PhD Thesis summary

# Predicting binding affinity of small molecules in ligand-receptor systems using neural networks.

Wykorzystanie sieci neuronowych do przewidywania aktywności związków niskocząsteczkowych w układach ligand-receptor.

Marta Stępniewska-Dziubińska

**Supervisor:**
Dr hab. Paweł Siedlecki

Department of Bioinformatics
Institute of Biochemistry and Biophysics
Polish Academy of Sciences
Warsaw, Poland

Warsaw, 2020

# Thesis summary

This thesis is an interdisciplinary work, that combines computer-aided drug design (CADD) with machine learning (ML). The main focus of this work is usage of deep neural networks for assessing small molecule ability to bind to a molecular target of interest. This concept is illustrated with two manuscripts: "Development and evaluation of a deep learning model for protein–ligand binding affinity prediction" and "Improving detection of protein-ligand binding sites with 3D segmentation". In both works we propose new ways of addressing fundamental problems in structure-based drug design with 3D convolutional neural networks based on architectures used in computer vision. Network built for the first project – Pafnucy – predicts $pK_i$ values for protein-ligand complexes. The model from the second work – Kalasanty – detects binding sites at protein surfaces. The first manuscript also describes a featurization mechanism that was used in both models – algorithm for converting 3D molecular structures into a representation that can be used as an input for neural networks or other *in silico* methods.

To showcase advantages and flaws of deep learning methods for CADD, I compare it to two alternative approaches: classical, semi-manual CADD methods; and classical ML models combined with extensive feature engineering. As examples for these approaches I use two manuscripts I co-authored: "DeCAF – Discrimination, Comparison, Alignment Tool for 2D PHarmacophores" and "Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions". The first manuscript is an example of classical CADD method for ligand-based design, while the second one describes featurization mechanism developed for structure-based binding affinity prediction and its usage with different classes of ML methods.

The manuscripts are accompanied with an introduction, in which I compare and contrast presented approaches. I also briefly introduce fundamental concepts related to machine learning, focusing mostly on deep artificial neural networks, and computer-aided drug design and how these two fields can be combined.

# Streszczenie rozprawy

Niniejsza rozprawa stanowi interdyscyplinarną pracę łączącą komputerowe projektowanie leków (CADD, ang. *computer-aided drug design*) z uczeniem maszynowym (ML, ang. *machine learning*). Jej głównym tematem jest wykorzystanie głębokich sieci neuronowych w projektowaniu związków niskocząsteczkowych zdolnych do oddziaływania z wysokim powinowactwem z wybranym celem molekularnym. Koncept ten ilustruję za pomocą dwóch publikacji naukowych: "Development and evaluation of a deep learning model for protein–ligand binding affinity prediction" oraz "Improving detection of protein-ligand binding sites with 3D segmentation". W obu pracach zaprezentowaliśmy nowe sposoby zaadresowania fundamentalnych problemów w projektowaniu leków opartych o strukturę (ang. *structure-based drug design*), wykorzystując trójwymiarowe konwolucyjne sieci neuronowe oparte o architektury wykorzystywane w wizji komputerowej. Sieć stworzona na potrzeby pierwszego projektu – Pafnucy – służy do przewidywania wartości $pK_i$ dla kompleksów białko-ligand. Model przedstawiony w drugiej pracy – Kalasanty – wykrywa miejsca wiążące na powierzchni białek. Pierwsza z prac opisuje również reprezentację danych wykorzystywaną przez oba modele – trójwymiarową siatkę, opisującą gruboziarnisty rozkład cech atomowych w przestrzeni 3D.

Głębokie sieci neuronowe kontrastuję z powszechnie stosowanymi alternatywami: klasycznymi modelami wykorzystywanymi w CADD oraz podejściem opartym o uczenie maszynowe i inżynierię cech. Jako przykłady wykorzystuję publikacje naukowe: "DeCAF – Discrimination, Comparison, Alignment Tool for 2D PHarmacophores" oraz "Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions". Pierwszy z manuskryptów stanowi przykład klasycznej metody stosowanej w projektowaniu w oparciu o ligandy (ang. *ligand-based design*), natomiast drugi opisuje reprezentację danych stworzoną na potrzeby opisu oddziaływania białko-ligand oraz przykłady jego wykorzystania do przewidywania powinowactwa z pomocą modeli uczenia maszynowego.

Publikacje opatrzone są wstępem, w którym porównuję prezentowane podejścia oraz umieszczam je w szerszym kontekście. Opisuję również podstawowe pojęcia związane z uczeniem maszynowym, ze szczególnym uwzględnieniem głębokich sieci neuronowych, oraz komputerowym projektowaniem leków oraz to, jak te dwie dziedziny przeplatają się ze sobą.