Machine Learning for drug-target interaction estimation in Drug Discovery: from replication to data integration analysis

Piervito Creanza

Vittorio Pio Remigio Cozzoli

Introduction

The *“Machine Learning for drug-target interaction estimation in Drug Discovery: from replication to data integration analysis”* is a pivotal initiative aimed at enhancing the adaptability of the GenScore neural network to novel data sets. This project is a testament to the innovative strides being made in the realm of bioinformatics and computational medical chemistry.

It emerged from the necessity to gain a profound comprehension of neural network models utilized in bioinformatics and computational medical chemistry.

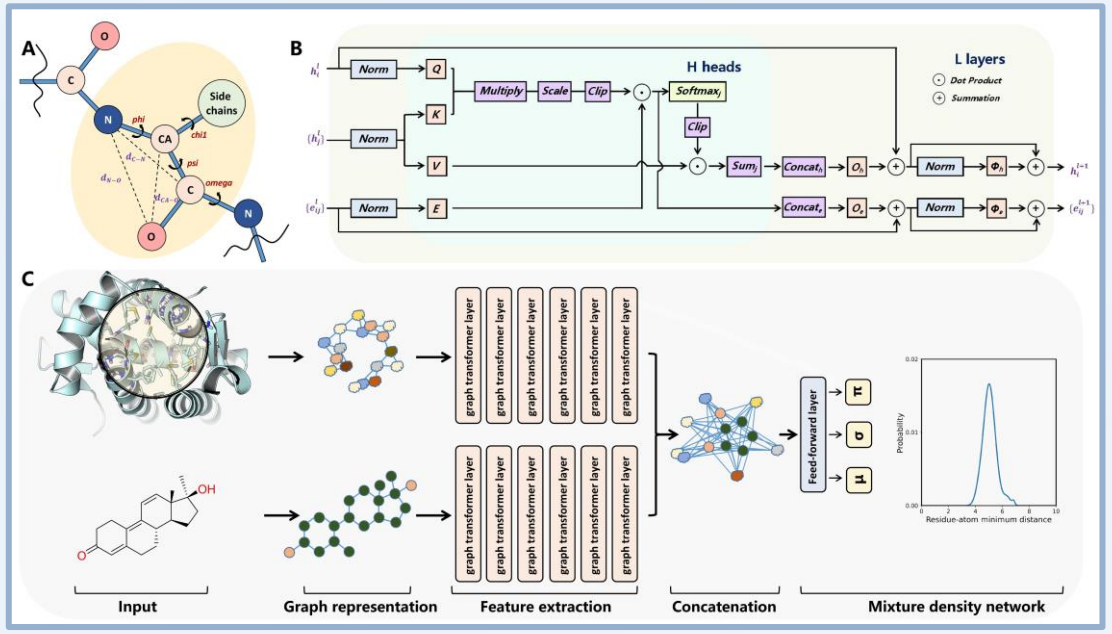
The goal is to refine these models to effectively handle new and diverse data sets.

