Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: **Informatyka Techniczna (ITE)**
Specjalność: **Systemy informatyki w medycynie (IMT)**

PRACA DYPLOMOWA INŻYNIERSKA

Opracowanie i walidacja modelu predykcyjnego opartego na architekturze grafowej do oceny powinowactwa małych cząsteczek do białka BRD4

Kacper Wieczorek

Opiekun pracy
prof. dr hab. inż. Michał Woźniak

Słowa kluczowe: bioinformatyka, grafowe sieci neuronowe,

WROCŁAW (rok złożenia pracy dyplomowej)

Wrocław University of Science and Technology
Faculty of Information and Communication Technology

Field of study: **Computer Engineering (ITE)**
Speciality: **Information technology systems in medicine (IMT)**

BACHELOR THESIS

Development and validation of a predictive model based on graph architecture for assessing the affinity of small molecules for BRD4 protein

Kacper Wieczorek

Supervisor
prof. dr hab. inż. Michał Woźniak

Keywords: bioinformatics, graph neural networks

WROCŁAW (year of submission of the diploma thesis)

Streszczenie

Dodaj streszczenie pracy w języku polskim. Staraj się uwzględnić wymienione na stronie tytułowej słowa kluczowe. Uwaga przedstawiony rekomendowany szablon dotyczy pracy dyplomowej pisanej w języku angielskim. W przeciwnym wypadku, student powinien samodzielnie zmienić nazwy „Chapter" na „Rodział" itp stosując odpowiednie pakiety systemu LaTeXoraz ustawienia w pliku *latex-settings.tex*.

Abstract

Streszczenie w języku angielskim.

Spis treści

1. Wprowadzenie

Współczesna farmakologia nieustannie poszukuje nowych i coraz skuteczniejszych leków na największe zagrożenia dla populacji, w tym nowotwory. Do niedawna proces ten opierał się jedynie na kosztownych, powolnych i obarczonych wysokim ryzykiem niepowodzenia badaniach laboratoryjnych. Obecnie coraz większą rolę odgrywają rozwiązania oparte na sztucznej inteligencji. Wykorzystanie uczenia maszynowego we wstępnych fazach rozwoju pozwala znacząco przyśpieszyć wprowadzenie leku na rynek.

1.1. Sformułowanie problemu

Przestrzeń lekopodobna jest szacowana na 10^{60} [6] związków chemicznych. Zbiór ten jest zbyt duży, aby przetestować nawet ułamek możliwych rozwiązań dla danego problemu, dlatego trwają prace nad narzędziami, które usprawnią ten proces. Leki małocząsteczkowe to związki chemiczne, które wchodzą w reakcje z wybranymi białkami i modyfikują maszynerię białkową. Cząsteczka aktywna jest inhibitorem, który uniemożliwia danemu celowi zajście w reakcję z innymi związkami, dzięki czemu można kontrolować konkretne procesy.

Poszukiwanie nowych leków sprowadza się zatem do znalezienia cząsteczek o jak najsilniejszym powinowactwie do danego obiektu. Metody takie jak HTS (High-Throughput Screening) czy FBDD (Fragment-based drug discovery) zyskują coraz większe znaczenie. [10]

Ostatnie postępy w rozwoju metod uczenia maszynowego stwarzają wiele nowych możliwości i zastosowań dla modeli AI, w tym GNN. GNN dobrze wpasowuje się w ten problem, ponieważ grafowa reprezentacja cząsteczki lepiej zachowuje informację o strukturze przestrzennej i relacjach między atomami niż inne proste formaty, które sprowadzają skomplikowaną strukturę 3D cząsteczki do postaci listy cech. [12]

Dodatkowym wyzwaniem w bioinformatyce jest drastyczne niezbalansowanie danych (liczba cząsteczek aktywnych jest znikoma w porównaniu do nieaktywnych), co utrudnia skuteczne uczenie modeli.

W związku z powyższym sformułowano następujący problem badawczy:
Czy zastosowanie architektury Grafowych Sieci Neuronowych (GNN), uwzględniającej topologię cząsteczki, pozwala na skuteczną predykcję aktywności inhibitorów białka BRD4?

Moja praca ma na celu odpowiedzieć na problem kosztownego i czasochłonnego przeszukiwania ogromnej przestrzeni chemicznej.

1.2. Znaczenie białka BRD4 w onkologii

Białko BRD4 (bromodomain containing 4), z uwagi na to, że odpowiada m.in. za transkrypcję dna i powielanie komórek, jest istotnym i intensywnie badanym celem terapeutycznym w onkologii. Jego inhibicja pozwala zahamować proces replikacji komórek nowotworowych.

