Research, Society and Development, v. 11, n. 8, e40311831222, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i8.31222

Affinity estimation models of proteins for intelligent drug design based on pseudoconvolutions and nonlinear regressors

Modelos de estimativa de afinidade de proteínas para design inteligente de drogas com base em pseudoconvoluções e regressores não lineares

Modelos de estimación de afinidad de proteínas para el diseño inteligente de fármacos basados en pseudoconvoluciones y regresores no lineales

Received: 05/31/2022 | Reviewed: 06/10/2022 | Accept: 06/12/2022 | Published: 06/24/2022

Laila Barros Campos
ORCID: https://orcid.org/0000-0001-6687-1202
Universidade de Pernambuco, Brazil
E-mail: lbc@ecomp.poli.br
Janderson Romário Borges da Cruz Ferreira
ORCID: https://orcid.org/0000-0002-8410-7344
Universidade de Pernambuco, Brazil
E-mail: jrb@ecomp.poli.br
Wellington Pinheiro dos Santos
ORCID: https://orcid.org/0000-0003-2558-6602
Universidade Federal de Pernambuco, Brazil
E-mail: wellington.santos@ufpe.br

Abstract
Purpose: The emergence of new viruses and, consequently, new diseases make the rapid and precise design of new drugs increasingly necessary. With the availability of large databases of proteins and affinity measures, it is possible to build scoring functions for predicting molecular affinity. These functions are fundamental to intelligent drug design. Objective: In this work, we propose a scoring function to predict affinity between two proteins. The method is based on extracting features by transfer learning on sequences represented on pseudo-convolutions. Method: The pseudo-convolutions organize the sequences into base neighborhood distributions. Each distribution is then represented by an image. Two proteins are then transformed into two images that are concatenated together, forming the third image. Through deep transfer learning, this resulting image is then represented by a vector of attributes, which have dimensionality reduced by Random Forest. Finally, the vector of attributes reduced is applied to a regression learning machine that returns the degree of affinity of the two proteins. Results: We used the Affinity Benchmark Version 2 database. 145 complexes were used for model training and 35 for testing. The results showed a performance equal to or better than the state-of-the-art methods of evaluating protein affinity, considering the correlation coefficients of Pearson, Spearman and Kendall. The best results were 0.66, 0.70, and 0.52. Conclusion: The proposed method can characterize protein sequences so that the binding affinity between two proteins can be estimated without simulating the three-dimensional structure of the complex.
Keywords: Affinity markers; Amino acids, peptides and proteins; Artificial intelligence.

Resumo
Propósito: O surgimento de novos vírus e, consequentemente, novas doenças torna cada vez mais necessária a produção rápida e precisa de novos medicamentos. Com a disponibilidade de grandes bancos de dados de proteínas e medidas de afinidade, é possível construir funções de pontuação para prever a afinidade molecular. Essas funções são fundamentais para o design inteligente de medicamentos. Objetivo: Neste trabalho, propomos uma função de pontuação para prever a afinidade entre duas proteínas. O método é baseado na extração de características por transferência de aprendizado em sequências representadas em pseudo-convoluções. Método: As pseudo-convoluções organizam as sequências em distribuições de vizinhança de base. Cada distribuição é então representada por uma imagem. Duas proteínas são então transformadas em duas imagens que são concatenadas, formando a terceira imagem. Por meio de *deep transfer learning*, essa imagem resultante é então representada por um vetor de atributos, que tem a dimensionalidade reduzida por Random Forest. Por fim, o vetor de atributos reduzido é aplicado a uma máquina de aprendizado de regressão que retorna o grau de afinidade das duas proteínas. Resultados: Usamos o banco de dados Affinity Benchmark Versão 2. 145 complexos foram usados para treinamento do modelo e 35 para teste. Os resultados mostraram um desempenho igual ou superior aos métodos de avaliação de afinidade de proteínas no estado da arte, considerando os coeficientes de correlação de Pearson, Spearman e Kendall. Os melhores resultados foram 0.66, 0.70 e 0.52. Conclusão: O método

proposto pode caracterizar sequências proteicas de forma que a afinidade de ligação entre duas proteínas possa ser estimada sem simular a estrutura tridimensional do complexo.
Palavras-chave: Marcadores de afinidade; Aminoácidos, peptídeos e proteínas; Inteligência artificial.

