3rd International Conference on Chemo and BioInformatics

Kragujevac, September 25-26, 2025, Serbia

ICCBIKG 2025

International Conference on Chemo and BioInformatics

Book of Proceedings

ORGANIZERS AND SPONSORS

Republic of Serbia
MINISTRY OF SCIENCE, TECHNOLOGICAL DEVELOPMENT AND INNOVATION

MC LABOR
LABORATORIJSKA I PROCESNA OPREMA

Investigating protein folding as a target for neurodegenerative disorders using a biological activity prediction model based on deep learning

Thomas Papikinos[1], Marios Krokidis[1,2], Aris Vrahatis[1,2], Panagiotis Vlamos[1,2] Themis Exarchos[1,2,*]

[1] Bioinformatics and Human Electrophysiology Laboratory, Department of Informatics, Ionian University, Corfu, Greece
[2] Institute of Digital Biomedicine, University Center for Research and Innovation, Ionian University, Corfu, Greece
** Correspondence:* exarchos@ionio.gr

DOI: 10.46793/ICCBIKG25.319P

Abstract : Protein misfolding is a hallmark feature of neurodegenerative disorders (NDs), playing a central role in their pathogenesis by disrupting cellular proteostasis and leading to neuronal degeneration. Molecular chaperones, such as heat shock proteins, are crucial in maintaining protein homeostasis by assisting in proper protein folding, preventing aggregation, and facilitating the clearance of misfolded proteins. A machine learning framework based on neural networks has been developed that predicts how much a given compound can enhance the activity of a target protein. This approach leverages large-scale biological data to connect chemical space with functional outcomes, providing a systematic way to explore therapeutic potential across diverse compounds. The model can be applied to screen drug libraries for compounds that increase the activity of ND-related chaperones, potentially offering therapeutic effects and also aims to aid in the acceleration of drug repurposing efforts for NDs, contributing to a better understanding of therapeutic options targeting protein homeostasis.

Keywords: drug repurposing, neurodegenerative disorders, protein folding, biological activity prediction, drug-target interaction prediction

1. Introduction

Protein folding (PF) is the process by wich a chain of amino acids folds into a stable three-dimensional structure, and when it fails to occur properly, the protein can become dysfunctional or toxic. Protein misfolding is central to neurodegenerative disorders (NDs). Moreover, misfolded proteins, due to their nature, can accumulate and clump into aggregates that disrupt cellular function and contribute to neuronal damage [1]. Therefore, targeting the PF process represents a potentially viable therapeutic strategy for NDs. Among the key components involved in maintaining proper protein folding are heat shock proteins (HSPs), which play a crucial role in protein folding, assembly, disassembly, and clearance. Their expression is primarily regulated by heat shock factor 1 (HSF1), a transcription factor that drives HSP gene expression [2]. Disruptions in HSF1 function adversely affect protein homeostasis and are strongly associated with diseases, including NDs. The aim of this study was the development of a machine learning model

that predicts drug-protein activities, which can be used to identify small molecules that increase HSF1 activity, ultimately enhancing HSP expression which can help restore proteostasis in neurodegenerative conditions.

2. Materials and Methods

The model was trained on 60,591 drug-protein pairs labeled with the $\log_{10}$ of their activity value, sourced from the database BindingDB [3], and randomly split into training, validation and test sets in proportions of 70%, 10% and 20%, respectively. Drugs and target proteins were represented with their simplified molecular input line entry system (SMILES) ASCII representation and amino acid sequence respectively, while activity values were denoted by EC_{50} (the concentration of a drug, expressed in nanomolar units (nM), required to produce half of the maximal biological response of the target protein).

The architecture of the model was as follows: drugs were encoded with a transformer and proteins with a convolutional neural network. Then, the concatenation of those encodings was fed into a multilayered perceptron for 30 epochs with the output being the predicted $\log_{10}$ value of the pair's EC_{50} in nM. Finally, its performance was evaluated on the unseen test set. In order to demonstrate its robustness, the model was also externally validated on the database ChEMBL [4] which contains 78,805 drug-target activities.

3. Results

The regression metrics of the model were: mean square error (MSE) = 0.73, Pearson correlation = 0.79 and concordance index = 0.8. Moreover, a compound can be considered a strong activator of a protein if its EC_{50} value is 1000 nM or less [5,6] since greater potency corresponds to requiring lower concentrations to achieve the desired effect. Due to this, the model was also validated as a classification problem, with the positive case being when $EC_{50} \leq 1000$ nM and the negative otherwise. The classification metrics are presented in Table 1.

Table 1. Classification metrics.

	Unseen test set from BindingDB	External validation on ChEMBL
Accuracy	0.83	0.71
Precision	0.85	0.77
Recall	0.91	0.82
F1 Score	0.88	0.8

4. Conclusion

The model demonstrated a reasonable level of precision on the unseen BindingDB test set (0.86), suggesting that predicted active compounds are likely to show measurable activity, which makes it a potentially useful tool for preliminary drug screening. To explore its applicability, we applied the model to a set of 2,587 drugs approved by the U.S. Food and Drug Administration (FDA) and found that several of the top-scoring