Ponadto, białko to posiada dobrze zdefiniowaną, głęboką kieszeń wiążącą, co czyni je jeszcze bardziej atrakcyjnym celem w rozwoju leków. Znalezienie ligand o silnym powinowactwie dla tego celu, pozwoli także na zminimalizowanie efektów ubocznych powstałych przez wiązania z innymi cząsteczkami o podobnej strukturze chemicznej do białka BRD4.

1.3. Cel pracy

Celem pracy jest zaprojektowanie, implementacja i walidacja potoku przetwarzania danych oraz modelu opartego na architekturze grafowych sieci neuronowych (GNN) w celu klasyfikacji małych cząsteczek pod kątem zdolności do swoistego wiązania z białkiem BRD4.

Do największych wyzwań należą silne niezbalansowanie danych oraz problem generalizacji modelu.

W zakres prac wchodzą:

- Akwizycja i filtracja danych z bazy BELKA.
- Implementacja konwersji cząsteczek z formatu SMILES do postaci grafowej
- Implementacja i trening modelu klasyfikacyjnego
- Walidacja modelu z uwzględnieniem oceny generalizacji.
- Analiza wyników oraz opracowanie wniosków z badań.

1.4. Struktura pracy

Praca dzieli się na 7 rozdziałów. Rozdział 2 przedstawia tło teoretyczne. Opis wykorzystanych metod w ekstrakcji danych i uczeniu modelu został przedstawiony w 3. Rozdział **??** opisuje projekt systemu i implementację potoku danych.

2. Przegląd Literatury

Zarys:

2.1. Wirtualny Screening

[11]

- Metody bazujące na strukturze białka
- Metody bazujące na strukturze ligandu
- Farmakofory

2.2. Zastosowanie GNN w biologii i medycynie [13]

2.3. Porównanie formatów cząsteczek: SMILES, Fingerprints, Grafy

2.4. Rodzaje GNN, dlaczego GAT?

3. Materiały i Metody

W tym rozdziale przedstawiono metody badawcze oraz narzędzia wykorzystane w trakcie uczenia. Na początku opisano, dlaczego zrezygnowano ze zbioru danych treningowych BindingDB na rzecz bazy BELKA oraz jak przygotowao dane do treningu modelu. Dalej zdefiniowano architekturę modelu predykcyjnego oraz wyjaśniono dobór metryk walidacyjnych, które najlepiej oceniają jego skuteczność.

3.1. Analiza danych

3.1.1. Charakterystyka zbioru danych

Baza BindingDB [5] jest największym publicznie dostępnym zbiorem treningowym, zawierającym dokładne wartości sił wiązań ligandów z białkami. Jest to niezwykle wartościowa baza, jednak powstała ona poprzez aglomerację wyników z wielu publikacji, co oznacza, że wartości powinowactwa (IC_{50}, K_i, K_d) zostały wyznaczone w różnych warunkach eksperymentalnych, a sam zbiór nie jest homogeniczny. Baza danych stwarza wiele możliwości rozwoju, jednak prawidłowe określenie progu odcięcia w celu binaryzacji etykiet wymaga dokładnej wiedzy z dziedziny farmakologii oraz chemii medycznej. W celu uniknięcia wprowadzenia szumu do danych treningowych ostatecznie zrezygnowano z bazy BindingDB na rzecz zbioru BELKA [1], który został udostępniony na platformie Kaggle na potrzeby konkursu pod tytułem "NeurIPS 2024 - Predict New Medicines with BELKA". Firma Leash Biosciences stworzyła go, wykorzystując technologię DEL (DNA-Encoded Libraries), która pozwoliła przetestować wszystkie cząsteczki w niezmiennych warunkach oraz zapewnić homogeniczność danych.

Dane zostały udostępnione w formacie csv i parquet. Zawieraja one numer identyfikujący, 3 bloki budulcowe w formacie smiles, pełną strukturę molekuły w formacie SMILES, nazwę białka docelowego oraz wartość binarną określającą czy dana para tworzy wiązanie. [8]

3.1.2. Ekstrakcja danych

Dla celu BRD4 wyekstrahowano 98 415 610 rekordów z pliku train.csv z bazy BELKA, z których ok. 0,466% to cząsteczki aktywne.

