Aine

Tietojenkäsittelytieteen kandiohjelma

koneoppimisen menetelmät lääkkeiden/kemiallisessa syntetisoinnin mallennuksessa

Heikki Pulli

22.10.2021

Matemaattis-luonnontieteellinen tiedekunta

Helsingin yliopisto

Ohjaaja(t)

Tarkastaja(t)

Yhteystiedot

PL 68 (Pietari Kalmin katu 5)
00014 Helsingin yliopisto

Sähkopostiosoite: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

Sisällys

1 Johdanto

Uusien lääkkeiden tuottaminen on pitkä ja kallis prosessi. Tavallisesti aika, jossa lääke ensin löudetään, testataan ja hyväksytään useiden eri testien läpi, vaihtelee 10-12 vuoden välillä ja hinta on 1-2 miljardin dollarin välillä. [4].

Lääkefirmat ovatkin alkaneet selvittää, kuinka eri koneoppimisen malleja voidaan hyödyntää lääketutkimuksessa nopeuttamaan suurimpia pullonkauloja. [4]. Eri mallit voivat esimerkiksi karsia kaikista harkinnasta olevista lääkkeistä vain lupaavimmat kandidaatit, joilla on mahdollisuus päästä testeistä läpi tuotantoon. Tai koneoppimismalleja voidaan hyödyntää täysin uusien lääkeaineiden etsinnässä, joilla on halutut lääkkeelliset että fysikaaliset ominaisuudet. [**!lÃďhde!**]

Uuden lääkkeen löytäminen sairauteen on pitkä ja kallis prosessi. Tähän kuuluu useita eri vaiheita ja eri vaiheet vievät eri määrän rahaa ja aikaa. Nämä ovat sairauden aiheuttajan tunnistaminen, tähän vaikuttavan lääkkeen tunnistaminen, lääkkeen optimointi, lääkkeen ominaisuuksien analysointi ja kliiniset testit. Näiden jälkeen lääke joko hyväksytään myyntiin tai ei. Nämä eri vaiheet vievät tavallisesti 10 - 12 vuotta ja hintaa tälle tulee noin 2.8 miljardia dollaria. Vaiheiden pitkän keston ja suuren hinnan takia tutkijat ja lääkefirmat ovatkin alkaneet tutkia mahdollisia keinoja, jotka nopeuttaisivat tai halventaisivat tätä lääkkeen kehityksen prosessia.

Koneoppimisenmallit ovat nousseet houkuttelevaksi vaihtoehdoksi, joka voisi nopeuttaa tätä prosessia. viimeisen kymmenen vuoden aikana saatavilla olevan laadukkaan datan määrä on kasvanut merkittävästi ja uusia tehokkaampia koneoppimismalleja on kehitetty, joita voidaan hyödyntää lääketutkimuksessa. Kehitetyt mallit ovatkin näyttäneet, että koneoppimismallit ovat tehokkaita työkaluja, joita voidaan hyödyntää kaikissa lääketutkimuksen prosessin vaiheissa.

Tässä tektissä paneudutaan syvemmin koneoppimismalleihin, joita käytetään uusien lääkkeiden tunnistaminseen ja näiden tunnistettujen lääkkeiden syntetisoinnin suunnitteluun.

2 Uusien lääkkeiden löytäminen

Yksi ensimmäisistä lääketutkimuksen prosessin osa-aluiesta on uusien lääekyhdisteiden löytäminen joko uusiin tai jo tunnettuihin tauteihin. [4] Tämä on kuitenkin ollut tavallisesti hidas prosessi ja uuden toimivan yhdisteen löytäminen on kestänyt kahdesta kolmeen vuotta. Lisääntynyt datan määrä on kuitenkin mahdollistanut tämän osa-alueen nopeuttamisen koneoppimismallien avulla. Tähän ongelmaan on kehitetty useita eri koneoppimismalleja. Mariya Popovan, Olexandr Isayev ja Alexander Tropsha tutkimusryhmä on kehittänyt mallin, joka ehdottaa uutta yhdistettä perustuen mallin syötteenä saamaan ominaisuus vektoriin. [7] Shahar Harelin ja Kira Radinskyn tutkijaryhmä puolestaan ovat kehittäneet mallin, joka luo uusia yhdisteitä, jotka perustuvat syötteenä annettuun prototyyppi yhdisteeseen. [5]

Jotta koneoppimismalleja voidaan hyödyntää lääketutkimuksessa täytyy olla saatavilla tarpeeksi dataa tutkittavasta aiheesta. [4] Viimeisimmän kymmenen vuoden aikana saatavilla olevan datan määrä on kasvannut merkittävästi kehitettyjen tietopankkien takia. Näitä ovat esimerkiksi PubChem ja ChEMBL.