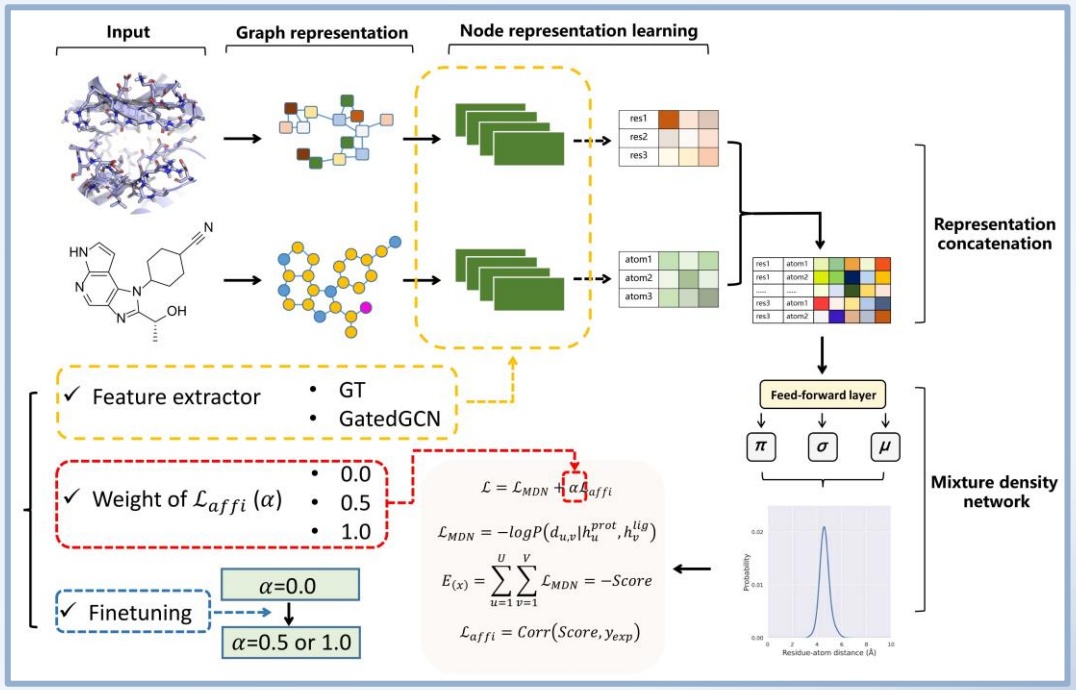
Related work

The journey from RTMScore to GenScore represents a significant evolution in the development of scoring functions for protein-ligand interactions. RTMScore, developed using machine learning and a unique residue-based graph representation strategy, set a new benchmark in the field. However, the challenge of achieving robust performance and wide applicability remained due to the task-specific nature of most scoring functions.

Addressing this challenge, the same research group introduced GenScore, an evolution of RTMScore. GenScore incorporates a new parameterization strategy that introduces an adjustable binding affinity term into the training of a mixture density network. This innovative approach not only improves scoring and ranking performance but also maintains superior docking and screening power.

The study underscores the potential utility of this innovative parameterization strategy and the resulting scoring framework in future structure-based drug design. It demonstrates that a model's performance can be balanced through an appropriate approach, paving the way for advancements in bio-research and drug design.





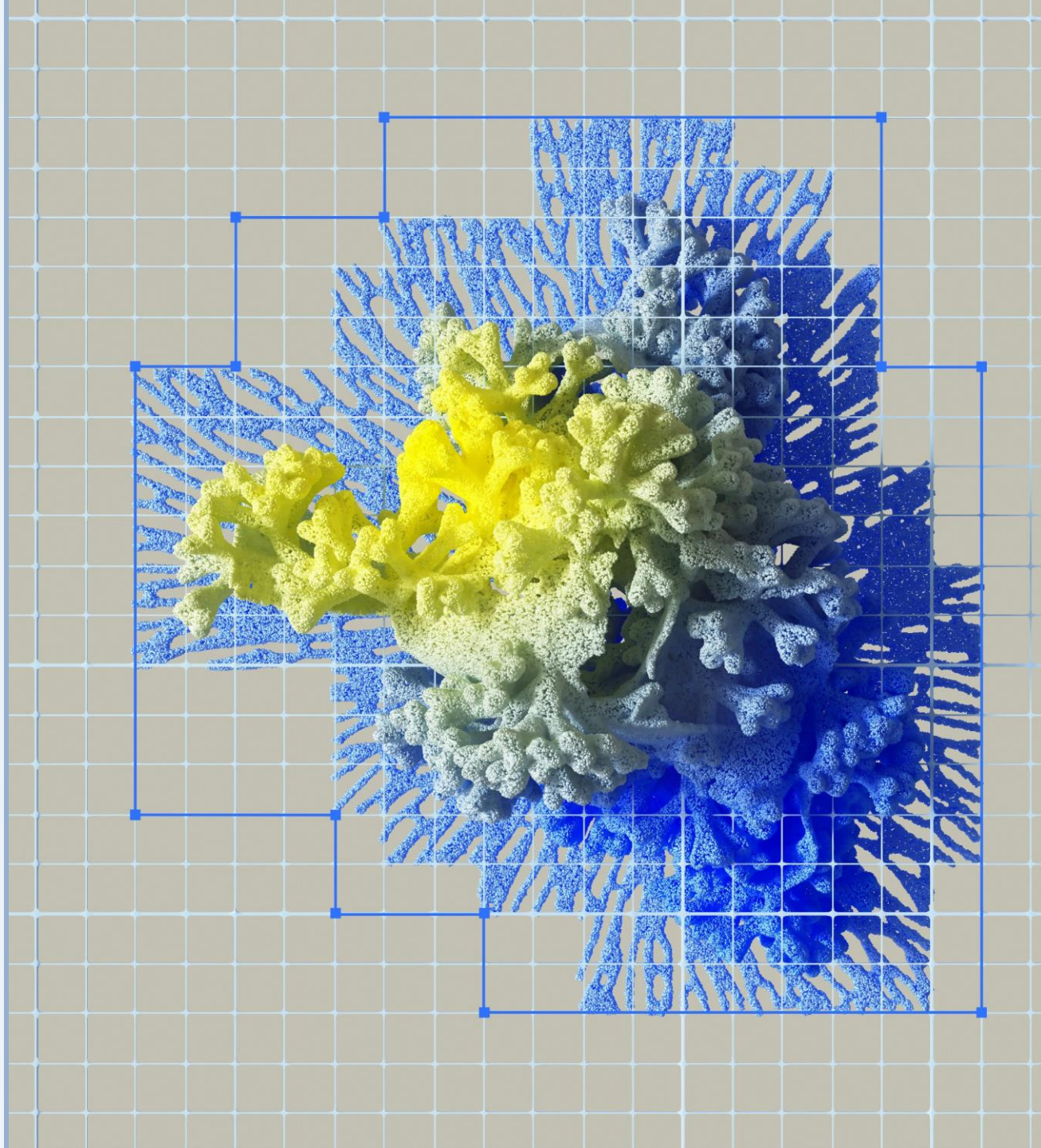
Innovation of our project

This project represents a significant evolution in the application of neural networks to manage protein and ligand interaction data.