Z uwagi na rozmiar zbioru i ograniczone zasoby opisane w podpunkcie 3.2.1 zdecydowano się na przeprowadzenie undersamplingu. W tym procesie pozostawiono wszystkie rekordy klasy pozytywnej (456964 rekordów) oraz 1370892 przykładów z klasy negatywnej, zachowując proporcję 1:3.

W celu zaoszczędzenia pamięci usunięto z pliku niewykorzystane kolumny *buildingblock1_smiles* oraz *buildingblock2_smiles*, co pozwoliło zmiścić się w ramach platformy Kaggle, gdzie pliki wyjściowe nie mogą przekraczać 20 GB.

3.2. Architektura systemu

System zaimplementowano w języku Python (wersja 3.10) z wykorzystaniem bibliotek:

- **RDKit:** Do operacji chemicznych i ekstrakcji cech. [7]
- **PyTorch Geometric (PyG):** Do budowy i treningu sieci GNN. [3]
- **Scikit-learn:** Do walidacji i obliczania metryk. [9]
- **Polars** Do szybkiego wczytywania dużych plików CSV i Parquet

3.2.1. Środowisko obliczeniowe

Wszystkie obliczenia zostały wykonane na platformie Kaggle. W środowisku przysługuje 16 GB pamięci RAM, a czas sesji jest ograniczony do 12 godzin. Przysługuje również limit 30 godzin tygodniowo dla sesji wykorzystujących akceleratory GPU i TPU. Do treningów modelu wykorzystano kartę GPU P100, która w porównaniu do TPU v5e-8 jest prostsza w skonfigurowaniu i nie powoduje narzutu kodu do jej obsługi.

3.2.2. Architektura modelu

Zastosowano architekturę GAT opartą na warstwach atencji. Architektura: GAT (Graph Attention Network). Pozwala sieci "skupić uwagę"na ważnych atomach a zignorować tło.

- Wejście: Graf z cechami wierzchołków ($N \times 70$) i krawędzi ($E \times 6$) - atomy stanowią wierzchołki, a wiązania chemiczne krawędzie
- Warstwy ukryte: 3 warstwy `GATConv` z nieliniowością ReLU. [4]
- Agregacja: Global Mean (+Max?)Pooling (transformacja grafu do wektora).
- Klasyfikator: Warstwa liniowa (Linear) zwracająca logit prawdopodobieństwa wiązania.
- relu? sformalizowy zapis matematyczny mechanizmu atencji w GAT?

3.3. Metody walidacji

W celu poprawnej oceny zdolności modelu do generalizacji, zbiór danych został podzielony na trening i walidację [2]na podstawie informacji z kolumny *buildingblock3_smiles*.

Takie podejście w przeciwieństwie do podziału na bazie szkieletów Bemis-Murcko bądź podziału losowego, zapewnia rozłączność strukturalną między zbiorem treningowym a testowym oraz ochronę przed wyciekiem danych (data leakage).

Czy zagłębić się bardziej w Scaffold Split i dlaczego zawodzi? (stały rdzeń, Bemis-Murcko usuwa łańcuchy boczne)

Głównymi metrykami użytymi do walidacji modelu są precision i recall.

4. Implementacja

4.1. Potok przetwarzania danych

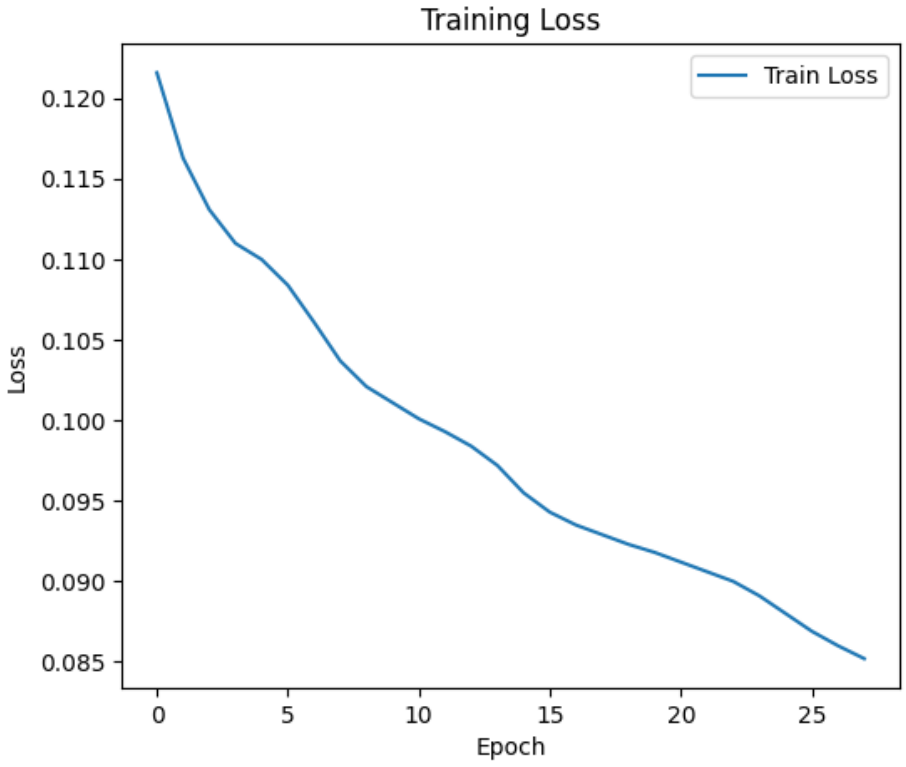
]

