14 January 2025 · CC-BY 4.0

Peer Review

Review of: "Evaluation of Molecular Docking by Deep Learning and Random Forests: A Hybrid Approach Based on Pseudo-Convolutions"

Juyong Yoon[1]

[1]. KIST Europe, Saarbrücken, Germany

Manuscript Review: Evaluation of Molecular Docking by Deep Learning and Random Forests: A Hybrid Approach Based on Pseudo-Convolutions

The manuscript explores a novel hybrid approach using pseudo-convolutions and Random Forests for protein-protein interaction predictions, aiming to address computational bottlenecks in molecular docking. While the methodology is innovative and presents valuable potential in reducing computational costs, several critical issues must be addressed to improve the scientific rigor and applicability of this study.

The authors provide a clear and explicit description of the workflow, including data preparation, pseudo-convolution feature extraction, and classifier design. This clarity enhances the reproducibility of the study. By avoiding 3D modeling and leveraging sequence-based features, the study introduces an interesting avenue for simplifying docking predictions. The manuscript demonstrates an innovative approach to protein-protein docking predictions using machine learning. However, significant refinements are required to address concerns regarding data relevance, biological plausibility, and structural considerations. Addressing these issues will substantially strengthen the study's contribution to computational docking and drug discovery. To improve the manuscript's suitability for publication in the journal of Organoid, certain enhancements are necessary, as outlined below:

Major concerns:

doi.org/10.32388/CCJQVB

1) The title of the article is misleading, as the authors focused exclusively on prediction methods for protein-protein interactions. The study did not provide a substantial evaluation of any existing deep learning-based or other methods. Additionally, the term 'molecular docking' refers to the docking of a wide variety of substances, not just proteins. Therefore, the authors must specify the particular type of docking emphasized in this study.

2) The overall methodology of this paper is well-defined and explicitly described, which I found good as a reviewer. However, the dataset (RNAseq) subjected to this study is very questionable, taking into account the fundamentals of any interactions that occur in a biological system. In protein-protein interaction (PPI) and docking studies, some amino acid pairs are less likely to interact effectively due to steric hindrance (Isoleucine-Proline), electrostatic repulsion (Aspartic acid - Glutamate, Lysine - Arginine), or unfavorable physicochemical properties. These pairs typically involve residues with similar charge or size that would not favorably align in the interaction interface. The classifier doesn't take into account these physiological criteria while training the model. A simple convolution filter applied to overlapping nucleic acids doesn’t represent the amino acid nature.

Also, it requires three consecutive ribonucleic acids to translate into one amino acid (e.g.: Arg [AGA or AGG]). The pseudo convolution filter also doesn’t take into account these properties; also, the authors don’t explain whether the model considers the biological translation properties of amino acids. Therefore, this proposed method lacks the very fundamental potential as well as explainability.

3) The dataset combines Affinity Benchmark 3 (docking pairs) and Negatome 2 (non-docking pairs). However, the authors should verify and discuss the compatibility of these datasets, as they originate from different experimental contexts. Furthermore, cross-validation using an independent dataset is recommended to validate the generalizability of the approach.

4) This workflow is also decisive as the study doesn’t consider the 3D structural properties of a protein to train the classifier.

Minor concerns:

1) Explain whether the model considers codon usage or amino acid translation when generating features. This would clarify the relevance of RNA sequence data in docking predictions.

2) While overall sensitivity is high, overall specificity is comparatively lower. The authors could explore techniques like cost-sensitive learning or rebalancing to enhance performance on non-

docking cases.

3) Ensure the citation style aligns with standard practices, and include recent works in molecular docking and machine learning to contextualize the study better.

Declarations

Potential competing interests: No potential competing interests to declare.