compounds are either already prescribed or under clinical investigation for NDs, consistent with the potential benefits of increased HSF1 activity. Compounds with the lowest 8 predicted EC_{50} values are displayed in Table 2. Among its other functions, HSF1's primary role is the regulation of HSP synthesis in response to stress.HSPs help stabilize unfolded proteins, disassemble protein aggregates, and direct misfolded proteins toward degradation, positioning HSF1 as a key factor that must be regulated in NDs [7]. Indicatively, Sirolimus (i.e. rapamycin), the FDA-approved immune-modulator (predicted EC_{50}: 32 nM), is a potent neuroprotective agent in several experimental models. By inhibiting mTOR, it influences autophagy which is beneficial for NDs [8]. In a recent pilot phase 1 study, rapamycin was not detected in cerebrospinal fluid before or after treatment while short-term, low-dose treatment was generally well-tolerated by older adults with mild cognitive impairment or mild dementia, clinically diagnosed as Alzheimer's disease [9]. Siponimod (predicted EC_{50}: 45 nM), is an approved drug for multiple sclerosis, as it modulates the S1P receptor, preventing white blood cells from attacking the central nervous system [10]. A phase 3 clinical trial indicated that siponimod significantly reduced (versus placebo) ganglion cell and inner plexiform layers thinning at month 24 and also suggested that optical coherence tomography measurement of retinal atrophy could serve as a non-invasive potential biomarker for assessing treatment effects on neurodegeneration in secondary progressive MS [11].

Table 2. Top 8 drugs with the lowest predicted EC_{50} values.

Drug name	Predicted EC_{50} in nM	Original drug indication
Micafungin	1.0	Fungal infections
Linzagolix	3.0	Uterine fibroids
Nizatidine	5.4	Stomatch ulcers
Sodium stibogluconate	6.6	Leishmaniasis
Ledipasvir	7.3	Hepatitis C
Bacitracin	7.5	Skin infections

Beyond the previous use case, the model could be integrated into high-throughput screening (HTS) pipelines which typically involve testing hundreds of thousands of compounds, making it a costly and resource-intensive process. Because the model has a promising recall value (0.91), it is less likely to discard true actives, meaning it can serve as a prescreening step that reduces the chemical space explored by HTS. To demonstrate this, when it was applied to the Enamine HLL-100 database (https://enamine.net/compound-libraries/diversity-libraries) of 100,160 compounds [12], imposing an EC_{50} activity cutoff of 1000 nM reduced the set to 44,556 candidates. This kind of reduction demonstrates practical value, as it could allow large-scale screening efforts to focus on a more manageable subset without substantially increasing the risk of missing potential hits. Finally, it is noted that the model is target agnostic, meaning that it can be used with any drug-target pair, and is available to the public in the address stated in the data availability section.

Data availability: The model created for this study as well as all related data can be found on https://doi.org/10.6084/m9.figshare.30018616.

Acknowledgments: This work was partially supported by the framework of the Action 'Flagship Research Projects in challenging interdisciplinary sectors with practical applications in Greek industry', implemented through the National Recovery and Resilience Plan Greece 2.0 and funded by the European Union—NextGenerationEU (project code: TAEDR-0535850).

References

[1] P. Sweeney *et al.*, "Protein misfolding in neurodegenerative diseases: implications and strategies," *Transl. Neurodegener.*, vol. 6, p. 6, Mar. 2017, doi: 10.1186/s40035-017-0077-5.

[2] J. Barna, P. Csermely, and T. Vellai, "Roles of heat shock factor 1 beyond the heat shock response," *Cell. Mol. Life Sci. CMLS*, vol. 75, no. 16, pp. 2897–2916, May 2018, doi: 10.1007/s00018-018-2836-6.

[3] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, "BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1045–D1053, Jan. 2016, doi: 10.1093/nar/gkv1072.

[4] B. Zdrazil *et al.*, "The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods," *Nucleic Acids Res.*, vol. 52, no. D1, pp. D1180–D1192, Jan. 2024, doi: 10.1093/nar/gkad1004.

[5] V. N. T. La, S. Nicholson, A. Haneef, L. Kang, and D. D. L. Minh, "Inclusion of Control Data in Fits to Concentration–Response Curves Improves Estimates of Half-Maximal Concentrations," *J. Med. Chem.*, vol. 66, no. 18, pp. 12751–12761, Sept. 2023, doi: 10.1021/acs.jmedchem.3c00107.

[6] K. Katsuno *et al.*, "Hit and lead criteria in drug discovery for infectious diseases of the developing world," *Nat. Rev. Drug Discov.*, vol. 14, no. 11, pp. 751–758, Nov. 2015, doi: 10.1038/nrd4683.

[7] R. Gomez-Pastor *et al.*, "Abnormal degradation of the neuronal stress-protective transcription factor HSF1 in Huntington's disease," *Nat. Commun.*, vol. 8, no. 1, p. 14405, Feb. 2017, doi: 10.1038/ncomms14405.

[8] J. Bové, M. Martínez-Vicente, and M. Vila, "Fighting neurodegeneration with rapamycin: mechanistic insights," *Nat. Rev. Neurosci.*, vol. 12, no. 8, pp. 437–452, Aug. 2011, doi: 10.1038/nrn3068.

[9] M. M. Gonzales *et al.*, "Rapamycin treatment for Alzheimer's disease and related dementias: a pilot phase 1 clinical trial," *Commun. Med.*, vol. 5, no. 1, p. 189, May 2025, doi: 10.1038/s43856-025-00904-9.

[10] L. Cao *et al.*, "Siponimod for multiple sclerosis," *Cochrane Database Syst. Rev.*, vol. 2021, no. 12, Nov. 2021, doi: 10.1002/14651858.CD013647.pub2.

[11] P. Vermersch *et al.*, "Effect of siponimod on retinal thickness, a marker of neurodegeneration, in participants with SPMS: Findings from the EXPAND OCT substudy," *Mult. Scler. Relat. Disord.*, vol. 94, p. 106259, Feb. 2025, doi: 10.1016/j.msard.2025.106259.