Laboratory of Bioinformatics and Computational Genomics LB!GO
Faculty of Mathematics and Information Science, Warsaw University of Technology
ul. Koszykowa 75, 00-662 Warsaw, Poland

Laboratory of Functional and Structural Genomics LFSG
Centre of New Technologies, University of Warsaw
Banacha 2c Street, 02-097 Warsaw, Poland

mobile: +48504726203, e-mail: Dariusz.Plewczynski@pw.edu.pl, www: https://plewczynski-lab.org

Warszawa,
12.06.2023

Warszawa, 12/06/2023

Prof. dr hab. Dariusz Plewczyński
Laboratorium Bioinformatyki i Genomiki Obliczeniowej,
Wydział Matematyki i Nauk Informacyjnych,
Politechnika Warszawska
Laboratorium Genomiki Funkcjonalnej i Strukturalnej
Centrum Nowych Technologii
Uniwersytet Warszawski

RECENZJA rozprawy doktorskiej mgr Tomasza Danela

Metody uczenia głębokiego w naukach farmaceutycznych

wykonanej w Katedrze Uczenia Maszynowego
Instytutu Informatyki i Matematyki Komputerowej
na Wydziale Matematyki i Informatyki
Uniwersytetu Jagiellońskiego

pod kierunkiem promotora
doktora habilitowanego Igora Podolaka

zgłoszonej do Rady Dyscypliny Naukowej
Informatyka Techniczna i Telekomunikacja (ITT)
Uniwersytetu Jagiellońskiego

Przedstawiona mi do recenzji praca doktorska jest wynikiem udanego zastosowania metod uczenia maszynowego, a dokładniej uczenia głębokiego, w naukach farmaceutycznych, w tematyce badawczej rozwijanej w paradygmacie opartym na danych. Autorowi udało się połączyć wyniki metod eksperymentalnych z modelami uczenia statystycznego wykorzystując geometryczne uczenie głębokie w projektowaniu leków.

Grafowe sieci neuronowe (GNN) stają się coraz bardziej nieodłącznym elementem w przewidywaniu właściwości małocząsteczkowych związków chemicznych w dużej skali. Wynika to w dużej mierze z ich zdolności do skutecznego modelowania struktury molekularnej, co jest kluczowe w badaniach naukowych, przemysłowych i medycznych, szczególnie w kontekście projektowania leków. Celem przedstawionego mi doktoratu jest pokazanie roli i zastosowania GNN w kontekście kodowania informacji przestrzennej leków i białek w grafach molekularnych. Konformacja trójwymiarowa związku chemicznego odgrywa istotną rolę w określaniu kompatybilności inhibitora z białkiem docelowym, z którym ma się wiązać testowany, lub projektowany lek. Znalezienie odpowiedniego sposobu kodowania informacji przestrzennych w ramach GNN jest kluczowe dla poprawy skuteczności przewidywania własności molekularnych.

Doktorat skupia się na różnych aspektach wykorzystania GNN w kontekście projektowania leków, tj. chemoinformatyki. Przedstawiona w pracy dyskusja obejmuje różne modele komputerowe, rozpoczynając od tych, które skupiają się na reprezentacji grafowej związków chemicznych, a kończąc na modelach przewidujących ich wiązanie z białkiem oraz modelach generujących nowe związki małocząsteczkowe. Szczególny nacisk został położony na opracowanie mapowania geometrii dwu- i trójwymiarowej efektywnie reprezentującego cechy cząsteczek w kontekście uczenia głębokiego. Doktorat opiera się na różnych opublikowanych badaniach, w których prelegent brał udział, ilustrując różne strategie i techniki stosowane w procesie kodowania i przetwarzania informacji przestrzennej za pomocą GNN. Celem była poprawa wyników teoretycznych w przewidywaniu interakcji białek i związków małocząsteczkowych, oraz generowaniu nowych kandydatów na leki.

Praca doktorska składa się z cyklu dziewięciu prac opublikowanych w międzynarodowych i wysoko punktowanych czasopismach lub materiałach konferencyjnych. Jedna praca została opublikowana w źródle ocenianym na 200 punktów ministerialnych, cztery prace ukazały się w źródłach mających 140 punktów, cztery w 100 punktowych. W sumie mamy więc 9 manuskryptów, co

recenzent uważa za wynik wyjątkowy, nawet uwzględniając współautorski charakter prowadzonych badań. Publikacje konferencyjne ukazały się przy okazji konferencji klas CORE A, zaś dwa doniesienia w czasopismach. Dodatkowo doktorantowi udało się uczestniczyć w grantach badawczych, czy też kierować własnymi grantami.