Resumen

Propósito: La aparición de nuevos virus y, en consecuencia, de nuevas enfermedades hace cada vez más necesaria la producción rápida y precisa de nuevos fármacos. Con la disponibilidad de grandes bases de datos de proteínas y medidas de afinidad, es posible crear funciones de puntuación para predecir la afinidad molecular. Estas funciones son fundamentales para el desarrollo inteligente de fármacos. Objetivo: En este trabajo, proponemos una función de puntuación para predecir la afinidad entre dos proteínas. El método se basa en la extracción de características por transferencia de aprendizaje en secuencias representadas en pseudoconvulsiones. Método: Las pseudoconvulsiones organizan secuencias en distribuciones de vecindario base. Cada distribución se representa mediante una imagen. Luego, dos proteínas se transforman en dos imágenes que se concatenan, formando la tercera imagen. A través del aprendizaje de transferencia profundo, esta imagen resultante se representa en un vector de atributos, que he reducido dimensionalmente por Random Forest. Finalmente, el vector de atributos reducido se aplica a un algoritmo de regresión que devuelve el grado de afinidad de las dos proteínas. Resultados: Utilizamos la base de datos Affinity Benchmark Versión 2. Se utilizaron 145 complejos para entrenar el modelo y 35 para probar. Los resultados mostraron un desempeño igual o superior a los métodos de evaluación de afinidad de proteínas de última generación, considerando los coeficientes de correlación de Pearson, Spearman y Kendall. Los mejores resultados fueron 0.66, 0.70 y 0.52. Conclusión: El método propuesto puede caracterizar secuencias de proteínas de modo que se pueda estimar la afinidad de unión entre dos proteínas sin simular la estructura tridimensional del complejo.
Palabras clave: Marcadores de afinidad; Aminoácidos, péptidos y proteínas; Inteligencia artificial.

1. Introduction

Industrialization and urbanization processes, climate change, the rapid evolution of viruses, and greater interaction between (humans and animal species from different biomes have contributed to a greater spread of viruses and, consequently, of diseases (Tian et al., 2018; Baca-Carrasco et al., 2016). These factors, combined with the increasingly connected nature of the world, with increasingly intense flows of human populations, arising from the needs of trade, tourism and migratory flows linked to wars and catastrophes, have contributed to the emergence of epidemiological outbreaks, epidemics and even pandemics, such as the recent Covid-19 pandemic, caused by the SARS-CoV-2 coronavirus (Tian et al., 2018; Baca-Carrasco et al., 2016; Nicola et al., 2020).

The recent human history of fighting viruses and epidemics has contributed to the construction of voluminous databases that list proteins and their degree of affinity. These bases are very useful for designing new drugs. Traditional techniques for drug design are essentially trial-and-error techniques and are time- and resource-expensive (Khamis et al., 2015) processes. Traditional drug development typically involves 5 to 15 years of research, from molecular design to human testing, consuming millions of dollars (Khamis et al., 2015; Hung & Chen, 2014; Baldi, 2010; Katara, 2013).

The smart drug project takes advantage of the existence of extensive protein databases built over decades of research. Smart drug design also benefits from recent advances in Computer Science, in particular Computational Intelligence and, more specifically, Evolutionary Computing and Machine Learning, saving time and resources over traditional techniques (Zhang et al., 2019; Khamis et al., 2015; Ballester & Mitchell, 2010). This type of project demands good scoring functions, capable of abstracting biochemical and biophysical experiments as much as possible by predicting the geometric conformation of the evaluated proteins (Zhang et al., 2019; Khamis et al., 2015; Ballester & Mitchell, 2010).

In this work, we propose a new protein affinity assessment function, based on the representation of proteins through pseudo-convolutions. These pseudo-convolutions organize the sequences into base neighborhood distributions. Each distribution is then represented by an image. Two proteins are then transformed into two images that are concatenated together, forming a third image. Through deep transfer learning, this resulting image is then represented by a vector of attributes, which had dimensionality reduced by Random Forest. The vector of attributes reduced is applied to a regression learning machine that returns the degree of affinity of the two proteins. To validate our proposal, we used the Affinity Benchmark Version 2 database.

This database contains 7 types of complexes, 145 of which are used for model training and 35 for testing. The results showed a performance equal to or better than the state-of-the-art methods of evaluating protein affinity, considering the correlation coefficients of Pearson, Spearman and Kendall, and the Mean Squared Error (RMSE).

