STRESZCZENIE PRACY DOKTORSKIEJ

Magdalena Wiercioch, mgr

Promotor: Prof. dr hab. Jacek Tabor, Uniwersytet Jagielloński

Opracowanie uniwersalnych metod reprezentacji danych do poprawy jakości metod uczenia maszynowego z przykładowym zastosowaniem w chemii

W ostatnich latach modele bazujące na głębokim uczeniu sieci neuronowych okazały się przydatne w dziedzinie uczenia reprezentacji (ang. *representation learning*). Wyraźny sukces takich algorytmów widoczny jest w obszarze wizji komputerowej czy rozpoznawania głosu, gdzie zaproponowane komputerowe metody osiągnęły wyniki poziomu eksperta. Niemniej jednak okazuje się, że wykorzystanie dostępnych metod do modelowania związków chemicznych nie jest trywialnym zadaniem. Powodów takiego stanu jest kilka, ale należy do nich fakt, iż molekuła jest zupełnie innym obiektem aniżeli na przykład obraz. Ponadto w problemach chemicznych liczba poetykietowanych danych treningowych jest wyraźnie mniejsza (ang. *inductive bias*).

Dlatego głównym wyzwaniem postawionym w pracy doktorskiej jest zaproponowanie metod uczenia reprezentacji, które mają mieć zastosowanie m.in. w procesie projektowania leków. Wychodząc naprzeciw postawionemu zadaniu autorka pracy przedstawiła trzy algorytmy pozwalające na uzyskanie komputerowej reprezentacji molekuły. W celu oceny przydatności modeli, wykonano szereg eksperymentów obejmujących zadania klasyfikacji oraz regresji.

Uściślając, w pracy autorka zaproponowała rozwiązania trzech następujących zagadnień informatycznych.

- *Zadanie klasyfikacji*. W tym obszarze autorka używa utworzonego modelu do wykrywania związków chemicznych aktywnych biologicznie.
- *Zadanie klasyfikacji, regresji and interpretowalności*. W tym obszarze autorka używa utworzonego modelu do przewidywania toksyczności związków chemicznych.
- *Uczenie reprezentacji grafów i ciągów bazujący na koncepcji głębokiego uczenia*. W tym obszarze autorka używa utworzonego modelu do sprawdzenia czy molekuła i biologiczny cel wchodzą z sobą w interakcję.

Praca składa się z wstępu, czterech rozdziałów, podsumowania oraz dodatku.

Rozdział I - Wprowadzenie

Na przestrzeni lat naukowcy zauważyli, że rodzaj i jakość komputerowej reprezentacji danych wybranych do trenowania algorytmu uczenia maszynowego wpływa bezpośrednio na ostateczne wyniki modelu. Poza tym nie jest tajemnicą, iż skuteczność algorytmu jest skorelowana z typem problemu. Opierając się tych obserwacjach, jednym z kierunków jest próba wyboru zbioru atrybutów danych, który może charakteryzować dane wejściowe, a następnie użycie ich w zadaniach klasyfikacji czy regresji. Tego typu podejście będące w pewnym stopniu nawiązaniem do teorii reprezentacji i uczenia reprezentacji przynosi w miarę zadowalające efekty, zwłaszcza w przypadku obszarów, gdzie dostępna jest znikoma liczba danych.

W ciągu lat pojawiło się również pojęcie uczenia reprezentacji, gdzie zadaniem jest dostarczenie parametrycznej funkcji zwracającej wektor cech opisujących dane. Celem jest wyodrębnienie abstrakcyjnych własności danych, które mają umożliwić algorytmom poprawę jakości predykcji dla nowych danych. Co więcej, zagadnienie uczenia reprezentacji jest tym bardziej istotne, iż typy danych, z którymi aktualnie pracują systemy uczące (takie jak obrazy czy teksty) charakteryzują się wielowymiarowością. Okazuje się, że metody redukcji wymiaru nie są tutaj zbyt przydatne.