Przesłany mi do recenzji doktorat składany jest w dyscyplinie informatyki technicznej i telekomunikacji (ITT), co w moim mniemaniu jest uprawnione, ponieważ doktorant opracował zaawansowane modele generatywne, sieci grafowe, uczenie głębokie i tym podobne, które to metody umożliwiają istotny postęp zarówno w informatyce, ale również zastosowaniach w naukach farmaceutycznych. Tego typu badania prowadzone są w dopiero od niedawna istniejącej ścieżce nauk cheminformatycznych, w odpowiedzi na wyzwania techniczne i informatyczne związane m.in. z projektowaniem leków, procedurami wirtualnego przesiewania. Wymaga to opracowania algorytmów specyficznych dla chemicznej charakterystyki wyzwań, co nie umniejsza skali trudności po stronie technicznej, przeciwnie – opracowanie przydatnych algorytmów informatycznych działających na danych rzeczywistych jest często dużo trudniejsze w mniemaniu recenzenta niż praca wyłącznie teoretyczna.

Przedmiotem mojej oceny, jest rozprawa doktorska, w tym przypadku zszywka dziewięciu prac współautorskich. Rozprawa w moim przekonaniu w pełni spełnia warunki określone w art. 187 ustawy z dnia 20 lipca 2004 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki (Dz. U. z 2014 r., poz. 665 ze zm.). Rozprawa ta, a raczej publikacje i dokonania naukowe doktoranta, prezentują oryginalność rozwiązanego unikalnego problemu naukowego, ogólną wiedzę teoretyczną, ale również aplikacyjną (specyficzną dla dziedziny cheminformatyki) kandydata w dyscyplinie informatyki technicznej, a także umiejętność prowadzenia pracy naukowej na wysokim poziomie.

Rozprawa doktorska mgr Tomasza Danela pt. pt. "*Metody uczenia głębokiego w naukach farmaceutycznych*" została przygotowana w Katedrze Uczenia Maszynowego Instytutu Informatyki i Matematyki Komputerowej na Wydziale Matematyki i Informatyki Uniwersytetu Jagiellońskiego pod kierunkiem dr hab. Igora Podolaka. Badania zostały zrealizowane w ramach stypendium *Descartes*, co umożliwiło Doktorantowi skorzystanie z unikalnej ścieżki rozwoju, wzbogacając tym samym zakres rozprawy doktorskiej. Ponadto, współpraca z firmą Ardigen Sp. z o.o. umieściła doktoranta w interdyscyplinarnym i międzynarodowym otoczeniu współpracy na pograniczu

nauki i biznesu, oraz dostarczyło mu unikalnych i przydatnych pojęć, narzędzi i technik, które zastosował w swojej pracy badawczej.

Autor prezentuje tezy swojej rozprawy doktorskiej na 22 stronach w autoreferacie, w którym podkreśla znaczenie wybranych metod bioinformatycznych oraz opracowanych narzędzi i algorytmów informatycznych. Zaproponowane przez autora metody i narzędzia są zgodne z rozwijanym w środowisku cheminformatycznym nurtem badawczym opierającym się na wykorzystaniu komputerów w projektowaniu leków, dodatkowo jednak wzbogacone zostały za pomocą algorytmów uczenia głębokiego. Doktorant zaprezentował zaawansowany poziom identyfikacji ciekawych białkowych celów terapeutycznych, przygotowania zestawów danych doświadczalnych, oraz rozwijania własnego oprogramowania bioinformatycznego. W ramach dostępnego czasu i nakładu pracy, dorobek naukowy doktoranta jest godny uwagi, a nawet wyróżniający i nie budzi żadnych wątpliwości recenzenta.

W jednostronicowym wstępie Doktorant prezentuje skrótowo swój dorobek. Praca doktorska koncentruje się na opracowywaniu metod uczenia głębokiego z naciskiem na zastosowanie w problemach farmaceutycznych, dążąc do skrócenia procesu projektowania leków i redukcji kosztów. Praca zwraca szczególną uwagę na modele grafowe w zastosowaniu do przetwarzania danych medycznych, biologicznych i chemicznych, które ze względu na swoją specyfikę wymagają zindywidualizowanego podejścia. Związki chemiczne reprezentowane są tutaj jako grafy z cechami wierzchołków i krawędzi odpowiadającymi atomom i wiązaniom cząsteczki chemicznej. Pierwszych sześć publikacji cyklu dotyczy sieci grafowych i modelowania związków chemicznych, wprowadza nową architekturę sieci, omawia reprezentację związków, generowanie nowych związków, zastosowanie sieci w kontroli metabolizmu, interpretowalną sieć grafową i optymalizację związków chemicznych. Dodatkowo, doktorat omawia również w kolejnych trzech pracach zastosowanie sieci neuronowych do przetwarzania danych z obrazowania medycznego i w farmacji, na przykładzie analizy zdjęć.