3. Related Works

Gomes et al. (2021) presented a new method, pseudo-convolutional machines, for representation of genomic sequences based on the analysis of the relationship between nitrogenous bases, then used machine learning algorithms to classify DNA samples from twenty-five viruses represented by the new method. The experiments were carried out in four different scenarios: The first aimed to evaluate the efficiency of the proposed method by carrying out experiments with twenty-five different viruses, including SARS-CoV-2. Five types of classifiers were tested: IBk, Multilayer Perceptron (MLP), Naive Bayes (NBC), Random Forest and Support Vector Machines (SVM). The second scenario was a binary classification due to the fact that the highest number of false positives for SARS-CoV-2 are from Coronaviridae in the multiclass scenario. The third scenario was classification tests done with SARS-CoV-2 and three other viruses with similar symptoms to each other. The fourth scenario was evaluated using the same methodology as the actual SARS-CoV-2 test, having three classes for classification: SARS-CoV-2 (the test target), GRCh38 (the healthy human reference) and Coronaviridae (a sample virus control system). The results obtained in all test scenarios evidenced the ability of the proposed approach to identify viruses using DNA sequences with high accuracy. Furthermore, the result of the real test scenario experiment indicates that one can optimize the molecular diagnosis of Covid-19 by combining reverse transcription followed by polymerase chain reaction (RT-PCR) with pseudo-convolutional machines for feature extraction.

Guedes et al. (2021) proposed the DockTScore, which is a set of new empirical scoring functions to estimate protein-ligand binding affinity. DockTScore has two steps: decryption and prediction. In the decryption step it is necessary that the physics-based interaction terms that contribute to the binding free energy are explicitly entered. DockTScore descriptors are based on the force field (MMFF94s) (Halgren, 1996), and have been trained and validated on large, properly curated high quality datasets. The prediction step was evaluated with machine learning algorithms that consider linear and non-linear data, the algorithms used were, Multiple Linear Regression (MLR) (Lai et al., 1978), Support Vector Machine (SVM) for Regression (SMOReg) (Shevade et al., 2000) and Random Forest (RF) (Breiman, 2001). It was concluded that DockTScore scoring functions perform similarly to the current highest-rated scoring functions in terms of binding energy prediction and ranking on four Database of Useful Decoys: Enhanced (-E) datasets. Finally, the authors concluded that the proposal will be useful in drug creation projects in the computational simulation stage for several proteins, as well as for specific targets, such as proteases and protein-protein interactions.

Wójcikowski et al. (2017) pitted machine learning scoring functions against classical scoring functions, comparisons were made in terms of virtual screening and affinity prediction. According to the authors, scoring functions based on Random Forest currently have one of the best performances in affinity prediction. Three scoring functions based on the algorithm were developed, each function being trained in a best-scoring ligand pose (ie, the lowest score found using matching fitting programs). The number of trees established for the experiments was 500. The number of features to be considered when looking for the best split in each RF tree was optimized using out-of-bag (OOB) predictions (James et al., 2013). The results of this work show that the scoring functions based on machine learning achieved affinity prediction accuracy of up to 88.6\% in all target proteins of the DUD-E base, which can be considered as a good performance. On the other hand, in the context of structure-based virtual screening, coupling algorithms using machine learning-based scoring functions were able to accurately predict binding affinity (Pearson Correlation = 0.56).

Wang & Zhang (2017) developed a new scoring function applying Random Forest to parameterize a correction measure to the original AutoDock Vina score. The Vina scoring function consists of six components: two steric Gaussian terms, a repulsion term, a hydrogen bonding (HB) term, and a twist count factor. It is possible to vary the parameters of these components to improve the position estimation and affinity prediction of the coupling to be analyzed (Trott & Olson, 2010). The new Δ_{vina}RF scoring function, which employs twenty descriptors in addition to the AutoDock Vina scoring, outperformed the classic scoring functions on all CASF-2013 and CASF-2007 benchmarks tests, including affinity prediction, ranking, fitting and screening potency tests. In terms of affinity prediction, the Δ_{vina}RF score significantly outperforms the AutoDock Vina as well as all scoring functions that were tested in the original benchmarking articles from 2013 and 2007. It achieves the best coefficients Pearson correlation coefficients of 0.686 and 0.732 for the CASF-2013 and CASF-2007 benchmarks respectively, and significantly improves on AutoDock Vina, which has corresponding Pearson correlation coefficients of 0.557 and 0.566, respectively.