W klasycznym podejściu do uczenia maszynowego dane są przetwarzane, a następnie podaje się je na wejście algorytmu uczącego, który zwraca rezultat. Zatem tzw. część ucząca odnosi się do wyznaczenia rozwiązania w oparciu o dane wcześniej rzutowane do przestrzeni o mniejszym wymiarze niż oryginalnie. W tym obszarze niejako przewagę stanowi podejście stosowane w uczeniu głębokim, gdzie komputerowa reprezentacja powstaje na podstawie danych wejściowych jako, w pewnym sensie, uboczny efekt algorytmu. Dodatkowo, rozwój sprzętowy, jaki można zauważyć, szczególnie z obszarze kart graficznych, spowodował, że tradycyjne metody związane z selekcją cech w wielu dziedzinach stają się mało atrakcyjne. Dlatego sporo uwagi poświęca się teraz na działania związane z inżynierią modeli/architektur, co bezpośrednio nawiązuje do uczenia reprezentacji. Potwierdzeniem tej tendencji jest mnogość rozwiązań bazujących na głębokich architekturach, które powstały w ostatnich latach i okazały się sukcesem.

Mówiąc o uczeniu reprezentacji, trudno jest podać jednoznaczną definicję, co to oznacza ‘uczyć reprezentację’. Z pewnością intuicja oraz wcześniejsze badania podpowiadają, że dobra reprezentacja ułatwia proces uczenia. Prawie dziesięć lat temu Bengio, Courville i Vincent [1] przeanalizowali cechy dobrej reprezentacji. Zgodnie z ich obserwacjami można wskazać kilka kluczowych własności, którymi charakteryzuje się odpowiednia reprezentacja. Należą do nich: lokalna gładkość, spójność, hierarchiczność. Poza tym, istotne jest aby reprezentacja odzwierciedlała relacje występujące pomiędzy obiektami.

Jednocześnie można zauważyć, że problem reprezentacji związków chemicznych jest szeroko podejmowany już przez kilka dekad [2]. Generalnie molekuła przedstawiana jest w postaci grupy atomów połączonych wiązaniem. Niestety, taka forma jest niewystarczająca, żeby zrozumieć prze-

strzeń chemiczną i wskazać rozwiązanie wielu zagadnień chemicznych, takich jak przewidywanie własności molekuł [3]. Autorka przyznaje, że pojawiło się wiele propozycji w kontekście komputerowej reprezentacji związku chemicznego, aczkolwiek jest wyraźna potrzeba wykonywania dalszych badań w tym zakresie. Uściślając, w pracy reprezentacja molekuły $\mathcal{R}$ rozumiana jest jako rzutowanie molekuły $\mathcal{M}$ na pewien zbiór X.

Niemniej jednak pomimo wielu zalet podejść bazujących na głębokim uczeniu, wyzwaniem badawczym pozostaje wykorzystanie tych metod w problemach występujących w chemii. Jednym z głównych powodów są dane. Po pierwsze, przestrzeń chemiczna jest ogromna, a szacuje się na około 10^{60} [4] molekuł. Wskazuje to na potrzebę eksploracji nowych rozwiązań dla różnorodnych zastosowań. Z drugiej strony należy mieć świadomość, iż liczba poetykietowanych danych treningowych jest ograniczona jeśli mowa o konkretnym problemie chemicznym. Dodatkowym utrudnieniem jest tzw. *bias* i niezrównoważenie danych. Niestety w odróżnieniu do dziedzin takich jak wizja komputerowa czy przetwarzanie języka naturalnego, w przypadku chemii zgromadzenie etykiet dla danych jest bardziej złożonym procesem. Przede wszystkim wymaga wykonania eksperymentów - kosztownych czasowo i finansowo.