The GenScore model, originally designed to work with specific protein pocket cutouts, is being retrained to handle a different type of input data.

The innovative approach of this project lies in its investigation of how the model performs when presented with a completely different dataset than it was originally designed for. The new data features a slimmer protein cutout, making the input data more manageable and less computationally burdensome.

The ultimate goal of this project is to enhance GenScore's versatility and applicability by enabling it to work with both types of protein cutouts. This could potentially increase its usefulness in various scientific contexts. The project underscores the importance of adaptability in machine learning models and their potential for innovation in bio-research.



Project development

It was important to start with an introductory course on neural networks and the PyTorch library, ensuring a common understanding of the concepts and tools necessary for the project. We then analyzed the GenScore source code, focusing on the model training and the ETL process of the data.

Identifying the Python requirements of the project was a challenge, due to obsolete or invalid packages in the provided list. Because of this, we had to manually cross-reference the dependencies to find substitute packages compatible with their system.

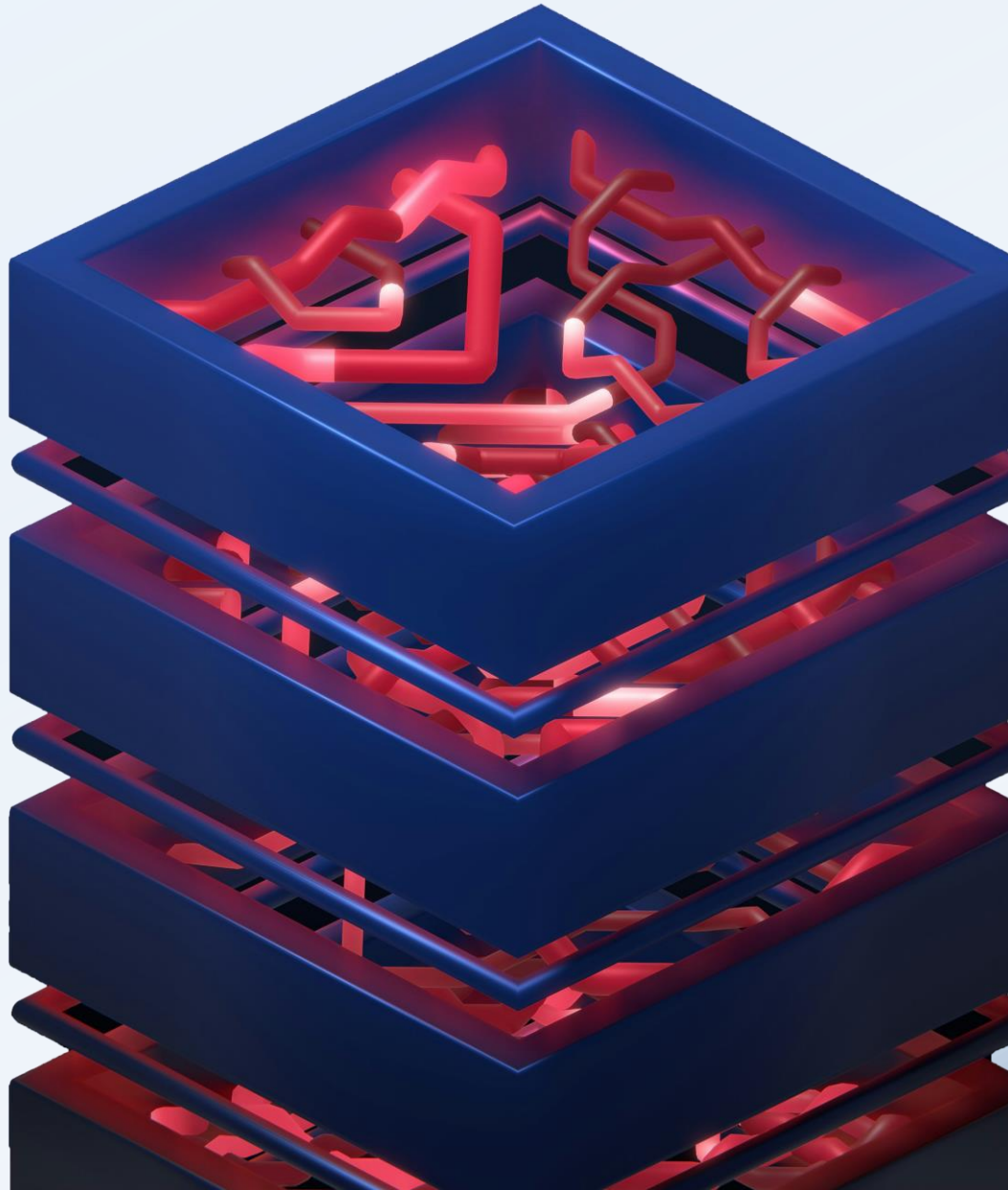
A virtual machine for neural network training and a set up for remote code execution was configured. We reverse-engineered the *mol2graph_rdmda_res.py* script to make it compatible with our new dataset, resolving bugs and introducing new parameters and functions.

