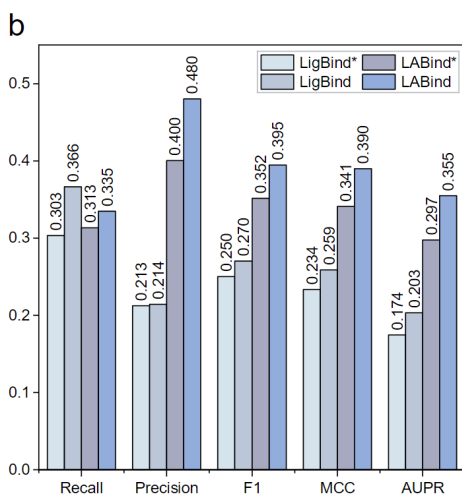


Fig. 2 ... b Overall performance metrics of LABind and LigBind on the DS3 dataset. '*' represents the use of ESMFold-predicted structures as the structural input. The results of LigBind are obtained from LigBind-G program.

4. The Reviewer comments that

Page 3. "Given the relatively fixed structure of small molecules..." I don't know what is trying to be said here - small molecules have greater diversity of chemical structures than macromolecules formed from a limited number of building blocks.

Response:

We sincerely apologize for any confusion resulting from our unclear or incorrect

description. We aim to highlight that there is a difference between protein-macromolecule interactions and protein-ligand (small molecule) binding. The use of the term “relatively fixed structure” was problematic, and we have revised this description on page 3:

... Additionally, as a partner of protein binding, ligand can be either biomacromolecules (such like proteins, RNAs) or other small molecules (including ions). We assume that protein-biomacromolecules binding may exhibit many different properties comparing with protein-small-molecule binding [22]. In this study, we focused on predicting protein binding sites for small molecules and ions.

5. The Reviewer comments that

Page 7. You use "picking set" without defining what it is.

Response:

Thank you for your careful review. We have added a brief explanation and a detailed introduction link to section 4.1.3 to provide access to the information about the picking set.

The DS3 picking set (see “4.1.3” for details) was used solely to tune hyperparameters and determine the binary classification threshold based on the model’s outputs, ensuring that no information from the test set influenced the training process.

The definition of the DS3 picking set is as follows:

... Data released after this date that contained ligands present in the training set were used as the picking set, whereas those containing ligands not observed in the training set were used as the test set. ...

6. The Reviewer comments that

2.4. "we filtered out predicted structures with a TM-score <0.5" Are you saying that 3b is only showing proteins where the TM-score is >0.5? Why is 3d only showing TM-score < 0.5 proteins? I don't see how the figure supports the claim (AUPR distribution of ESMFold is superior to that of OmegaFold) - 3a shows most examples have TM-score > 0.5 so wouldn't those dominate the AUPR distribution (which is not shown)?

Response:

Thank you for pointing out this ambiguous section. Figure 3 analyses how the performance of LABind is relied on the high-quality protein structures, and identifies which protein structure predictor is good to LABind. In response to your comment, we now re-assembly Figure 3 as following description.

To clarify, Figure 3a and 3b include all predicted structures on DS1 test set. Structures with low TM-scores were not filtered out. We retrained and tested LABind using the predicted structures from ESMFold and OmegaFold on the DS1 dataset, respectively. Figure 3a presents a violin plot of the TM-scores for each predicted structure in the DS1

test set, providing an overview of the overall accuracy of the predictions. Figure 3b presents the AUPR metrics obtained from evaluations using various predicted structures on the DS1 dataset. Figure 3c provides a further analysis focusing on low-quality predicted structures. We selected structures with TM-score < 0.5 to analyze the impact of low-quality structures on binding site identification. As our original assessment, which relied solely on the visual comparison of the linear regression lines, lacked sufficient rigor, we have modified the corresponding description on page 9:

Fig. 3 analyzes how the performance of LABind is relied on the high-quality protein structures, and identifies which protein structure predictor is good to LABind. The distributions of TM-scores for ESMFold and OmegaFold on the DS1 test set is presented in Fig. 3a. The average TM-score of ESMFold is higher than that of OmegaFold on the DS1 test set (0.861 vs 0.824). As shown in Fig. 3b, when LABind was retrained on the DS1 dataset using predicted structures from ESMFold and OmegaFold, respectively, it is evident that LABind based on native structures outperforms those based on predicted structures. This is understandable, as native structures provide more accurate surface and spatial information. Specifically, the average AUPR achieved using ESMFold and OmegaFold is lower than that of the native structure by 0.055 and 0.066, respectively. Although the performance using OmegaFold-predicted structures is suboptimal, the average AUPR remains 0.065 higher than that of LigBind. Additionally, as shown in Fig. 3c, we select predicted structures with a TM-score < 0.5 on the DS1 test set to examine the robustness of the LABind for low-quality predictions. On the DS1 test set, OmegaFold contains 252 proteins with a TM-score < 0.5 , while ESMFold contains 152 proteins with a TM-score < 0.5 . As visualized in the Fig. 3c, due to the higher number of low-quality proteins in OmegaFold, the linear regression line for LABind based on ESMFold-predicted structures consistently lies above that based on OmegaFold-predicted structures.