Problemem badawczym podjętym w pracy doktorskiej jest uczenie reprezentacji. W tym celu autorka skupiła się na trzech zadaniach informatycznych: klasyfikacji, regresji oraz głębokim uczeniu na grafach i ciągach. Dodatkowo przeprowadzono badania w zakresie interpretowalności wyników. Ewaluacja zaproponowanych algorytmów została dokonana w serii komputerowych eksperymentów, które koncentrowały się na rozwiązaniu wybranych konkretnych problemów występujących w chemii, tj. przewidywania aktywności biologicznej związków chemicznych, przewidywania właściwości molekuł oraz wykrywania interakcji molekuła a biologiczny cel.

Rozdział II - Kontekst: cheminformatyka i uczenie reprezentacji dla danych chemicznych

Rozdział II wprowadza niezbędne podstawy wyjaśniające dlaczego podjęty temat jest istotny i przedstawia informacje dotyczące dostępnych komputerowych modeli molekuł oraz wyzwań stojących przed cheminformatyką. Autorka wskazuje, wzmacniając swoje stanowisko pozycjami z literatury, że projektowanie leków nie jest prostym zagadnieniem. Przede wszystkim czas i pieniądze stanowią ograniczenie pełnego rozwoju dziedziny [5]. Jednym z dowodów tej koncepcji jest fakt, iż powstanie leku, który można zakupić w aptece zabiera co najmniej 5-10 lat z kosztem sięgającym miliardów dolarów. Tym bardziej, cheminformatyka, a zwłaszcza techniki oparte na głębokich sieciach neuronowych upatrywane są jako oręże do wychodzenia naprzeciw zadaniom *Komputerowo wspomaganego projektowania leków (ang. Computer Aided Drug Design, CADD).*

Najważniejszym argumentem przemawiającym za metodami uczenia głębokiego jest możli-

wość otrzymania reprezentacji wprost z danych wejściowych. Dlatego też współcześnie prace naukowe koncentrują się na dwóch ścieżkach wykorzystania uczenia głębokiego w zależności od rodzaju danych podawanych na wejście modelu. Zatem rozważa się podejścia bazujące na tekstach i algorytmy uwzględniające teorię grafów. Pierwszy wariant bierze po uwagę powszechną płaską reprezentację molekuł w postaci ciągów SMILES, gdzie związek chemiczny zakodowany jest za pomocą zestawu znaków zestawionych zgodnie z ustaloną gramatyką [6]. Drugą opcją jest przedstawienie molekuły w postaci grafu z węzłami odnoszącymi się do atomów i krawędziami skojarzonymi z wiązaniami chemicznymi. Koncepcja, w której związek chemiczny utożsamiany jest z grafem prowadzi do architektur opartych o grafowe sieci neuronowe [7, 8].

Faktycznie, metody grafowe okazują się być przydatne i mają udział w wielu znaczących odkryciach. Niemniej jednak ciągle jest potrzeba ich rozwijania, w szczególności z uwzględnieniem potrzeb stawianych przez chemię. Jednym z obszarów do badań jest uwzględnienie małej liczby poetykietowanych danych chemicznych wobec jednoczesnego zapotrzebowania głębokich modeli na dane. Eksperymenty chemiczne pochałaniają czas, wiążą się z błędami i koniecznością powtórzeń. Poza tym wymagają wiedzy eksperckiej [9, 10]. Ograniczona liczba danych powoduje z kolei problemy związane z uczeniem. Częstym zjawiskiem jest zatem przeuczenie (ang. overfitting) i problem uzyskania zdolności generalizacji [11]. Dlatego oczekuje się, że kolejne modele będą uwzględniać wymienione wyzwania poprzez skalowalność i wysoką dokładność predykcji dla nowych danych. Kolejnym zagadnieniem, które powinno być bardziej szczegółowo brane pod uwagę jest dostarczenie szczegółowych informacji związanych ze związkiem chemicznym do systemu uczącego. Niewątpliwie warto aby nowe modele uwzględniały występujące zależności obecne w związku chemicznym. W szczególności mowa jest o włączeniu do algorytmów modułów opisujących zależności strukturalne pomiędzy atomami i wiązaniami oraz interakcji.