Następnie zaprezentowane są podstawowe pojęcia i stosowana przez Doktoranta notacja. Są to: graf nieskierowany ocechowany, splotowa/konwolucyjna sieć neuronowa, grafowe splotowe sieci neuronowe i model generatywny. Graf nieskierowany ocechowany to struktura, w której wierzchołki odpowiadają atomom, a krawędzie wiązaniom związku chemicznego, zaś cechami wierzchołków i krawędzi są deskryptory atomów i wiązań. Splotowa sieć neuronowa to architektura używana do przetwarzania

obrazów, reprezentowanych jako 3-wymiarowe tensory, i wykorzystuje pojęcie sąsiedztwa do filtrowania obrazów. Grafowe splotowe sieci neuronowe działają analogicznie, z tym, że pojęcie sąsiedztwa definiuje się na podstawie sąsiedztwa wierzchołków w grafie. Model generatywny uczy się rozkładu prawdopodobieństwa danych z celem generowania nowych przykładów, które pasują do rozkładu danych wejściowych, w tym przypadku związki chemiczne, które spełniają dodatkowe założenia związane z celem ich projektowania.

Następnie zaprezentowany jest cykl prac obejmujących dziewięć pozycji wraz z jednozdaniowymi opisami wkładu autorskiego:

- **Tomasz Danel,** Przemysław Spurek, Jacek Tabor, Marek Śmieja, Łukasz Struski, Agnieszka Słowik, Łukasz Maziarka. "*Spatial graph convolutional networks*". International Conference on Neural Information Processing. Springer. 2020, pp. 668–675. CORE A, MNiSW 140 pkt.
- Agnieszka Pocha, **Tomasz Danel,** Sabina Podlewska, Jacek Tabor, Łukasz Maziarka. "*Comparison of atom representations in graph neural networks for molecular property prediction*". 2021 International Joint Conference on Neural Networks (IJCNN). IEEE. 2021, pp. 1–8. CORE A, MNiSW 140 pkt.
- **Tomasz Danel,** Jan Łęski, Sabina Podlewska, Igor Podolak. "*Docking-based generative approaches in the service of finding new drug candidates*". Drug discovery today (2023). IF 8.369, MNiSW 200 pkt.
- **Tomasz Danel,** Agnieszka Wojtuch, Sabina Podlewska. "*Generation of new inhibitors of selected cytochrome P450 subtypes – in silico study*". Computational and Structural Biotechnology Journal (2022). IF 6.155, MNiSW 100 pkt.
- Dawid Rymarczyk, Daniel Dobrowolski, **Tomasz Danel**. "*ProGReST: Prototypical Graph Regression Soft Trees for Molecular Property Prediction*". SIAM International Conference on Data Mining (SDM). 2023. CORE A, MNiSW 140 pkt.
- Łukasz Maziarka, Agnieszka Pocha, Jan Kaczmarczyk, Krzysztof Rataj, **Tomasz Danel,** Michał Warchoł."*Mol-CycleGAN: a generative model for molecular optimization*". Journal of Cheminformatics (2020). IF 8.489, MNiSW 100 pkt.
- **Tomasz Danel,** Marek Śmieja, Łukasz Struski, Przemysław Spurek, Łukasz Maziarka. "*Processing of incomplete images by (graph) convolutional neural networks*". International

Conference on Neural Information Processing. Springer. 2020. CORE A, MNiSW 140 pkt.

- Agnieszka Galanty, **Tomasz Danel,** Michał Węgrzyn, Irma Podolak, Igor Podolak. "*Deep convolutional neural network for preliminary in-field classification of lichen species*". Biosystems engineering (2021). IF 5.002, MNiSW 100 pkt.
- Łukasz Struski, **Tomasz Danel**, Marek Śmieja, Jacek Tabor, and Bartosz Zieliński. "*SONGs: Self-Organizing Neural Graphs*". 2023 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE. 2023. CORE A, MNiSW 140 pkt.

Praca pierwsza „Przestrzenne grafowe sieci splotowe” przedstawia nową architekturę grafowej sieci neuronowej, przestrzenną grafową sieć splotową (Spatial Graph Convolutional Network, SGCN), której celem jest uogólnienie klasycznej sieci konwolucyjnej oraz sieci grafowej. W przeciwieństwie do klasycznych sieci grafowych, które nie rozróżniają sąsiednich wierzchołków, SGCN wprowadza do grafu pojęcie przestrzeni d-wymiarowej, umieszczając w niej poszczególne wierzchołki. Wyniki eksperymentów wskazują na wyższą skuteczność SGCN, z augmentacjami, w porównaniu do innych metod grafowych na wybranych chemicznych zbiorach danych. W kontekście związków chemicznych, SGCN wykorzystuje informacje o pozycji atomów, co przynosi dodatkową wartość. Wykorzystanie tej architektury pozwala na analizę związków chemicznych, które można łatwo umieścić w przestrzeni 3-wymiarowej, umożliwiając w modelowaniu chemicznym użycie konformacji molekularnych.