Opracowano autorską metodę konwersji `smiles_to_graph`. Proces ekstrakcji cech inspirowany był rozwiązaniami State-of-the-Art (m.in. z konkursu BELKA).

5. Badania eksperymentalne

Metryki oceny Ze względu na niezbalansowanie klas, jako główną metrykę decyzyjną przyjęto **Average Precision (AP)**, która lepiej niż AUC-ROC oddaje zdolność modelu do szeregowania cząsteczek aktywnych na szczycie listy rankingowej.

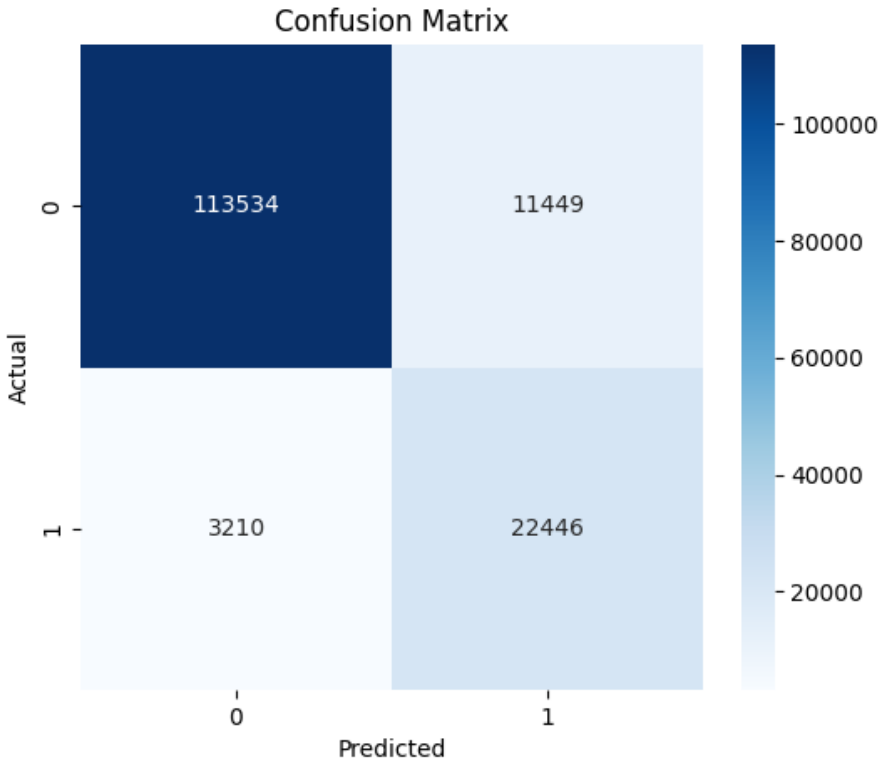
5.1. Wyniki eksperymentów



Rysunek 5.1: Krzywa strat (Loss Curve)