W oparciu o przeanalizowane potrzeby w pracy autorka zaproponowała trzy architektury oparte na koncepcji głębokich sieci neuronowych, które powodują utworzenie reprezentacji związku chemicznego. Modele zostały przedstawione w Rozdziale III, Rozdziale IV oraz Rozdziale V. Skuteczność rozwiązań potwierdzają eksperymenty bazujące na zadaniach klasyfikacji i regresji.

Rozdział III - Uczenie Hybrydowe dla Klasyfikacji

Ważnym początkowym etapem w procesie powstawania leku jest ocena aktywności związku chemicznego [12]. Z punktu widzenia informatycznego, zadanie to sprowadza się do problemu klasyfikacji. W efekcie powstało sporo komputerowych metod służących do detekcji związków aktywnych. Wiele z tych podejść zalicza się do technik QSAR (ang. Quantitative Structure-Activity Relationship), w których działania opierają się na znalezieniu zależności pomiędzy strukturą a na przykład aktywnością biologiczną [13]. Inne metodologie bazują na podobieństwie molekuł [14].

Ogromny wkład w dziedzinę mają również tradycyjne algorytmy uczenia maszynowego operujące na chemicznych deskryptorach [15].

Okazuje się jednak, że klasyczne systemy uczące mają wady, do których zalicza się konieczność pracy nad reprezentacjami wektorowymi i związana z tym wiedza ekspercka. Naprzeciw tym ograniczeniom stają algorytmy głębokiego uczenia. Niestety, aktualnie w praktyce metody wykorzystujące sieci neuronowe również ujawniają szereg wyzwań. Jedną kwestię stanowi potrzeba dużych i dobrej jakości zbiorów treningowych. Innym zagadnieniem jest fakt, iż zwiększenie głębokości sieci wiąże się ze zwiększeniem wymagań na zasoby obliczeniowe. Ponadto dotychczasowe modele nie eksplorują w pełni informacji strukturalnej związanej z molekułą, przez co pomijane są istotne zależności chemiczne. W Rozdziale III autorka proponuje pierwszą metodę odpowiadającą na wymienione oczekiwania nazwaną **H**ybrid **D**eep **N**eural **N**etwork (HybNN).

Metoda

Rozdział prezentuje algorytm **H**ybrid **D**eep **N**eural **N**etwork (HybNN), który jest nowym podejściem łączącym sieć BiGRU oraz grafową sieć neuronową. W architekturze zawarto pięć głównych elementów składowych: moduł reprezentacji słowa, moduł reprezentacji grafu, sieć BiGRU, moduł reprezentacji przestrzennej, moduł wyjściowy. Moduł reprezentacji grafu odpowiada za przekształcenie związku chemicznego do postaci grafu. Następnie moduł reprezentacji przestrzennej służy do zebrania informacji odnoszących się do własności molekuły. Z kolei moduł reprezentacji słowa pozwala na identyfikację podsłów składających się na SMILES. Przy tym sieć BiGRU pozwala na wyodrębnienie informacji lokalnej i globalnej z SMILES. Zgromadzone cechy podlegają operacji konkatenacji, a następnie przekazywane są do modułu wyjściowego, który odpowiada za predykcję.

Rezultaty

Przydatność zaproponowanej metody HybNN została potwierdzona serią eksperymentów. W tym celu autorka posłużyła się czternastoma zbiorami danych dostępnymi w bazie PubChem [16]. Każdy zbiór zawiera informację dotyczącą aktywności lub braku aktywności molekuły względem receptora.

Zaprojektowane w pracy eksperymenty sprawdzają skuteczność HybNN pod względem kilku aspektów: 1) zdolność predykcji na zbiorze walidacyjnym; 2) zdolność predykcji na zbiorze testowym; 3) analiza szybkości procesu uczenia; 4) analiza wpływu różnorodności zbioru treningowego na wyniki; 5) analiza wartości funkcji kosztu; 6) analiza wpływu zmennej liczby warstw GRU na wyniki; 7) analiza wpływu braku kluczowych komponentów architektury na wyniki. W Rozdziale III porównane zostały rezultaty HybNN z innymi uznanymi aktualnie podejściami. HybNN wyka-

zuje przewagę dla wszystkich użytych zbiorów danych.