2.1 virtual screening

Erillaisten kemiallisten yhdisteiden avaruus on suuri. On arvioitu, että erillaisia kemiallisia yhdisteitä, jotka voivat esiitnyä huoneen lämmössä ja nesteessä, voi olla välillä $10^{18} - 10^{180}$. [10] Lääkkeeksi käyvien yhdisteiden määrä on taas puolestaan arvioitu olevan koko luokkaa 10^{60}. [10] Tämä itsessään esittää tarpeen tehokkaille algoritmeille ja menetelmille, jotka auttavat karsimaan tästä suuresta määrästä kemiallisia yhdisteitä vain lupaavimmat.

Virtual screening (VS) on joukko menetelmiä uusien lääkkeiden löytämiseksi. VS menetelmillä tarkoitetaan yleisesti prosesseja, joissa käydään läpi suuria tietokantoja dataa, jotta löydetään haluttu yhdiste. [10]

2.1.1 Koneoppimisen käyttökohteet VS:ässä

Ongelman kuvaaminen koneoppimisongelmana [5, 6]

2.2 Kehitettyjä koneoppimismalleja

2.2.1 Prototyyppiin perustuva lääkkeen suunnittelu

- Prototype based drug design [5]

2.2.2 Deep generative autoencoder

- Deep generative autoencoder [6]

3 Uusien lääkkeiden syntetisointi

Yhdisteen syntetisoinnin suunnittelulla tarkoitetaan prosessia, jossa määritellään, kuinka haluttu yhdiste voidaan tuottaa synteettisesti saatavilla olevista lähtöaineista. [3] Retrosynteesi analyysillä tarkoitetaan puolestaan menetelmää, jonka avulla löydetään halutun yhdisteen tuottamiseen tarvittavat lähtöaineet. Retrosynteesi toimii siis toiseen suuntaan kuin syntetisointi. Retrosynteesissä yhdiste pilkotaan rekursiivisesti pienempiin lähtöaineisiin kunnes jäljellä on vain saatavilla olevia lähtöaineita.

Tavallisesti yhdisteen retrosyntetisointi on vaatinut suorittavalta kemistiltä usean vuoden kokemusta ja tietoa saatavilla olevista lähtöaineista ja eri reaktioista. Tätä on pyritty automatisoimaan eri CASP -menetelmien avulla (Computer-Aided Synthesis Planning). Ensimmäiset CASP -menetelmät perustuivat heuristisiin algoritmeihin, joissa kemistit käsin koodasivat, miten eri lähtöaineet reagoivat keskenään ja mikä on reaktion lopputuote. Tämä on kuitenkin osoittautunut toivottomaksi yritykseksi massiivisen datan määrän takia.

Kehitys koneoppimismenetelmissä on kuitenkin tarjonnut uuden lähestymis tavan CASP -menetelmien keshitykseen. Sen sijaan, että kemistit loisivat heuristisia malleja, niin uudet koneoppimismallit koulutetaan saatavilla olevan datan avulla. Tämä on todettu merkittävästi enemmän toteutettavaksi lähestymistavaksi.

Koneoppismallien käyttö ja koulutus ei ole kuitenkaan täysin ongelmaton lähestymistapa myöskään. Ongelmaan liittyen dataa ei välttämättä ole saatavilla ja datan hankkiminen voi olla kallis operaatio. Tätä varten on kehitetty tietopankkeja, jotka sisältävät massiivisia määriä dataa tietystä aiheesta, esim. Reaxys kemiallisesista raektioista.

3.1 Lääkkeen retrosyntetisoinnin haastavuus

Retorsyntetisoinnin tekee hankalaksi fakta, että yhdiste voidaan muodostaa sadoilla tai tuhansilla eri tavoilla. Tämä ongelma toistuu rekursiivisesti, kun yhdiste pilkotaan yhdisteisiin, jotka keskenään reagoidessa muodostavat alkuperäisen yhdisteen. Pienille ja yksinkertaisille yhdisteille tämä vaihtoehto avaruus on pienempi, mutta yhdisteen koon kasvaessa eri tapojen määrä muodostaa haluttu yhdiste kasvaa eksponentiaalisesti.