After this stage, we conducted two training phases:

- the first one, using the new smaller cutout protein dataset, where we encountered a problem with non-conformable matrix multiplication due to a bug in the original graph conversion script, which we fixed.
- the second one on about 2K protein-ligand pairs from the PDBbind database using firstly *energy* and secondly *-logKd/Ki* labels.

Finally, we evaluated the model using the CASF-2016 Benchmark, focusing on the metrics of *Forward Screening Power* and *Power Scoring*.

Despite the absence of documentation and some bugs in the original code, we were able to successfully retrain the neural network and make significant improvements to the code.



Used Tools

The realization of the project required the use of a multitude of tools. In particular:

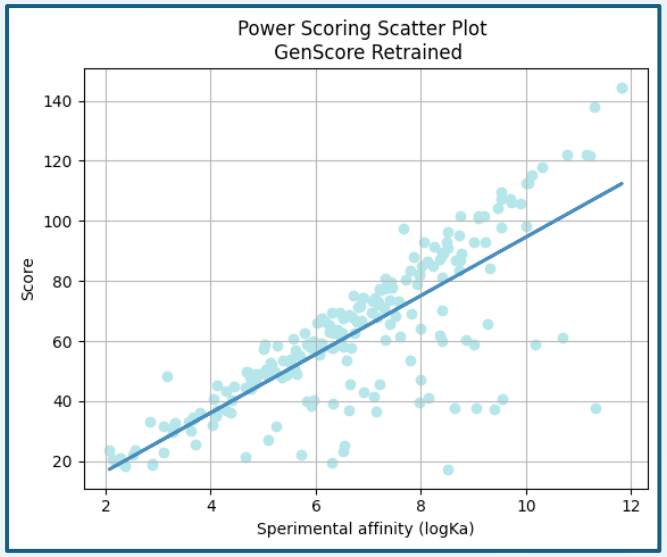
- **PyCharm:** For writing the code and remote execution, we chose to use the PyCharm IDE, as extensively described in the previous section.
- **Terminal:** To interact with the VM and execute remote commands, SSH was used along with the respective terminal apps, including for example iTerm2 and Termius.
- **Termius:** For a graphical management of the files present on the VM through SFTP, making their management easier. In addition to this, Termius provides autocomplete functions for terminal sessions.
- **GitHub:** To manage the versioning functions of the project and archive its source code *(repo link)*.
- **Google Colab:** In order to try ad hoc scripts in an extemporaneous manner and not connected to the project files, such as those for generating scatter plots for the validation of the retrained model.