7. The Reviewer comments that

"For proteins with accurate native structures" - so they have inaccurate native structures? Do you mean without experimentally determined structures?

Response:

Thank you for pointing out the error. We meant proteins without experimentally determined structures. We have revised this description to clarify this point on page 9:

For proteins without experimentally determined structures from wet-lab, we chose predicted structures from dry-lab approaches like ESMFold, instead.

8. The Reviewer comments that

2.6. "we disregarded complexes identified by LABind as having no binding sites" All complexes in CASF-2016 have binding sites, so these examples are failures of LABind and should be acknowledged and quantified as such. Need to specify that you are doing whole protein docking with a box created around the whole protein. Unclear if multiple binding sites are docked to or only one. Would be interesting to know how "Vina with binding sites" compares to docking using the cognate ligand (e.g., autobox_ligand in smina/gnina).

Response:

Thank you for your insightful and constructive comment, which helped us identify and address the lack of rigor in our initial experimental validation. In response, we have re-conducted the docking experiments with a more comprehensive and transparent evaluation protocol.

The CASF-2016 dataset contains 285 protein-ligand complexes, each consisting of a one-to-one mapping between a protein and its cognate ligand; thus, there are no cases with multiple ligands per complex. We applied LABind to predict binding sites for all complexes and found that 16 of them had no predicted binding site on any chain. These are considered failure cases of LABind and are now clearly acknowledged and quantified in our analysis. As a result, 269 complexes were retained for docking experiments.

To objectively assess the impact of predicted binding sites, we reported to use smina as the docking engine in three different settings:

- (1) Blind docking: A docking box was created to encompass the entire protein structure, centered on the geometric center of the protein.
 - (2) Binding site-guided docking: Docking boxes were generated based on the spatial center and extent of LABind-predicted binding residues for the specific ligand in each complex.
 - (3) Cognate ligand-guided docking (autobox_ligand): Docking boxes were automatically defined by the position of the native ligand using the --autobox_ligand option in smina.
- This setup allows us not only to examine how helpful the binding sites predicted by LABind for smina blind docking, but also to explore how far LABind will have chance to improve.

We have revised Section 2.8 (Page 12) to provide detailed descriptions of these three experimental setups, including how the search boxes were defined, and clarified that docking was performed to one pocket per complex, since each complex contains only one ligand.

To further investigate the potential application of binding sites, we used LABind predicted binding sites to assist in molecular docking tasks, to both examine how helpful these predictions are for smina blind docking and explore the potential for further improvement of LABind. We selected CASF-2016 dataset [38] to evaluate performance on molecular docking tasks and used smina [39] as the docking program. Subsequently, LABind failed to identify binding sites on 16 complexes, resulting in 269 protein-ligand complexes for the further analysis. Since all complexes in the CASF-2016 dataset consist of a single protein and a single ligand, we utilized the average position and spatial extent of the binding residues predicted by LABind to determine the center and radius of the search box for the specific ligand in each complex (binding site-guided docking). For comparison, we employed both blind docking and cognate ligand-guided docking. Blind docking was performed using the geometric center of the protein and a box that encompasses the whole structure. Cognate ligand-guided docking was performed based on the position of the native ligand using the --autobox_ligand option in smina. It is noted that this option means using almost known binding site. Supplementary Note S6 provides a detailed exposition of three experimental details and procedures. We assessed the accuracy of predicted structures, with a strict RMSD threshold of 2 Å considered a successful prediction. The success rate of the top 20 structures was used to compare the effectiveness of predictions without and with the binding sites predicted by LABind. As shown in Fig. 5, docking using LABind-predicted binding sites improves the average top-20 success rate by 19.1% over blind docking. However, it still underperforms by 21.9% compared to docking with the cognate

ligand, indicating room for further improvement. These results demonstrate the potential of LABind in enhancing blind docking scenarios and highlight directions for future enhancement.

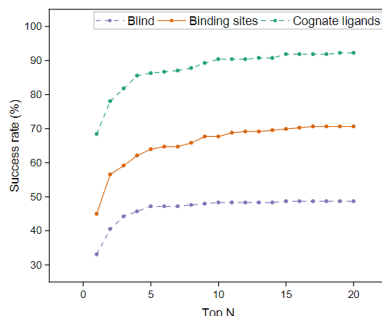


Fig. 5 The top N success rate under three different experimental conditions. ‘Blind’, ‘Binding sites’, and ‘Cognate ligands’ represent the success rates of blind docking, binding site-guided docking, and cognate ligand-guided docking, respectively.