5.2. Dyskusja wyników

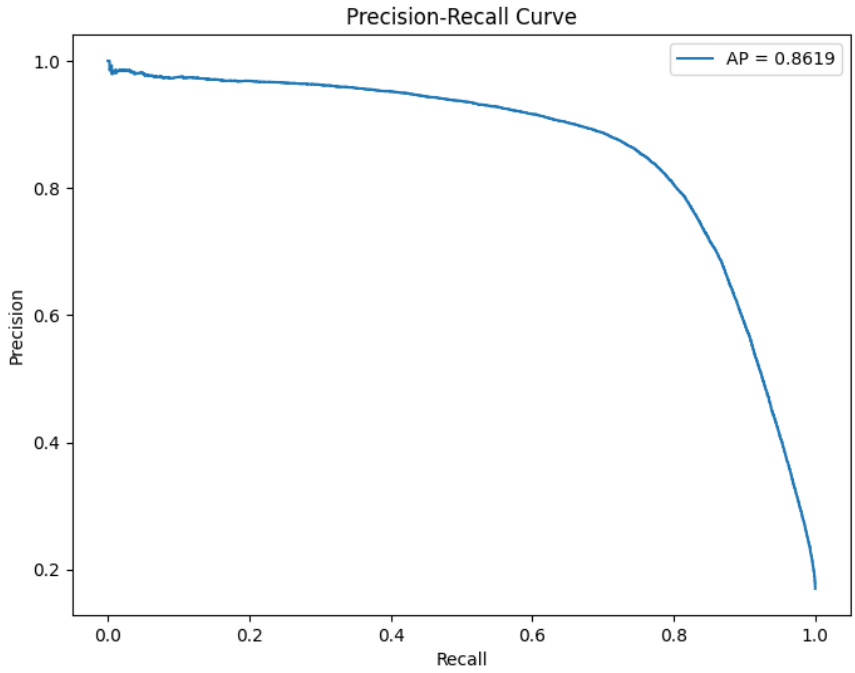
Zakończenie, podsumowuje najważniejsze wnioski, wskazuje obszar potencjalnego zastosowania pracy. Rezultaty pracy mają charakter poznawczy, mogą mieć charakter użytkowy.



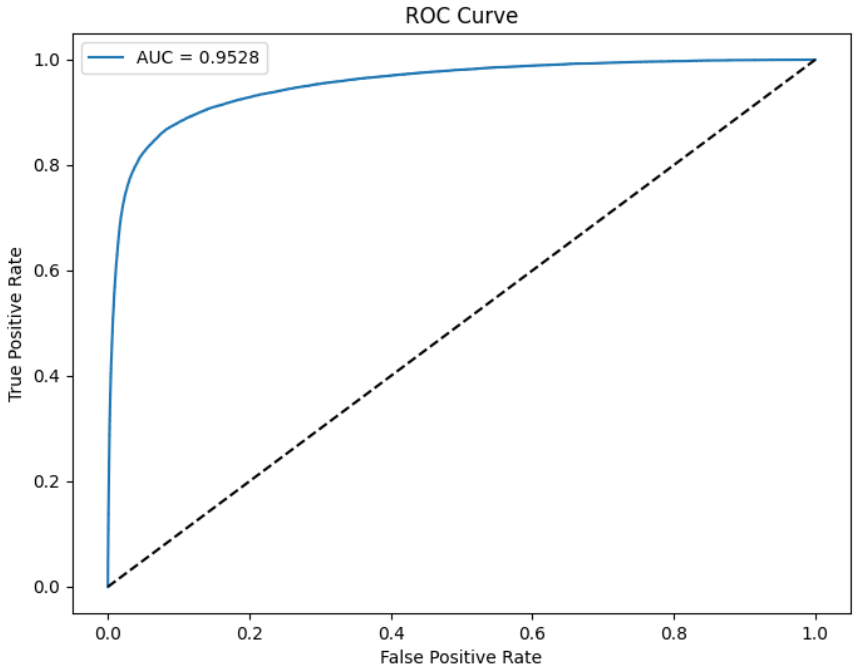
Rysunek 5.2: Tablica pomyłek (Confusion Matrix)

Należy dokonać analizy uzyskanych wyników. Rezultaty powinny charakteryzować się oryginalnością, a nawet w pewnym stopniu nowatorstwem. Praca zawiera (...). Zostało pokazane (...). Eksperymenty wykazały (...). Tu piszemy wnioski i obserwacje.