Ragoza et al. (2017) developed scoring functions based on the Convolutional Neural Network (CNN). The proposed model was trained to classify couplings using 3D grid representation of protein-ligand structures during coupling. The CNN scoring functions were divided into two training sets, one focused on pose prediction and the other on coupling affinity. Due to the fact that the input of the CNNs were originally images discretized in pixels, it was necessary to carry out a transformation in the 3D structural data. In the pose prediction task, the database used to train the functions was the CSAR-NRC HiQ, with the addition of the CSAR-HiQ Update. To evaluate the proposed models the databases were split using 3-times cross validation for the pose prediction task as well as for the affinity prediction task. In order to avoid evaluating the models on targets similar to the training set, the training and testing folds were constructed by grouping data based on target families rather than individual targets. The results in the pose prediction task show that the proposed CNN obtained an area under the ROC Curve equal to 0.815 while the coupling program Autodock Vina reached only 0.645. From the perspective of affinity prediction, the following were evaluated: proposed CNN, Autodock Vina, Random Forest Score, NNScore (Durrant & McCammon, 2011), the results respectively were: 0.868, 0.716, 0.622 and 0.584. After analyzing the results, it was evidenced that the scoring functions based on CNN have the potential to surpass current methods. In addition, it is still possible to improve CNN models, such as training with larger datasets covering a range of goals (e.g. pose classification, affinity prediction, virtual triage etc.).

Vreven et al. (2015) updated their protein-protein docking and affinity prediction benchmarks. The new benchmarks consist of high-quality, non-redundant structures of protein-protein complexes along with the unbound structures of their components. Fifty-five new complexes were added to the coupling benchmark, 35 of this set had binding affinities experimentally measured. Currently the updated coupling position estimate and affinity prediction benchmarks now contain 230 and 179 entries, respectively. The authors concluded that after adding the new samples the test became more challenging for position estimation and affinity prediction algorithms: Success rates of structure prediction and correlations with experimentally obtained affinities are lower than reported using versions previous benchmark. Since updating these benchmarks, they have been widely used by the community to improve coupling algorithms between proteins, and also to understand biomolecular interactions.