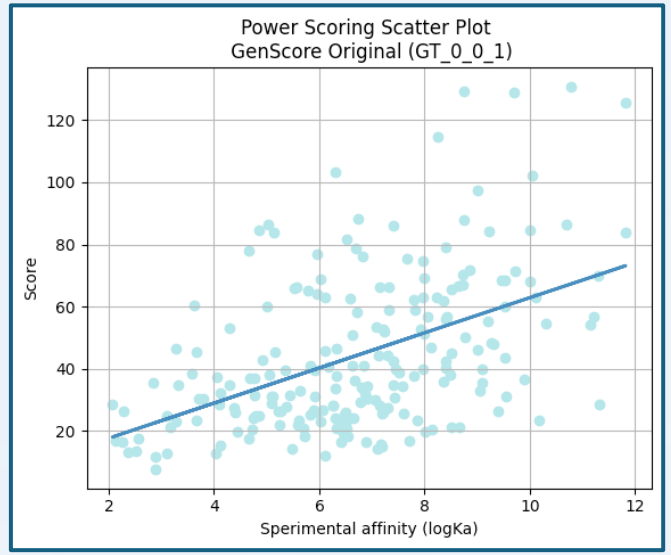
Results: first training

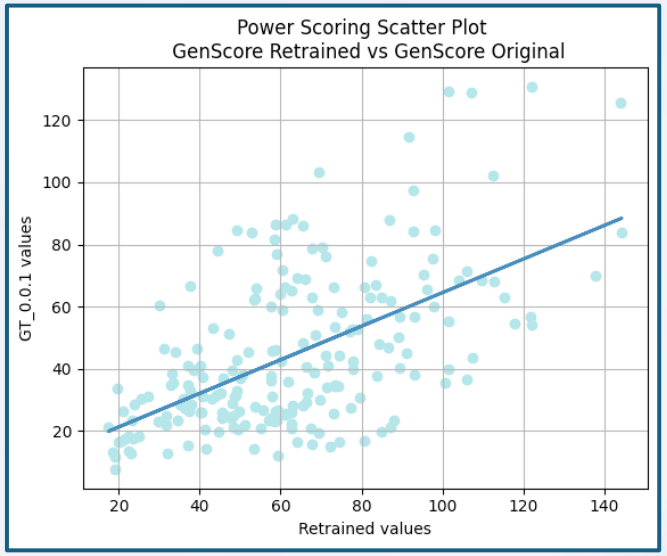
The first retraining showed very good results.

As can be seen in the following scatterplots, there is a very strong positive correlation between the model predictions in the power scoring test and the experimental results.

Furthermore, the forward screening power test shows results on par with the original model trained in the GenScore research paper.





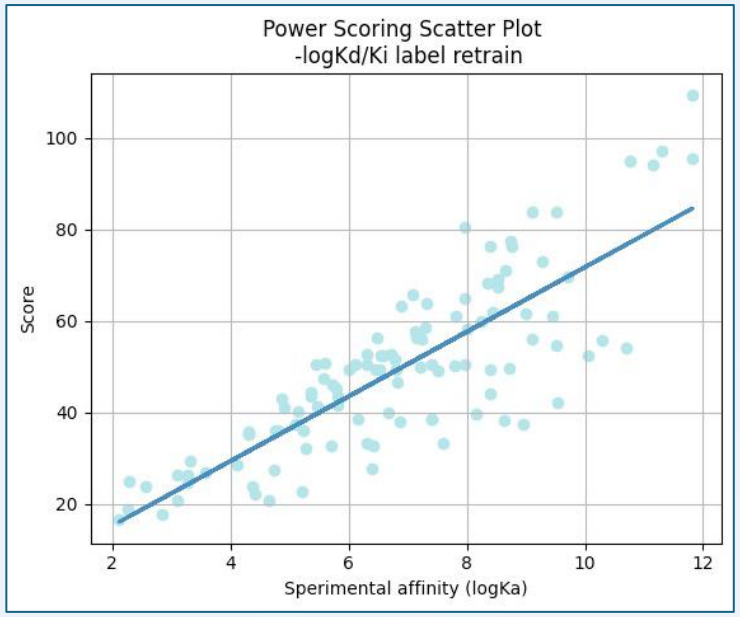


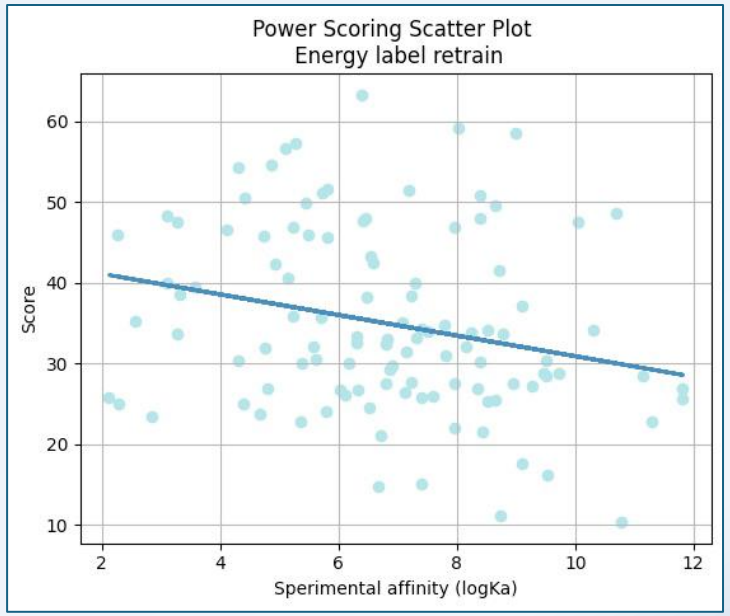
Results: second training

The second retrain did not show the same good results of the first one. The correlation between the model predicted scores and the experimental results is a weak, negative one. This was not completely unexpected, in fact there are multiple reasons that could have led to so poor results:

- **Size of the training set:** since an energetical label wasn't available for each of the PDBind complexes, only a set of about 2K protein-ligand pairs was used. This is a relatively small number, if compared either to the original amount of protein-ligand pairs used to train the original model or the number of parameters of the neural network itself.
- **Quality of the energy labels:** the used energy labels had already shown a fairly low positive correlation for massive quantity of data in other experiments, thus negatively affecting our results.
- **Model inadequacy:** since the model was conceived for *-logKd/Ki* labels, using a different set could have compromised its results.

It would be interesting to perform a new training when a larger set of *energy labels* will be available.