Rozdział IV - Uczenie Zorientowane na Podgrafy dla Problemów Uczenia Nadzorowanego

Niewątpliwie wyzwaniem uczenia maszynowego i głębokiego uczenia jest rozmiar zbioru treningowego. Powoduje to brak możliwości wyucznenia modelu, przeuczenie, słabe zdolności do generalizacji i interpretowalności. Jednocześnie wiele badań wskazuje, że toksyczność jest częstym powodem wykluczającym molekuły w drodze do powstania leku [17]. Używa się wprawdzie podejść typu *in vivo*, ale wymagają one długotrwałych eksperymentów laboratoryjnych i skutkują wciąż zbyt małą liczbą molekuł o sprawdzonych własnościach [18]. W dodatku zauważalne są rozbieżności pomiędzy próbkami ludzkimi a zwierzęcymi. Dlatego coraz częściej wybiera się podejścia typu *in vitro* i wykorzystanie uczenia maszynowego.

Rozdział IV wychodzi w kierunku rozwiązania problemu małej liczby danych dla problemu toksyczności molekuł poprzez użycie podgrafów eksplorujących relacje pomiędzy atomami i globalne cechy molekuły.

Metoda

Autorka proponuje nową metodę nazwaną jako **S**ubgraph **E**ncoded **N**eural **N**etwork (SENN). W ogólnej perspektywie architekura sieci składa się z siedmiu komponentów. Na wejściu SENN znajduje się graf reprezentujący związek chemiczny. Do konstrukcji molekularnego grafu $\mathcal{G}$ autorka wybiera zestaw atrybutów chemicznych związanych z wierzchołkami i krawędziami. Warto dodać, że każdej krawędzi zostaje przypisana waga odpowiadająca za 'siłę' połączenia. Następnie graf jest przetwarzany w celu uzyskania zbioru podgrafów podanych dalej na wejście sieci konwolucyjnej. W ten sposób autorka uzyskuje reprezentację $out_{\mathcal{G}}$. Ta reprezentacja jest konkatenowana z wektorem globalnych cech molekuły out_{att}. Do używanych atrybutów należy masa cząteczkowa czy liczba wiązań kowalencyjnych. W celu wyodrębnienia atrybutów wykorzystano pakiet ChemoPy [19]. Połączone reprezentacje przekazywane są do zestawu warstw liniowych z dropout, a następnie do warstwy decyzyjnej.

Autorka zwraca uwagę, że jedną z kluczowych operacji jest uzyskanie reprezentacji grafowej. W Rozdziale IV opisane są szczegóły podejścia, w którym dla grafu $\mathcal{G}$ wyznacza się podgrafy o długości krawędzi co najwyżej k. Do każdego takiego podgrafu i węzłów przypisuje się wektor jednostkowy. W kolejnych warstwach sieci reprezentacje poszczególnych węzłów są aktualizowane. W szczególności każdy wektor jest zastępowany poprzez średnią poszczególnych wektorów węzłów sąsiadujących. Następnie wykonywana jest transformacja liniowa, wyliczone reprezentacje poszczególnych węzłów są uśredniane i powstaje d_0-wymiarowa reprezentacja grafu wejściowego.

Rezultaty

Do weryfikacji skuteczności metody SENN wykonano badania w zakresie klasyfikacji i regresji. Do klasyfikacji autorka wykorzystała zbiór danych w formacie SDF-ów i SMILES pochodzący z konkursu Tox21 Data Challenge [20]. Natomiast do zadania regresji posłużyły dane udostępnione przez autorów pracy naukowej dotyczącej toksyczności molekuł [21].