3. Methodology

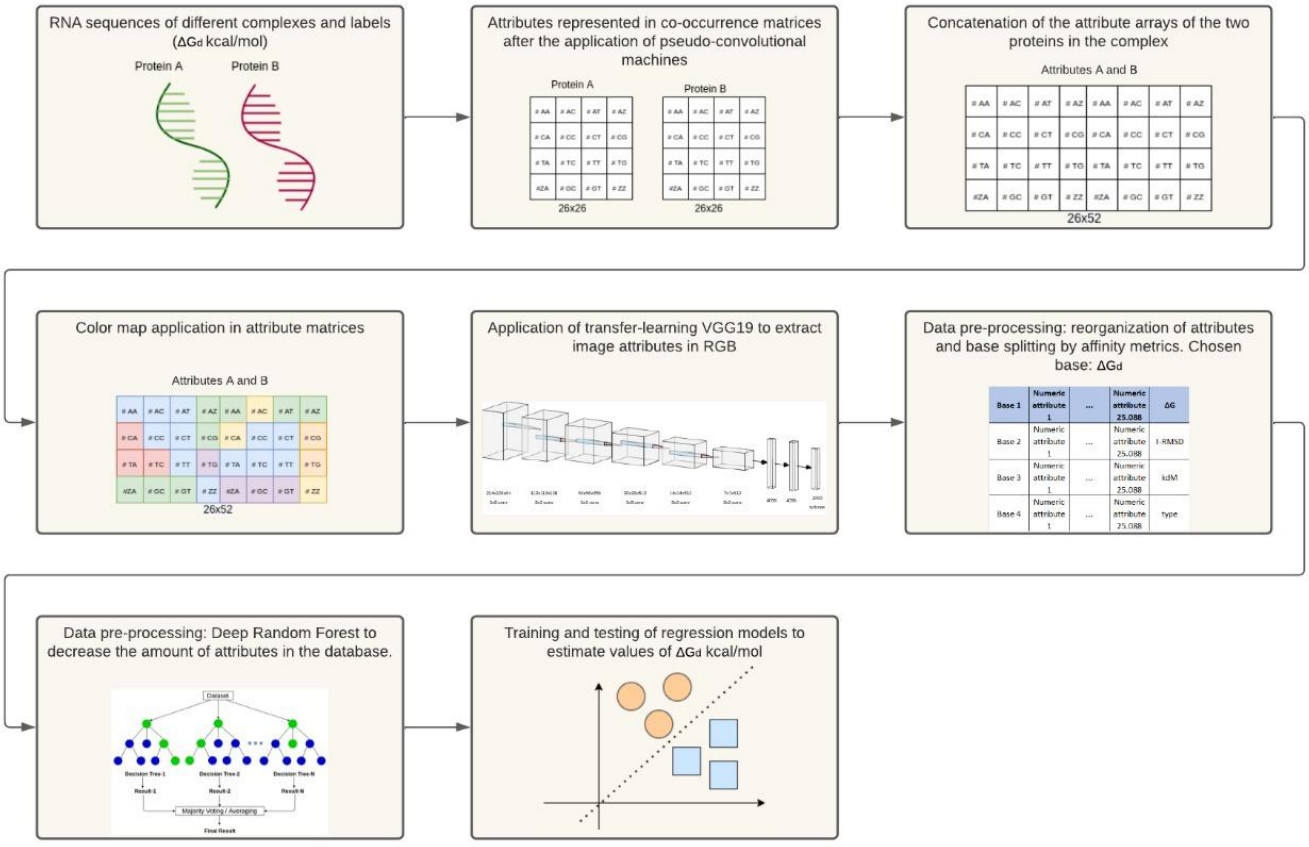
3.1 Proposed Method

This work proposes a new method for estimating the affinity between RNA sequences of two proteins, consisting of two main steps: extraction of the characteristics (attributes) of the sequences and methods of estimating the affinity between them. With this method, we are bringing an alternative to estimate affinity between proteins without using 3D simulations of the structures of the complexes.

Data extraction is done in two stages: pseudo-convolution and transfer-learning. The first pseudo-convolution step

follows the same approach presented by Gomes et al. (2021) for extracting characteristics from the DNA sequences of viruses for the diagnosis of Covid-19 and, for this work, we aimed to use this strategy to generate matrices of co-occurrence for the sequences of each protein in the complexes. RNA sequences consist of combinations of nucleotide pairs, e.g. (AA, AB, CD, etc.), forming a 26 x 26-dimensional matrix. These pair combinations are defined by reading the sequences from left to right, and, for each pair formed, its occurrence is counted. For example, if the pair AA is identified, the cell from column A and row A increases. This process is repeated for each sequence of the two proteins, resulting in two co-occurring matrices for each concatenated complex. Then, the two matrices are concatenated, resulting in a new matrix with dimensions 26 x 52. A color map is applied to them to represent them on the RGB scale. After this process, the attributes are extracted through the transfer-learning method with a VGG19 (deep neural network trained on ImageNet dataset) (Deng et al., 2009). Finally, A Random Forest with ten trees was applied to reduce the number of attributes. This step was crucial due to the many attributes coming from VGG19. As a result, the Random Forest decreased significantly from 25,088 to 2,110 attributes.

Figure 1 - Boxplot of the top 3 regressors: a) coefficient correlation of Pearson. b) coefficient correlation of Spearman.



Source: The Authors.

3.2 Database

The database used for training and evaluating the regression models in this work was the Affinity Benchmark Version 2, part of an update of the affinity and docking benchmark presented by Vreven et al. (2015), where its samples can be found in the database. PDBbind data. The database has 179 protein complexes of 7 different types to be used in affinity prediction. The new benchmarks consist of high-quality non-redundant structures of protein complexes. Table 1 presents a summary of this new

benchmark.

Table 1 - Demonstration of the organization of Affinity Benchmark 2.