Praca druga „Porównanie reprezentacji atomów w grafowych sieciach neuronowych przewidujących własności związków chemicznych” skupia się na analizie wpływu różnych reprezentacji cech atomów na wyniki przewidywań generowanych przez grafowe sieci neuronowe dla związków chemicznych. Pomimo skupiania się na udoskonaleniu architektury sieci, wybór użytych cech atomowych jest często pomijany, mimo że może mieć równie duży wpływ na jakość przewidywań. Przeprowadzono systematyczne porównanie różnych zestawów cech na kilku chemicznych zbiorach danych, z których wynika, że optymalne cechy mogą różnić się w zależności od problemu i sposobu podziału danych. Ustalono, że nie zawsze użycie wszystkich możliwych cech prowadzi do najlepszych wyników. Zauważono, że dla niektórych zbiorów danych usunięcie pewnych cech (np. ładunku formalnego i aromatyczności) może poprawić wyniki, a dodanie informacji o sąsiadach ciężkich i wodorach przynosi największą poprawę. Wyniki wskazują na istotność odpowiedniego doboru cech atomów dla skuteczności sieci grafowej.

Praca trzecia „Podejścia bazujące na dokowaniu w służbie odkrywania nowych kandydatów na lek” omawia wykorzystanie dokowania, narzędzia komputerowego służącego do oceny kandydatów, w procesie odkrywania nowych leków. Dokowanie, tworząc ranking związków chemicznych pod względem ich przewidywanego powinowactwa do celu (zwykle białka) oraz potencjalnej lokalizacji w kieszeni wiążącej białka, jest kluczowym etapem wirtualnego screeningu. Artykuł opisuje najnowsze metody generatywne integrujące dokowanie molekularne, które służą do tworzenia związków o zwiększonym powinowactwie do swojego celu. Informacje uzyskane z dokowania stanowią wartościowe dane, które mogą ukierunkować proces generatywny. Modele generatywne są uczone w sposób częściowo nadzorowany, gdzie etykiety generowane są za pomocą narzędzi do dokowania. Autorzy proponują taksonomię modeli generatywnych korzystających z dokowania, podzieloną na metody wykorzystujące kodowanie kieszeni wiążącej i te korzystające tylko z danych specyficznych dla danego celu biologicznego. Metody te mogą wykorzystywać różne techniki, takie jak algorytmy genetyczne, uczenie ze wzmocnieniem, czy optymalizacja przestrzeni ukrytej.

Praca czwarta „Generowanie nowych inhibitorów wybranych podtypów cytochromu P450 - studium in silico” prezentuje badania przeprowadzone dla wybranych podtypów cytochromu P450, enzymu odpowiedzialnego za metabolizm związków chemicznych w organizmie. Celem publikacji było znalezienie bliskich analogów istniejących inhibitorów tych enzymów i porównanie ich powinowactwa w dokowaniu molekularnym. Analiza obejmowała generowanie analogów z użyciem 15 małych grup chemicznych, które były kombinatorycznie dołączane do wszystkich możliwych atomów w pierścieniach związków pierwotnych. Przy pomocy sieci grafowej przewidziano zmiany wartości funkcji dokowania, formułując problem jako regresję wielowartościową na wierzchołkach grafu. W wyniku eksperymentu zaproponowano nowy protokół generowania inhibitorów wybranych podtypów CYP, demonstrując skuteczność grafowej sieci neuronowej w proponowaniu podstawień związku, które poprawiają wyniki dokowania. Cała implementacja, wraz z bazą danych zadokowanych związków chemicznych, jest dostępna publicznie, a wyniki dokowania części bazy danych można zobaczyć w stworzonym przez autorów narzędziu online.

Praca piąta „ProGReST: prototypowe grafowe miękkie drzewa regre- syjne do przewidywania własności cząsteczek” przedstawia nowy interpretowalny model grafowy, ProGReST, wykorzystywany do przewidywania własności

cząsteczek, co ma istotne znaczenie w procesie projektowania leków. Model opiera się na koncepcji prototypów - zestawie cech lub fragmentów podobnych do reprezentantów w zbiorze treningowym, co w przypadku małych związków chemicznych może oznaczać konkretne grupy funkcyjne. ProGReST bazuje na grafowej sieci neuronowej, tworząc reprezentację ukrytą cząsteczki, składającą się z wektorów cech dla każdego z jej atomów. W swojej konstrukcji ProGReST jest miękkim drzewem decyzyjnym, które w każdym wierzchołku porównuje reprezentację atomów cząsteczki wejściowej z trenowaną częścią prototypową. W praktyce oznacza to, że możliwe jest kierowanie decyzji do obu potomków jednocześnie, a wartość własności związku oblicza się jako kombinację liniową prawdopodobieństw w liściach. Eksperymenty na pięciu zbiorach własności chemicznych wykazały bardzo wysoką skuteczność modelu, pokonując w większości zadań model odniesienia - prostą sieć grafową.