And we have revised the Supplementary Note S6 on page 10:

Note S6 The experimental details and procedures of molecular docking

For the 285 protein-ligand complexes in the CASF-2016 core set [18], we first performed a preprocessing step. Initially, all protein chains were separated. LABind predictions were then performed on individual chains. Complexes where no binding site was predicted for any chain were excluded. Ultimately, 16 proteins were excluded, retaining 269 protein-ligand complexes for analysis. To ensure that smina [19] can correctly perform molecular docking, we use the processing files it provides to process the protein and ligand. The command lines executed are as follows:

```
prepare_receptor4.py -i receptor.pdb -o receptor.pdbqt -A hydrogens -U nphs
prepare_ligand4.py -i ligand.mol2 -o ligand.pdbqt -A hydrogens -U nphs
```

When not utilizing predicted binding sites, docking was performed with the grid box centered on the geometric centroid of the protein and dimensions fully encompassing the entire structure. The binding sites predicted by LABind are used to reduce the size of the search box and determine the center position of the search box. The center of the search box is set at the geometric center of all the atoms of the binding residues, with the size of the box determined by the longest distance between atoms in each dimension plus 4 Å. Docking using the cognate ligand was implemented via *smina autobox ligand*. To evaluate prediction outcomes, we computed the RMSD between the native pose and docked poses using the Hungarian algorithm (implemented via the *spyrmsd* package [20]).

9. The Reviewer comments that

2.8 "protein compound embeddings" This was confusing since "compound" can be synonymous with ligand. Suggest changing to protein-DSSP embedding

Response:

Thank you for your helpful suggestion. We have revised all instances of “protein compound embedding” to “protein-DSSP embedding” in both the manuscript and the supplementary information.

10. The Reviewer comments that

Figure 7b. This figure is inadequately labeled. Need a legend where it is clear what the different colors/styles are.

Response:

Thank you for pointing out the error. We have added a description of the colors in the legend of Figure 7b on page 14:

b Performance improvement of LABind over LABind without ligand information on the six ligands test sets in the DS1 dataset. AVG represents average values across all ligands. MCC and AUPR are represented by blue and yellow, respectively. Light colors represent results obtained without using ligand information, while textured dark colors indicate performance improvements achieved with ligand information. The values indicate the improvement achieved by incorporating ligand information.

11. The Reviewer comments that

Page 16. "Sequence identity was reduced to 30%". I don't know what is being said here. Identity with what? How reduced. "Proteins with protein sequence identity below 30%..." identity with what? (presumably test set, but say it)

Response:

Thank you for pointing out this ambiguity. We have revised the description on page 17 to provide a more accurate and explicit explanation:

To ensure low sequence similarity between the training and test sets, redundant proteins sharing more than 30% sequence identity with any protein in the test set were removed using CD-HIT [43].

12. The Reviewer comments that

Page 18. "Gaussian radial basis function to compute the distances". RBFs don't compute distances, they encode them.

Response:

Thank you for carefully pointing out this inaccuracy. We have revised the sentence on page 20 to more accurately describe the role of the Gaussian radial basis function:

We use a Gaussian radial basis function $\text{RBF}(\cdot)$ to encode the pairwise atomic distances D , and then concatenate the resulting distance embeddings accordingly.

13. The Reviewer comments that

Page 19. W_O is defined but not used in any equation.

Response:

Thank you for bringing this issue to our attention. We apologize for the oversight. We have carefully reviewed all equations and revised Equation (3) on page 21 to correctly incorporate the term W_o as originally intended:

$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l)} + \mathbf{W}_O^{(l)} \left(\sum_{j \in N(i) \cup i} \alpha_{ij}^{(l)} \mathbf{W}_V^{(l)} (\mathbf{e}_{ij}^{(l)} \parallel \mathbf{h}_j^{(l)}) \right),$$

In addition, we have checked the entire manuscript to ensure consistency and accuracy across all formulas, including coherence with surrounding text and formatting. All modifications have been highlighted in the revised manuscript.

Response to Reviewer #3

We sincerely thank the reviewer for the valuable comments and helpful suggestions.

1. The Reviewer comments that

The authors were highly responsive to reviewer feedback and the result is a much stronger paper worthy of publication. Copy editing for English language usage would still be helpful.

Response:

Thank you for your positive feedback and valuable comments. We have reviewed the English language in the manuscript and corrected numerous grammatical errors, awkward phrasing, and typos. The revisions have been highlighted in the manuscript. We hope this revision makes the article more accessible and easier for readers to understand.