PDB Complex	Type	Protein A	Protein B	Kd (M)	dG	I-RMSD	Temperature (C°)	pH
3L5W_LH:I	A	C836 FAB	Interleukin-13	0.00000000005	-14	0	25	7.4
3L89_ABC:M	OR	Ad21 fiber knob	CD46 SCR1 and SCR2 domains	0.00000028400	-9	3	25	7.4
3LVK_AC:B	E	Cysteine desulfurase IscS	Sulfurtransferase tusA	0.00000030400	-9	1	25	7.4
4DN4_LH:M	A	CNTO888 FAB	MCP-1	0.00000000004	-14	1	25	7.1
4FQI_HL:ABEFCD	A	CR9114 FAB	H5N1 influenza virus hemagglutinin	0.00000000090	-13	1	30	7.4

Source: The Authors.

3.3 Experimental Setup

To evaluate our proposal, we selected four traditional regressors, Linear Regression, Random Forest, SVM, and MLP. The models were trained using cross-validation with k=10. The objective of the training was to predict the ΔG of two-protein complexes. Also, we defined several hyperparameters to train each regressor. The Table 2 shows the hyperparameters used to train these algorithms. Finally, to ensure that the model results were statistically comparable, each model generated by a hyperparameter set was trained and tested 30 times.

Table 2 - Hyperparameters used to Train the Regressors.

Regressors	Parameters	
Linear Regression	-	
Random Forest	Number of trees: 50, 100, 200 e 300	
Support Vector Machine (SVM)	C: 0.1, 1.0 e 10.0	
	Linear Kernel	
	Polynomial Kernel	Degree: 2 and 3
	RBF Kernel:	g: 0.01, 0.25 and 0.50
Multilayer Perceptron (MLP)	Hiden Layers	1
	Neurons: 2, 5, 10, 20 and 50	

Source: The Authors.

The evaluation metrics used in the process of training and validating the effectiveness of these algorithms were the means and standard deviations of the coefficients of Pearson, Spearman, Kendall, and the Mean Squared Error (RMSE).

4. Results and Discussion

Due to the many experiments performed and the metrics used, the results were organized in table and boxplots to facilitate understanding. In Table 4 are the results of all the regressors evaluated in this project. Also, Figure 2 shows only the three best regressors according to the coefficient correlation of Pearson, Spearman, and Kendall. In order to compare the performance of the models, the correlation coefficients were set in box plots. These boxplots contain only the three best configurations of each regressor to avoid a bigger image. In addition, the Linear Regression was included in the graphs for comparison purposes.

Table 4 - Results from all regression models were evaluated in the test dataset. The models that achieved the best results according to the defined metrics were bold.

	Pearson		Spearman		Kendall		RMSE	
Model	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**	**Avg**	**Std**
Linear Regression	0.2708	0.1458	0.2786	0.1435	0.2307	0.1189	2.7295	0.301
Random Forest (50)	0.6306	0.0939	0.6792	0.0839	0.4958	0.0733	2.2511	0.305
Random Forest (100)	0.6452	0.0905	0.6944	0.0874	0.5098	0.078	2.2349	0.301
Random Forest (200)	0.6501	0.0874	0.7031	0.0834	0.5183	0.0745	2.2302	0.29
Random Forest (300)	0.6538	0.0849	**0.7067**	**0.0812**	**0.5216**	**0.073**	2.2271	0.294
SVM (linear, C=0,1)	0.609	0.0864	0.6512	0.0909	0.47	0.0778	2.3203	0.258
SVM (linear, C=1,0)	0.6091	0.0863	0.6511	0.0908	0.4699	0.0777	2.3203	0.258
SVM (linear, C=10,0)	0.6091	0.0864	0.6515	0.0908	0.4703	0.0777	2.3202	0.258
SVM (poli 2, C=0,1)	0.5964	0.09	0.6623	0.0936	0.4778	0.0798	2.2658	0.294
SVM (poli 2, C=1,0)	0.5964	0.0899	0.6621	0.0936	0.4778	0.0797	2.2658	0.294
SVM (poli 2, C=10,0)	0.5964	0.0899	0.6621	0.0938	0.4777	0.08	2.2659	0.294
SVM (poli 3, C=0,1)	0.5456	0.1014	0.6376	0.1012	0.4558	0.0825	2.3892	0.300
SVM (poli 3, C=1,0)	0.5456	0.1015	0.6376	0.1012	0.4559	0.0825	2.3893	0.300
SVM (poli 3, C=10,0)	0.5456	0.1014	0.6374	0.101	0.4556	0.0824	2.3892	0.300
SVM (rbf, C=0,1, g=0,01)	**0.6645**	**0.0726**	0.6932	0.0819	0.5025	0.0715	2.2384	0.289
SVM (rbf, C=0,1, g=0,25)	0.0956	0.1966	0.529	0.1093	0.3911	0.0846	2.7977	0.281
SVM (rbf, C=0,1, g=0,50)	0.0851	0.1261	0.1688	0.1345	0.1381	0.1101	2.7976	0.281
SVM (rbf, C=1,0, g=0,01)	0.6521	0.0841	0.6997	0.0852	0.5168	0.0743	**2.1383**	**0.294**
SVM (rbf, C=1,0, g=0,25)	0.0883	0.1914	0.5363	0.1114	0.4	0.0873	2.8016	0.282
SVM (rbf, C=1,0, g=0,50)	0.0613	0.1622	0.1556	0.1419	0.1269	0.1161	2.8015	0.282
SVM (rbf, C=10,0, g=0,01)	0.649	0.0848	0.697	0.0855	0.5154	0.0752	2.1448	0.294
SVM (rbf, C=10,0, g=0,25)	0.0877	0.1951	0.5383	0.1102	0.4013	0.0869	2.8016	0.282
SVM (rbf, C=10,0, g=0,50)	0.0655	0.1593	0.1533	0.14	0.125	0.1145	2.8015	0.282
MLP (2 neurons)	-0.1048	0.2762	-0.1339	0.4357	-0.0938	0.3047	3.1242	0.551
MLP (5 neurons)	-0.1249	0.3347	-0.1411	0.399	-0.099	0.2788	3.157	0.589
MLP (10 neurons)	-0.0932	0.31	-0.1127	0.359	-0.0759	0.2509	3.2154	0.625
MLP (20 neurons)	-0.0691	0.3177	-0.0755	0.3669	-0.0537	0.2577	3.5006	1.29
MLP (50 neurons)	0.0435	0.268	0.0451	0.2977	0.0321	0.2062	3.5812	1.264