Conclusions

This project wants to propose two new approaches to the training of state-of-the-art convolutional neural networks conceived to calculate the degree of affinity between various protein-ligand complexes: on one hand, using datasets of cutout protein pockets, while, on the other hand, using energy labels rather than classical ones.

As can be seen in the ***RESULTS*** slide, the first of these new approaches led us to achieve very good scoring results, even though the GenScore model was not developed to work with this type of training set.

However, our second experiment didn't return the kind of results we hoped for.

It would be interesting to see if these same results and approaches, for better or worse, could be replicated with other similar models, such as RTMScore, the predecessor of GenScore.

This could be the goal of further research works.

Second Training's FSP (energy)	
Analysed metric	Value
Average enrichment factor among top 1%	1.01
Average enrichment factor among top 5%	1.21
Average enrichment factor among top 10%	1.19
The best ligand is found among top 1% candidates for **2** cluster(s) with a success rate of	3.6%
The best ligand is found among top 5% candidates for **6** cluster(s) with a success rate of	10.7%
The best ligand is found among top 10% candidates for **13** cluster(s) with a succes rate of	23.2%

Second Training's FSP (-logKd/Ki)	
Analysed metric	Value
Average enrichment factor among top 1%	1.43
Average enrichment factor among top 5%	1.27
Average enrichment factor among top 10%	1.16
The best ligand is found among top 1% candidates for **3** cluster(s) with a success rate of	5.4%
The best ligand is found among top 5% candidates for **7** cluster(s) with a success rate of	12.5%
The best ligand is found among top 10% candidates for **14** cluster(s) with a succes rate of	25.0%

First Training's FSP (-logKd/Ki)	
Analysed metric	Value
Average enrichment factor among top 1%	22.44
Average enrichment factor among top 5%	7.64
Average enrichment factor among top 10%	4.33
The best ligand is found among top 1% candidates for **34** cluster(s) with a success rate of	60.7%
The best ligand is found among top 5% candidates for **39** cluster(s) with a success rate of	69.6%
The best ligand is found among top 10% candidates for **41** cluster(s) with a succes rate of	73.2%

Acknowledgments

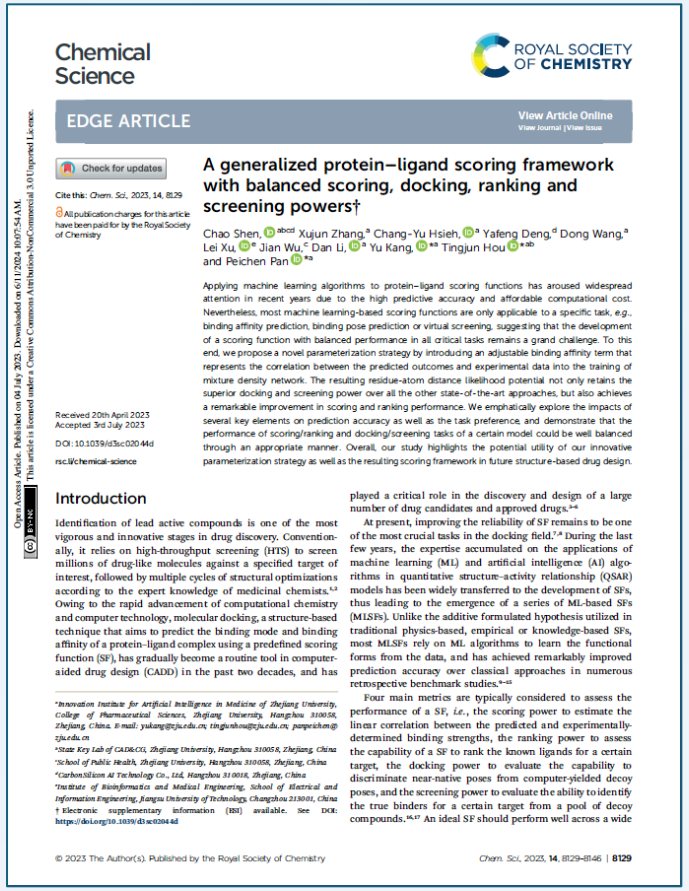
We thank Professor Gianluca Palermo, Davide Gadioli and Gianmarco Accordi for their support and dedication in helping us throughout the whole project.

We want to dedicate a special thank also to Politecnico di Milano, which allowed us to take part in this project and gave us the opportunity to see the world of academic research much more closely.