Praca szósta „Mol-CycleGAN: model generatywny służący do optymalizacji związków chemicznych” prezentuje Mol-CycleGAN, innowacyjny model generatywny wykorzystywany do optymalizacji związków chemicznych, co ma kluczowe znaczenie w procesie odkrywania nowych leków. Mol-CycleGAN to pierwszy model stosujący adaptację domen do problemu optymalizacji związków chemicznych, bazujący na architekturze JT-VAE - autoenkodera wariacyjnego kodującego molekuły w formie drzew podstruktur chemicznych. Autorzy traktują problem optymalizacji cząsteczek jako zadanie optymalizacji reprezentacji danych, z zastosowaniem dwóch autoenkoderów do generowania związków z jednej domeny na podstawie związków z drugiej domeny i odwrotnie. Skuteczność modelu sprawdzono na własności *penalized log P*, łączącej przewidywaną komputerowo lipofilowość związku i łatwość jego syntezy, a także na optymalizacjach strukturalnych, np. podmiany bioizosterycznej i aktywności biologicznej. Model wykazał zdolność do nauki cech strukturalnych obu domen i skuteczną optymalizację związków.

Praca siódma „Przetwarzanie niekompletnych obrazów przy pomocy (grafowych) sieci splotowych” opisuje nowatorskie podejście do przetwarzania niekompletnych obrazów przy użyciu grafowych sieci neuronowych (SGCN). Autorzy skupiają się na klasyfikacji oraz rekonstrukcji obrazów z brakującymi regionami, reprezentując obrazy jako grafy, gdzie wierzchołki to piksele, a krawędzie łączą sąsiadujące piksele. SGCN, działając jako enkoder w architekturze autoenkodera, pozwala na pominięcie brakujących pikseli podczas przetwarzania, utrzymując przy tym wysoką dokładność. Zasugerowane rozwiązanie pozwala na optymalizację procesu bez potrzeby sztucznego uzupełniania brakujących informacji. Dekodowanie odbywa się za pomocą sieci

dekonwolucyjnej, stopniowo rekonstruującej pełny obraz. Testy przeprowadzono na zestawach danych MNIST i SVHN, osiągając lepsze wyniki w klasyfikacji oraz rekonstrukcji obrazów niż przy zastosowaniu tradycyjnych sieci konwolucyjnych na sztucznie uzupełnionych obrazach.

Praca ósma „Głębokie sieci splotowe służące wstępnej terenowej klasyfikacji gatunków porostów" przedstawia innowacyjne podejście do klasyfikacji gatunków porostów przy użyciu głębokich sieci splotowych, z naciskiem na zastosowanie mobilne. Autorzy skupiają się na rodzaju *Cladoniaceae*, używając sieci splotowych do klasyfikacji zdjęć porostów do jednego z 12 gatunków. Zastosowane architektury sieci są dostosowane do urządzeń o ograniczonych możliwościach obliczeniowych, umożliwiając instalację na smartfonach badaczy terenowych lub kamerach dronów. Eksperymenty przeprowadzono na własnym zbiorze 1164 zdjęć porostów, pobranych z Internetu i ręcznie odfiltrowanych. Wykorzystano dwie mobilne architektury sieci - MobileNet v2 oraz SqueezeNet, a także wektory Fishera. Zastosowano procedurę augmentacji danych z uwagi na małą ilość dostępnych zdjęć. Najlepsze rezultaty - 61% dokładności w klasyfikacji 12-klasowej - osiągnięto za pomocą zmodyfikowanego modelu SqueezeNet.

Praca dziewiąta i ostatnia „SONG: samoorganizujące się grafy neuronowe" prezentuje nową architekturę sieci neuronowych - samoorganizujące się grafy neuronowe (SONG), która łączy zalety drzew decyzyjnych z elastycznością struktury grafu decyzyjnego. SONG jest modelem klasyfikacyjnym, który operuje na wektorowej reprezentacji danych, dokonując sekwencji decyzji binarnych w grafie decyzyjnym. Graf jest sterowany przez dwie macierze przejść, a decyzja o wyborze macierzy w danym wierzchołku jest zależna od danych wejściowych. Model jest w pełni różniczkowalny, co pozwala na trenowanie go z innymi architekturami sieci neuronowych. W celu poprawy wyników, zastosowano dodatkowo regularyzację oraz stochastyczną operację Gumbel softmax, zamiast tradycyjnego softmaxu. Eksperymenty pokazały, że SONG osiąga dokładność porównywalną do innych metod wykorzystujących drzewa decyzyjne, przy mniejszej liczbie wierzchołków decyzyjnych. Ponadto, dowiedziono teoretycznie i empirycznie, że nauczony SONG zbiega do rzadkiego acyklicznego grafu binarnego.