Source: The Authors.

Figure 2 shows that regardless of the correlation metric used, the Random Forest and SVM regressors achieve higher and more consistent performance since they have a lower standard deviation. On the other hand, MLP is heavily dependent on data and initialization. Given the nature of the problem, estimating protein affinity for drug creation, the models must be stable and less dependent on the data.

.

Figure 2 - Boxplot of the top 3 regressors:
a) coefficient correlation of Pearson. b) coefficient correlation of Spearman. c) coefficient correlation of Kendall.

Source: The Authors.

5. Conclusion

Given the experiments carried out, the results obtained in this work could satisfactorily fulfill the general objectives initially outlined. The investigated database derived from Affinity Benchmark Version 2 presented in (Vreven et al., 2015) was well explored, despite having few records to be used as a training and testing basis in the construction of a prediction model. In addition, the technique of extracting information from the complexes by transfer-learning resulted in a large number of attributes and was disproportionate to the number of database instances. For this, we use Deep Random Forest and cross-validation techniques to generate a greater proportionality between the number of attributes and the number of instances, contributing positively to the training and testing phases of the regression model.

The proposal to create a sub-architecture based on a pre-trained convolutional neural network to extract features from complex images was possible with pseudo-convolutional machines to represent the attributes in an image with the transfer-learning VGG19 to extract the attributes of these created images. These attributes were used as input data for the classical regression methods, finalizing the structure of the proposed prediction model.

Among the regressors used to predict the final affinity of the complexes, the ones that obtained the best results about the correlation coefficient metrics were the Random Forest and SVM configurations. The best configuration of Random Forest tested was with 300 trees. The best of SVM was with the exponential kernel function RBF, regularization parameter C of 0.1, and kernel coefficient gamma of 0.01.

As future works, we hope to investigate models deeply for extracting features from the images of protein complexes, also performing further studies with more extensive databases to obtain more satisfactory affinity prediction results. It is even

possible to carry out a more significant investigation of other regression methods in addition to those used in this work. We also intend to investigate other methods of representing protein sequences as signals or images. Additionally, we plan to investigate the modeling of the protein affinity assessment problem as a classification problem, transforming the most likely affinity levels, defined by stratification, into classes.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and the Brazilian agencies FACEPE and CNPq.