Eksperymenty obejmują porównanie SENN z pięcioma często stosowanymi metodami uczenia maszynowego i głębokiego uczenia. Rezultaty wskazują, że SENN uzyskuje zadowalające wyniki. Przykładowo dla jednego ze zbiorów danych związanych z klasyfikacją SENN ujawnia AUC-ROC wyższy (0.802 ± 0.006) niż inne algorytmy.

Rozdział V - Głębokie Uczenie Grafów i Ciągów

Wśród wielu zadań związanych z projektowaniem leków wiele uwagi przywiązuje się do identyfikacji interakcji lek - biologiczny cel [22]. Zwykle za taki cel uznaje się białka [23].

Niewątpliwie badanie interakcji lek - biologiczny cel jest znaczące z punktu widzenia poszukiwania leków. Sporo badaczy interpretuje to zadanie jako zadanie klasyfikacji. Mimo wielu opracowań i metod aktualne techniki nie są wystarczające. Pojawia się problem braku wiarygodnych danych. Dodatkowo sytuację komplikuje ograniczona liczba trójwymiarowych białkowych struktur. W konsekwencji dostępne metody nie przynoszą satysfakjonujących rezultatów.

W odpowiedzi na wymienione wady autorka w Rozdziale V proponuje architekturę sieci neuronowej nazwaną **T**riplet **E**ncoded **N**eural **N**etwork (TENN). TENN ma na celu przewidywanie interakcji lek - cel badawczy poprzez eksplorację topologii molekuły wraz z uwzględnieniem informacji semantycznej zawartej w reprezentacji tekstowej.

Metoda

Triplet **E**ncoded **N**eural **N**etwork (TENN) składa się z trzech głównych jednostek, których celem jest integracja różnorodnych informacji odnoszących się oddzielnie do związku chemicznego i białka. Każdy z trzech komponentów zwraca swoją reprezentację. Rezultatem jest reprezentacja wstawiona na wejście kilku liniowych warstw z dropout i predykcja.

Rezultaty

Ocena zaproponowanej metody oparta jest na danych z bazy BindingDB [24] zawierającej informację dotyczące interakcji białek z molekułami. Autorka posłużyła się również bazą DrugBank [25].

Główną przewagą TENN jest możliwość dostarczenia nisko - wymiarowej wektorowej reprezentacji i wskazanie prawdopobieństwa interakcji pomiędzy molekułą a białkiem. Wyniki przeprowadzonych przez autorkę eksperymentów wskazują, że jeśli wyodrębni się informację o globalnych cechach z tekstowej reprezentacji białek oraz molekuł, wówczas poprawia się skuteczność predykcji. Co więcej, zastosowane podejście można użyć do wykrywania bardziej złożonych interakcji. Porównanie algorytmu TENN z czterema znanymi modelami ujawnia jego przewagę. Autorka wskazuje, że chociaż zaproponowana metoda koncentruje się na zastosowaniu w chemii, to jest uniwersalna i może być użyta do badania interakcji pomiędzy innymi obiektami niż molekuły.

Rozdział VI - Podsumowanie

Przedstawione w pracy doktorskiej metody wprowadziły poprawę wyników predykcji w zadaniach związanych z chemią. Niemniej jednak autorka przyznaje, że nadal pozostaje wiele zagadnień, nad którymi można pochylić się. W pierwszym etapie należałoby skupić się na wyjaśnialności (ang. explainability) wprowadzonych metod. Innym ciekawym zadaniem byłaby poprawa zaproponowanych rozwiązań, na przykład wykorzystując podejście few-shot learning.

Dodatek

Rozdział stanowi opis najbardziej przydatnych dla pracy dostępnych modeli opartych o głębokie sieci neuronowe. W tym celu autorka najpierw przedstawia znane płytkie architektury, a następnie przechodzi do bardziej rozbudowanych, w tym grafowych sieci. Na zakończenie omówione zostają szczegóły uczenia nadzorowanego.

Literatura

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] W. J. Wiswesser, "107 years of line-formula notations (1861-1968)," *Journal of Chemical Documentation*, vol. 8, no. 3, pp. 146–150, 1968.