Poniżej postaram się w skrócie podsumować kluczowe osiągnięcia badawcze doktoranta, które zostały przedstawione przez autora i streszczone przez recenzenta.

- nowa architektura sieci neuronowej, przestrzenna grafowa sieć splotowa (SGCN), która łączy cechy klasycznej sieci konwolucyjnej i sieci grafowej, uwzględniając położenie wierzchołków w przestrzeni d-wymiarowej. Wyniki badań pokazują, że SGCN z augmentacjami daje lepsze wyniki niż inne metody grafowe na wybranych zbiorach danych chemicznych, sugerując, że efektywne wykorzystanie pozycji atomów dostarcza dodatkowej informacji. Dodatkowo, dodanie informacji o pozycji do cech atomowych w klasycznej sieci grafowej nie przynosi wyników porównywalnych z tymi uzyskiwanymi przez SGCN.
- pokazanie, że wybór odpowiedniego zestawu cech atomu może mieć równie duży wpływ na wyniki przewidywań modelu co dobór odpowiedniej architektury sieci neuronowej. Wyniki wskazują, że optymalne cechy zależą od konkretnego problemu i sposobu podziału danych. Użycie wszystkich możliwych cech nie zawsze prowadzi do najlepszych wyników modelu, co może być spowodowane przeuczeniem. Usunięcie niektórych cech, takich jak ładunek formalny i aromatyczność, może poprawić wyniki przewidywań w niektórych zestawach danych, podczas gdy dodanie informacji o sąsiadach ciężkich i wodorach często znacząco poprawia wyniki.
- nowe metody generatywne, które integrują dokowanie molekularne w celu tworzenia związków o polepszonym powinowactwie do swojego celu. W pracy zaproponowano nową taksonomię modeli generatywnych wykorzystujących dokowanie, podzieloną na te które korzystają z kodowania kieszeni wiążącej i te, które wykorzystują wyłącznie dane dokowania specyficzne dla danego celu biologicznego. Autorzy prezentują różne strategie, takie jak uczenie ze wzmocnieniem, algorytmy genetyczne, optymalizacja w przestrzeni ukrytej i metody filtrujące wygenerowane związki. Nowa klasyfikacja ma na celu umożliwić naukowcom lepsze zrozumienie i wykorzystanie takich algorytmów w przyszłych badaniach nad modelami generatywnymi.
- nowy protokół generowania inhibitorów wybranych cytochromów P450 za pomocą studium *in silico*. Wykorzystując sieć neuronową typu grafowego, autorzy skutecznie proponują modyfikacje związku, które poprawiają wyniki dokowania molekularnego. To podejście pozwala na zawężenie przeszukiwanej przestrzeni chemicznej i stanowi alternatywę dla generowania wszystkich kombinatorycznych pochodnych. Dodatkowo, autorzy udostępnili publicznie swoją implementację generatora związków, sieci grafowej oraz modeli wyjaśniających wraz z bazą danych dużej liczby zadokowanych związków chemicznych.

Przygotowali także narzędzie online umożliwiające wizualizację wyników dokowania części bazy danych.