References

- Chao Shen, Xujun Zhang, Yafeng Deng, Junbo Gao, Dong Wang, Lei Xu, Peichen Pan*, Tingjun Hou*, and Yu Kang*, **“Boosting Protein–Ligand Binding Pose Prediction and Virtual Screening Based on Residue–Atom Distance Likelihood Potential and Graph Transformer”**, *J. Med. Chem. 2022, 65, 15, 10691–10706*, doi: 10.1021/acs.jmedchem.2c00991

- Chao Shen, Xujun Zhang, Chang-Yu Hsieh, Yafeng Deng, Dong Wang, Lei Xu, Jian Wu, Dan Li, Yu Kang, Tingjun Hou, and Peichen Pan, **“A generalized protein–ligand scoring framework with balanced scoring, docking, ranking and screening powers”**, *Chem Sci. 2023 Aug 2; 14(30): 8129–8146*, doi: 10.1039/d3sc02044d



Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Chem. Sci.*, 2023, 14, 8129

All publication charges for this article have been paid for by the Royal Society of Chemistry

A generalized protein–ligand scoring framework with balanced scoring, docking, ranking and screening powers†

Chao Shen, [abcd] Xujun Zhang,[a] Chang-Yu Hsieh,[a] Yafeng Deng,[d] Dong Wang,[a] Lei Xu,[e] Jian Wu,[c] Dan Li,[a] Yu Kang,[a*] Tingjun Hou[*ab] and Peichen Pan[*a]

Applying machine learning algorithms to protein–ligand scoring functions has aroused widespread attention in recent years due to the high predictive accuracy and affordable computational cost. Nevertheless, most machine learning-based scoring functions are only applicable to a specific task, *e.g.*, binding affinity prediction, binding pose prediction or virtual screening, suggesting that the development of a scoring function with balanced performance in all critical tasks remains a grand challenge. To this end, we propose a novel parameterization strategy by introducing an adjustable binding affinity term that represents the correlation between the predicted outcomes and experimental data into the training of mixture density network. The resulting residue–atom distance likelihood potential not only retains the superior docking and screening power over all the other state-of-the-art approaches, but also achieves a remarkable improvement in scoring and ranking performance. We emphatically explore the impacts of several key elements on prediction accuracy as well as the task preference, and demonstrate that the performance of scoring/ranking and docking/screening tasks of a certain model could be well balanced through an appropriate manner. Overall, our study highlights the potential utility of our innovative parameterization strategy as well as the resulting scoring framework in future structure-based drug design.

Received 20th April 2023
Accepted 3rd July 2023

DOI: 10.1039/d3sc02044d

rsc.li/chemical-science

Introduction

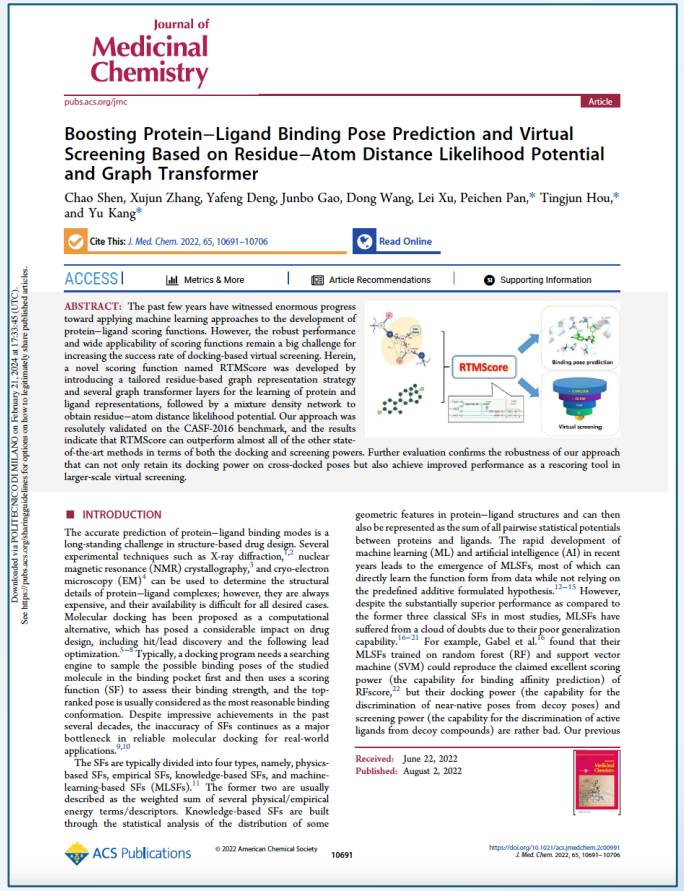
Identification of lead active compounds is one of the most vigorous and innovative stages in drug discovery. Conventionally, it relies on high-throughput screening (HTS) to screen millions of drug-like molecules against a specified target of interest, followed by multiple cycles of structural optimizations according to the expert knowledge of medicinal chemists.[1,2] Owing to the rapid advancement of computational chemistry and computer technology, molecular docking, a structure-based technique that aims to predict the binding mode and binding affinity of a protein–ligand complex using a predefined scoring function (SF), has gradually become a routine tool in computer-aided drug design (CADD) in the past two decades, and has played a critical role in the discovery and design of a large number of drug candidates and approved drugs.[3–6]