References

Baca-Carrasco, D.; Velasco-Hernández, J. X. (2016). Sex, mosquitoes and epidemics: an evaluation of zika disease dynamics. *Bulletin of Mathematical Biology*, 78 (11), 2228–2242.

Baldi, A. (2010). Computational approaches for drug design and discovery: An overview. *Systematic reviews in Pharmacy*, 1 (1), 99.

Ballester, P. J.; Mitchell, J. B. (2010). A machine learning approach to predicting protein– ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26 (9), 1169–1175.

Breiman, L. (2001). Random forests. *Machine learning*, 45 (1), 5–32.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *In 2009 ieee conference on computer vision and pattern recognition* (p. 248-255).

Durrant, J. D.; McCammon, J. A. (2011). Nnscore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling*, 51 (11), 2897–2903.

Gomes, J. C.; Masood, A. I.; Silva, L. H. d. S., da Cruz Ferreira, J. R. B., Júnior, A. A. F.; dos Santos Rocha, A. L.,; de Oliveira, L.C. P.; da Silva, N. R. C.; Fernandes, B. J. T.; Dos Santos, W. P. (2021). Covid-19 diagnosis by combining rt-pcr and pseudo-convolutional machines to characterize virus sequences. *Scientific Reports*,11 (1), 1–28.

Guedes, I. A.; Barreto, A. M. S.; Marinho, D.; Krempser, E.; Kuenemann, M. A.; Sperandio, O.; Dardenne, L. E.; Miteva, M. A. (2021). New machine learning and physics-based scoring functions for drug discovery. *Scientific Reports*, 11 (1), 3198.

Halgren, T. A. (1996). Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of Computational Chemistry*, 17 (5-6), 490-519.

Hung, C.-L.; Chen, C.-C. (2014). Computational approaches for drug discovery. *Drug development research*, 75 (6), 412–418.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Katara, P. (2013). Role of bioinformatics and pharmacogenomics in drug discovery and development process. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 2 (4), 225–230.

Khamis, M. A.; Gomaa, W.; Ahmed, W. F. (2015). Machine learning in computational docking. *Artificial Intelligence in Medicine*, 63 (3), 135–152.

Lai, T. L.; Robbins, H.; Wei, C. Z. (1978). Strong consistency of least squares estimates in multiple regression. *Proceedings of the National Academy of Sciences of the United States of America*, 75 (7), 3034.

Nicola, M.; Alsafi, Z.; Sohrabi, C.; Kerwan, A.; Al-Jabir, A.; Iosifidis, C.; Agha, M.; Agha, R. (2020). The socio-economic implications of the coronavirus pandemic (covid-19): A review. *International journal of surgery*, 78 , 185–193.

Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57 (4), 942–957.

Shevade, S.; Keerthi, S.; Bhattacharyya, C.; Murthy, K. (2000). Improvements to the smo algorithm for svm regression. *IEEE Transactions on Neural Networks*, 11 (5), 1188-1193.

Tian, H.; Hu, S.; Cazelles, B.; Chowell, G.; Gao, L.; Laine, M.; Li, Y.; Yang, H.; Li, Y.; Yang, Q.; Tong, X.; Huang, R.; Bjornstad, O. N.; Xiao H.; Stenseth, N. C. (2018). Urbanization prolongs hantavirus epidemics in cities. *Proceedings of the National Academy of Sciences*, 115 (18), 4707–4712.

Trott, O.; Olson, A. J. (2010). Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31 (2), 455–461.

Vreven, T.; Moal, I. H.; Vangone, A.; Pierce, B. G.; Kastritis, P. L.; Torchala, M.; Chaleil, R.; Jiménez-García, B.; Bates, P. A.; Fernandez-Recio, J.; Bonvin, A. M. J. J.; Weng, Z. (2015). Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*, 427 (19), 3031–3041.

Wang, C.; Zhang, Y. (2017). Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *Journal of computational chemistry*, 38 (3), 169–177.

Wójcikowski, M.; Ballester, P. J.; Siedlecki, P. (2017). Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports*, 7 (1), 1–10.

Zhang, Y.; Wang, Y.; Zhou, W.; Fan, Y.; Zhao, J.; Zhu, L.; Lu, S.; Lu, T.; Chen, Y.; Liu, H. (2019). A combined drug discovery strategy based on machine learning and molecular docking. *Chemical Biology & Drug Design*, 93 (5), 685–699.