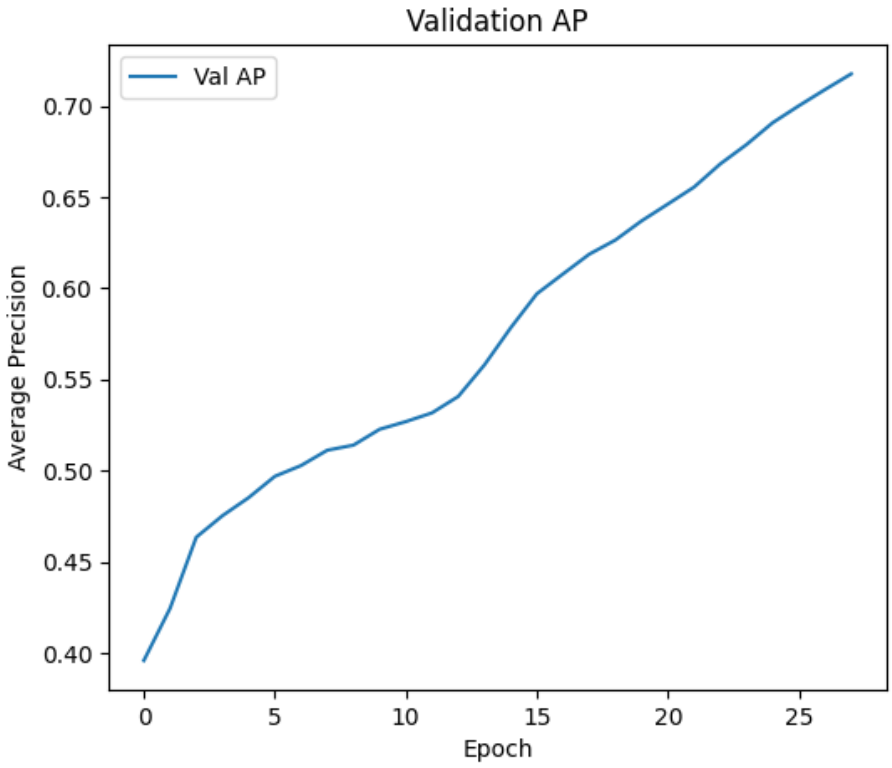
Interpretacja chemiczna wyników, analiza wpływu scaffold split na generalizację.



Rysunek 5.3: Krzywa PR (PR curve)



Rysunek 5.4: Krzywa ROC (ROC curve)



Rysunek 5.5: Krzywa VAL AP (Val AP curve)

6. Dyskusja

6.1. Interpretacja wyników w kontekście literaturowym

6.2. Analiza błędów.

6.3. Ograniczenia przyjętego podejścia.

- zrezygnowano z struktury 3D, zbyt złożona obliczeniowo
- undersampling klasy negatywnej nie tylko utracił dane, niepoprawna implementacja nadzorowanego undersamplingu mogła doprowadzić do wycieku danych?

7. Podsumowanie i wnioski

7.1. Osiągnięcia

Stworzenie SOTA modelu GAT.

7.2. Dalsze kierunki rozwoju

Złożoność problemu pozwala na rozwijanie projektu na wielu płaszczyznach. Jednakże najbardziej wpływowe zapewne będzie uruchomienie treningu na pełnym zbiorze treningowym. Wykorzystanie wszystkich rekordów

Ponadto uwzględnienie bloków budulcowych może poprawić generalizację algorytmu i

Bibliografia

[1] N. W. I. K. Q. Andrew Blevins, Brayden J Halverson. Belka: The big encoded library for chemical assessment.

[2] Encord. Training, validation, test split for machine learning datasets, n.d.

[3] M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[4] M. Fey and J. E. Lenssen. torch_geometric.nn.conv.GATConv – PyTorch Geometric 2.5.3 Documentation. https://pytorch-geometric.readthedocs.io/en/2.5.3/generated/torch_geometric.nn.conv.GATConv.html, 2024. Ostatni dostęp: 2024-12-06.

[5] 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 Research*, 44(D1):D1045–D1053, 2016.

[6] P. Kirkpatrick and C. Ellis. Chemical space. *Nature*, 432(7019):823–824, 2004.

[7] G. Landrum, P. Tosco, B. Kelley, et al. Rdkit: 2025.09.3, Sept. 2025.

[8] Leash Biosciences. Leash Bio - The Big Encoded Library for Chemical Assessment (BELKA). https://www.kaggle.com/competitions/leash-BELKA/data, 2024. Dostęp: 18.01.2026.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] D. E. Scott, A. R. Bayly, C. Abell, and J. Skidmore. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery*, 15(8):533–550, 2016.

[11] B. K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004.

[12] D. S. Wigh, J. M. Goodman, and A. A. Lapkin. A review of molecular representation in the age of machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1603, 2022.

[13] X.-M. Zhang, L. Liang, L. Liu, and M.-J. Tang. Graph neural networks and their current applications in bioinformatics. *Frontiers in genetics*, 12:690049, 2021.