- nowy interpretowalny model grafowy do przewidywania własności cząsteczek, nazwany ProGReST, który opiera się na koncepcji prototypów. Grafowa sieć neuronowa tworzy ukrytą reprezentację cząsteczki, jako miękkie drzewo decyzyjne, które w każdym węźle porównuje reprezentację atomów cząsteczki wejściowej z trenowalnym fragmentem prototypowym. Eksperymenty przeprowadzone na pięciu zbiorach własności chemicznych wykazały, że ProGReST przewyższa w każdym przypadku model odniesienia, jakim jest prosta sieć grafowa, a gdy siecią grafową jest RMAT, pokonał inne modele w czterech na pięć zadaniach, jednocześnie dodając interpretowalność przewidywań modelu.
- nowy model generatywny do optymalizacji związków chemicznych, nazwany Mol-CycleGAN, oparty na autoenkoderze wariacyjnym JT-VAE. Model ten wykorzystuje koncepcję adaptacji domen do problemu optymalizacji molekularnej, gdzie domenami są zbiory związków o niskich i wysokich wartościach wybranych własności chemicznych. Model składa się z dwóch autoenkoderów służących do generowania związków z jednej domeny na podstawie związków z drugiej domeny i odwrotnie. Funkcja kosztu CycleGAN-a zapewnia, że wygenerowane związki są wysoko podobne do rozpatrywanej domeny, a jednocześnie zachowuje duże podobieństwo między związkiem wejściowym a zmodyfikowanym.
- Mol-CycleGAN może mieć znaczący wpływ na proces odkrywania leków, umożliwiając skuteczne modyfikacje struktury molekularnej w celu poprawy pożądanych właściwości z minimalnymi zmianami struktury wyjściowej.
- opracowanie efektywnego systemu do klasyfikacji oraz uzupełniania niekompletnych obrazów z dowolnymi brakującymi regionami, wykorzystującego do tego celu grafowe sieci konwolucyjne (SGCN) w architekturze autoenkodera. Propozycja użycia SGCN w tego typu zadaniach to istotne rozwiązanie, które może znaleźć zastosowanie w wielu obszarach, szczególnie w medycynie, gdzie często spotykane są obrazy o wysokiej rozdzielczości i z brakującymi lub niewyraźnymi fragmentami.
- sposób reprezentacji obrazów jako grafów, w których wierzchołki to piksele, a nieskierowane krawędzie łączą sąsiednie piksele, pozwolił na pominięcie brakujących pikseli podczas przetwarzania przez sieć. W przeprowadzonych eksperymentach na zbiorach danych MNIST i SVHN, system wykazał lepsze wyniki w zadaniu klasyfikacji niż analogiczne sieci konwolucyjne działające na sztucznie uzupełnionych obrazach. Również

w zadaniu rekonstrukcji obrazów, regiony uzupełnione przez model wyraźnie przypominały oryginalne obrazy, a kształty były ostre i dobrze wpasowywały się w otaczającą je ramkę dostępnych pikseli.

- opracowanie skutecznego systemu klasyfikacji gatunków porostów za pomocą głębokich sieci splotowych, dostosowanych do urządzeń o ograniczonej mocy obliczeniowej i dostępnej pamięci, jak telefony komórkowe czy drony. Zastosowane metody uczą się na zdjęciach porostów pobranych z internetu i poddanych augmentacji, co pozwoliło na osiągnięcie 61% dokładności w klasyfikacji 12 gatunków porostów przy stosunkowo niewielkim zbiorze danych. Jest to pierwsze rozwiązanie tego typu skierowane na klasyfikację gatunków porostów, co ma duże znaczenie dla obszarów takich jak farmakognozja.
- stworzenie nowej architektury sieci neuronowych - samoorganizujących się grafów neuronowych (SONG), które łączą zalety drzew decyzyjnych z elastycznością struktury grafu decyzyjnego. Przeprowadzone eksperymenty pokazały, że SONG potrafi osiągnąć dokładność porównywalną do innych metod wykorzystujących drzewa decyzyjne, jednakże przy użyciu mniejszej liczby wierzchołków decyzyjnych. Dodatkowo, teoretycznie i empirycznie dowiedziono, że po pełnym procesie nauki, SONG zbiega do rzadkiego acyklicznego grafu binarnego. Ta unikalna architektura sieci neuronowych okazała się skuteczna w rozwiązaniu różnorodnych problemów klasyfikacyjnych.

Na koniec przedstawiony jest życiorys doktoranta oraz Bibliografia. Kandydat opublikował dodatkowo 6 prac niezależnych od tych już uwzględnionych w cyklu. Był wykonawcą w grancie OPUS „*Głębokie samoorganizujące się grafy neuronowe*”, oraz kierował grantem PRELUDIUM z NCN pod tytułem „*Połączenie symulacji molekularnej i uczenia głębokiego w projektowaniu leków de novo*”. Podobnie był wykonawcą w jednym grancie IDUB pt. „*Stabilność reprezentacji ukrytej molekuł w autoenkoderach*”, oraz kierował drugim grantem typu Uczelnia Badawcza „*Grywalizacja procesu projektowania leków*”. Na koniec należy wspomnieć o prowadzeniu przez Doktoranta zajęć dydaktycznych „*Uczenie maszynowe w projektowaniu leków*” od roku akademickiego 2021/2022.

Wnioski końcowe pracy doktorskiej obejmują:

- Kompleksowy projekt badawczy dotyczący uczenia głębokiego w cheminformatyce, farmacji i rozpoznawaniu obrazu.

- Kluczowe odkrycia, w tym opracowanie wydajnych i odpornych algorytmów uczenia głębokiego oraz polepszenie tradycyjnych metod uczenia maszynowego.

- Uogólnienie wniosków na podstawie obserwowanych wyników doświadczalnych przy użyciu uczenia statystycznego.

- Opracowanie zestawów danych do treningów uczenia maszynowego.

- Podsumowując, praca doktorska przyniosła solidne wyniki, ważne zbiory danych oraz nowatorskie algorytmy.