[3] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. molecular frameworks," *Journal of medicinal chemistry*, vol. 39, no. 15, pp. 2887–2893, 1996.

[4] R. S. Bohacek, C. McMartin, and W. C. Guida, "The art and practice of structure-based drug design: a molecular modeling perspective," *Medicinal research reviews*, vol. 16, no. 1, pp. 3–50, 1996.

[5] P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, *et al.*, "Rethinking drug design in the artificial intelligence era," *Nature Reviews Drug Discovery*, vol. 19, no. 5, pp. 353–364, 2020.

[6] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

[7] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 2, pp. 729–734, IEEE, 2005.

[8] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.

[9] L. David, J. Arús-Pous, J. Karlsson, O. Engkvist, E. J. Bjerrum, T. Kogej, J. M. Kriegl, B. Beck, and H. Chen, "Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research," *Frontiers in pharmacology*, vol. 10, 2019.

[10] E. B. Lenselink, N. Ten Dijke, B. Bongers, G. Papadatos, H. W. Van Vlijmen, W. Kowalczyk, A. P. IJzerman, and G. J. Van Westen, "Beyond the hype: deep neural networks outperform established methods using a chembl bioactivity benchmark set," *Journal of cheminformatics*, vol. 9, no. 1, pp. 1–14, 2017.

[11] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec, "Strategies for pre-training graph neural networks," *arXiv preprint arXiv:1905.12265*, 2019.

[12] P. Buchwald and N. Bodor, "Computer-aided drug design: the role of quantitative structure–property, structure–activity and structure–metabolism relationships (qspr, qsar, qsmr)," *Drugs Future*, vol. 27, no. 6, pp. 577–588, 2002.

[13] J. C. Dearden, "The history and development of quantitative structure-activity relationships (qsars)," in *Oncology: breakthroughs in research and practice*, pp. 67–117, IGI Global, 2017.

[14] F. R. Burden and D. A. Winkler, "New qsar methods applied to structure- activity mapping and combinatorial chemistry," *Journal of chemical information and computer sciences*, vol. 39, no. 2, pp. 236–242, 1999.

[15] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, *et al.*, "Qsar modeling: where have you been? where are you going to?," *Journal of medicinal chemistry*, vol. 57, no. 12, pp. 4977–5010, 2014.

[16] Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He, and J. Zhang, "Pubchem bioassay: 2017 update," *Nucleic acids research*, vol. 45, no. D1, pp. D955–D963, 2017.

[17] M. Hay, D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal, "Clinical development success rates for investigational drugs," *Nature biotechnology*, vol. 32, no. 1, pp. 40–51, 2014.

[18] G. J. Harry, M. Billingsley, A. Bruinink, I. L. Campbell, W. Classen, D. C. Dorman, C. Galli, D. Ray, R. A. Smith, and H. A. Tilson, "In vitro techniques for the assessment of neurotoxicity.," *Environmental health perspectives*, vol. 106, no. suppl 1, pp. 131–158, 1998.

[19] D.-S. Cao, Q.-S. Xu, Q.-N. Hu, and Y.-Z. Liang, "Chemopy: freely available python package for computational biology and chemoinformatics," *Bioinformatics*, vol. 29, no. 8, pp. 1092–1094, 2013.

[20] N. C. for Advancing Translational Sciences, "Tox21 data challenge 2014," 2014.

[21] K. Wu and G.-W. Wei, "Quantitative toxicity prediction using topology based multitask deep neural networks," *Journal of chemical information and modeling*, vol. 58, no. 2, pp. 520–531, 2018.

[22] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities," *Information Fusion*, vol. 50, pp. 71–91, 2019.

[23] Y. Feng, Q. Wang, and T. Wang, "Drug target protein-protein interaction networks: a systematic perspective," *BioMed research international*, vol. 2017, 2017.

[24] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D198–D201, 2007.

[25] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, *et al.*, "Drugbank 5.0: a major update to the drugbank database for 2018," *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2018.