At present, improving the reliability of SF remains to be one of the most crucial tasks in the docking field.[7,8] During the last few years, the expertise accumulated on the applications of machine learning (ML) and artificial intelligence (AI) algorithms in quantitative structure–activity relationship (QSAR) models has been widely transferred to the development of SFs, thus leading to the emergence of a series of ML-based SFs (MLSFs). Unlike the additive formulated hypothesis utilized in traditional physics-based, empirical or knowledge-based SFs, most MLSFs rely on ML algorithms to learn the functional forms from the data, and has achieved remarkably improved prediction accuracy over classical approaches in numerous retrospective benchmark studies.[9–15]

Four main metrics are typically considered to assess the performance of a SF, *i.e.*, the scoring power to estimate the linear correlation between the predicted and experimentally-determined binding strengths, the ranking power to assess the capability of a SF to rank the known ligands for a certain target, the docking power to evaluate the capability to discriminate near-native poses from computer-yielded decoy poses, and the screening power to evaluate the ability to identify the true binders for a certain target from a pool of decoy compounds.[16,17] An ideal SF should perform well across a wide

[a]Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, China. E-mail: yukang@zju.edu.cn; tingjunhou@zju.edu.cn; panpeichen@zju.edu.cn
[b]State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang, China
[c]School of Public Health, Zhejiang University, Hangzhou 310058, Zhejiang, China
[d]CarbonSilicon AI Technology Co., Ltd, Hangzhou 310018, Zhejiang, China
[e]Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou 213001, China

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3sc02044d



Journal of Medicinal Chemistry

pubs.acs.org/jmc

Article

Boosting Protein–Ligand Binding Pose Prediction and Virtual Screening Based on Residue–Atom Distance Likelihood Potential and Graph Transformer

Chao Shen, Xujun Zhang, Yafeng Deng, Junbo Gao, Dong Wang, Lei Xu, Peichen Pan,* Tingjun Hou,* and Yu Kang*

Cite This: *J. Med. Chem.* 2022, 65, 10691–10706

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The past few years have witnessed enormous progress toward applying machine learning approaches to the development of protein–ligand scoring functions. However, the robust performance and wide applicability of scoring functions remain a big challenge for increasing the success rate of docking-based virtual screening. Herein, a novel scoring function named RTMScore was developed by introducing a tailored residue-based graph representation strategy and several graph transformer layers for the learning of protein and ligand representations, followed by a mixture density network to obtain residue–atom distance likelihood potential. Our approach was resolutely validated on the CASF-2016 benchmark, and the results indicate that RTMScore can outperform almost all of the other state-of-the-art methods in terms of both the docking and screening powers. Further evaluation confirms the robustness of our approach that can not only retain its docking power on cross-docked poses but also achieve improved performance as a rescoring tool in larger-scale virtual screening.

INTRODUCTION

The accurate prediction of protein–ligand binding modes is a long-standing challenge in structure-based drug design. Several experimental techniques such as X-ray diffraction,[1,2] nuclear magnetic resonance (NMR) crystallography,[3] and cryo-electron microscopy (EM)[4] can be used to determine the structural details of protein–ligand complexes; however, they are always expensive, and their availability is difficult for all desired cases. Molecular docking has been proposed as a computational alternative, which has posed a considerable impact on drug design, including hit/lead discovery and the following lead optimization.[5−8] Typically, a docking program needs a searching engine to sample the possible binding poses of the studied molecule in the binding pocket first and then uses a scoring function (SF) to assess their binding strength, and the top-ranked pose is usually considered as the most reasonable binding conformation. Despite impressive achievements in the past several decades, the inaccuracy of SFs continues as a major bottleneck in reliable molecular docking for real-world applications.[9,10]

The SFs are typically divided into four types, namely, physics-based SFs, empirical SFs, knowledge-based SFs, and machine-learning-based SFs (MLSFs).[11] The former two are usually described as the weighted sum of several physical/empirical energy terms/descriptors. Knowledge-based SFs are built through the statistical analysis of the distribution of some geometric features in protein–ligand structures and can then also be represented as the sum of all pairwise statistical potentials between proteins and ligands. The rapid development of machine learning (ML) and artificial intelligence (AI) in recent years leads to the emergence of MLSFs, most of which can directly learn the function form from data while not relying on the predefined additive formulated hypothesis.[12−15] However, despite the substantially superior performance as compared to the former three classical SFs in most studies, MLSFs have suffered from a cloud of doubts due to their poor generalization capability.[16−21] For example, Gabel et al.[16] found that their MLSFs trained on random forest (RF) and support vector machine (SVM) could reproduce the claimed excellent scoring power (the capability for binding affinity prediction) of RFscore,[22] but their docking power (the capability for the discrimination of near-native poses from decoy poses) and screening power (the capability for the discrimination of active ligands from decoy compounds) are rather bad. Our previous

Received: June 22, 2022
Published: August 2, 2022