Pytania do doktoranta:

Opracowane metody prezentują interesujące i innowacyjne koncepcje, jednak nie są wolne od potencjalnych słabości i obszarów, które mogą wymagać dalszych badań. Prosiłbym o skomentowanie w trakcie obrony następujących ogólnych kwestii dla wybranych przykładów z całości cyklu publikacji:

- *Skomplikowany proces treningu*. Przykładowo metoda SONG wykorzystuje zaawansowane techniki regularyzacyjne i strategie optymalizacji, takie jak operacja Gumbel softmax, które mogą komplikować proces treningu i wymagać znacznej mocy obliczeniowej, którą warto oszacować porównując z osiąganym zyskiem.
- *Porównanie z innymi metodami*: Autorzy osiągają dokładność porównywalną do innych metod (np. wykorzystujących drzewa decyzyjne), ale eksplorują bezpośredniego porównania np. modelu SONG z innymi, bardziej zaawansowanymi technikami sieci neuronowych (BERT).
- *Zastosowanie do różnych typów danych*: Autorzy opisują zastosowania opracowanych metod do różnych problemów klasyfikacyjnych, jednak nie zawsze prezentują szczegółowe wyniki dla każdego z tych przypadków. Ciekawe jest skomentowanie zagadnienia uniwersalności i efektywności proponowanych metod w różnych scenariuszach i ich transferowalności.
- *Skrótowość i niejasności w opisie metod*: Opis metod jest skomplikowany i niejasny w niektórych miejscach, bardzo skrótowy - co może czasami utrudniać pełne zrozumienie proponowanego podejścia, szczególnie dla osób nieznajomych z obszarem uczenia głębokiego i typami danych.
- *Wyniki teoretyczne o zbieżności (np. SONG do rzadkiego acyklicznego grafu binarnego)*: teoretyczne i empiryczne dowody są opisywane, jednak w rzeczywistych warunkach ten proces może nie zawsze być gwarantowany lub może wymagać bardzo długiego czasu treningu.

Wnioski końcowe

W podsumowaniu oceny rozprawy doktorskiej pana mgr Tomasza Danela pt. "*Metody uczenia głębokiego w naukach farmaceutycznych*", pragnę wyrazić moje wysokie uznanie dla przedstawionej pracy. Suma wskaźników impact factor wszystkich publikacji wynosi 28.015, a suma punktów ministerialnych w Polsce wynosi 1200. Jest to wynik wyjątkowy, nawet jeśli uwzględnimy współautorskość zamieszczonych publikacji.

Biorąc pod uwagę czytelność i wartość naukową rozprawy doktorskiej, udane połączenie starannie opisanych narzędzi uczenia głebokiego, a także farmaceutycznie istotnych eksperymentów i fundamentalnych pytań badawczych związanych z cheminformatyką, uważam rozprawę doktorską pana mgr Tomasza Danela za znaczący wkład w dyscyplinę Informatyki Technicznej i Telekomunikacji.

Rozprawa doktorska spełnia warunki określone w Art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478, 619, 1630). Ponadto uważam, że rozprawa ta przewyższa powszechne i ustawowe wymagania stawiane rozprawom doktorskim, stanowi oryginalne rozwiązanie problemu naukowego, wykazuje ogólną wiedzę informatyczną oraz chemiczną kandydata w zakresie farmacji i technik bioinformatycznych, oraz demonstruje zdolność do samodzielnego prowadzenia pracy naukowej.

W związku z powyższym, mam przyjemność przedłożyć Radzie Dyscypliny Naukowej Informatyki Technicznej i Telekomunikacji Uniwersytetu Jagiellońskiego wniosek o dopuszczenie pana mgr Tomasza Danela do dalszych etapów przewodu doktorskiego.

Ponadto, zważywszy na wysoki poziom merytoryczny i obszerność rozprawy, jej staranne przygotowanie oraz klarowny sposób prezentacji tematyki badawczej, metodyki i wyników, wnioskuję o wyróżnienie rozprawy odpowiednią nagrodą.

Dariusz Plewczynski, PhD, Professor of Exact and Natural Sciences; Principal Investigator
Phone: +48 22 554 36 54 or +48 22 234 7219
e-mail: d.plewczynski@cent.uw.edu.pl or Dariusz.Plewczynski@pw.edu.pl www: https://plewczynski-lab.org

Laboratory of Functional and Structural Genomics LFSG
Centre of New Technologies, University of Warsaw; Banacha 2c Street, 02-097 Warsaw, Poland
Laboratory of Bioinformatics and Computational Genomics LB!GO
Faculty of Mathematics and Information Science, Warsaw University of Technology; ul. Koszykowa 75, 00-662 Warsaw, Poland