Tämän takia tarve tätä prosessia yleistäville koneoppimismalleille on suuri. Miksi lääkkeiden retrosyntetisointi on hankalaa? [2, 1]

3.2 Kehitettyjä apuvälineitä

3.2.1 3N-MCTS

3N-MCTS on kehitetty koneoppimismalli, joka etsii retrosynteesi polkuja yksinkertaisempiin ja saatavilla oleviin lähtöaineisiin [8]. 3N-MCTS:än kehitti Marwin Seglerin, Mike Preussin ja Mark Wallerin tutkijaryhmä. Kun retrosynteesi polku on varmennettu ja todettu toimivaksi, niin syötteenä annettu yhdiste on mahdollista syntetisoida laboratoriossa. 3N-MCTS koostuu kolmesta eri koneoppimismallista ja Monte Carlo -puuhaku algoritmista (**Monte carlo tree search, MCTS**). Neuroverkot on koulutettu avustamaan puuhaku algoritmia etenemään fiksuimpaan suuntaan, kun haku algoritmi etsii syntetisointi polkuja ja tarkistamaan, onko ehdotettu reaktio mahdollinen kyseisellä molekyylille.

Neuroverkot ovat hakupuun laajentumisen suuntaa ohjaava verkko (**Expansion policy network, EPN**), MCTS:än rollout toimintoa tukeva Rollout -verkko (**Rollout policy network, RPN**) ja verkko, joka tarkistaa, onko syntetisointi polku toteutettavissa (**In-scope filter network, IFN**).

Data, jolla neuroverkot koulutetaan, on peräisin Reaxys -tietokannasta. Reaxyksen omistaa Elsevier kustantamo. Reaxys -tietokannan sisältämä data koostuu säännöistä, jotka kertovat, mitkä lähtöaineet reagoivat keskenään, mikä reaktio on kysessä ja mikä on reaktion tuote. Näitä sääntöjä käytetään mallien kouluttamiseen. Reaxys sisältää yli 12.4 miljoonaa sääntöä. Mallien kouluttamiseen käytetyt säännöt sisältävät vain yksivaiheisia kemiallisia reaktioita ja reaktiossa on mukana vain yhdestä kolmeen lähtötuotetta. Eri mallien kouluttamiseen käytetiin eri kriteerein suodatettua dataa tietokannasta.

RPN:än kouluttamiseen valittiin datasta vain reaktiossa muuttuneet atomit ja liitokset (reaktiokeskus) ja lähimmät vierekkäiset atomit. Datasta suodatettiin pois sellaiset reaktiot, jotka ilmaantuivat alle 50 kertaa ennen vuotta 2015. EPN:än kouluttamiseen valittiin datasta vain reaktiokeskus. EPN:än datasta suodatettiin pois sellaiset reaktiot, jotka ilmenivät datassa alle kolme kertaa ennen vuotta 2015. Lopulliset reaktio määrät, joilla RPN ja EPN koulutettiin, olivat 17134 ja 301671. Näillä säännöillä EPN ja RPN koulutetaan toimimaan hakualgoritmia ohjaavina neuroverkkoina.

EPN on toteutettu Highway -neuroverkkona (Highway network, HN). HN on hyvin syvä neuroverkko tyyppi, joka saattaa jopa sisältää yli sata kerrosta [11].

RPN on neuroverkko, jossa on yksi piilotettu taso. RPN koulutettiin samalla tavalla kuin EPN.

IFN on neuroverkko, joka tarkistaa, onko EPN:än ja RPN:än valitsemat reaktio säännöt toteutettavissa. IFN koulutetaan sekä onnistuneiden että epäonnistuiden reaktioiden avulla. Koska epäonnistuneita reaktioita ei talleneta tietokantaan, niin kyseinen data generoidaan. Data generoidaan siten, että jos reaktiossa

$$A + B \rightarrow C$$

lähtöaineet A ja B muodostavat reaktiossa lopputuotteen C, niin lopputuotteita D, E, F, jne. ei muodostu (voisi selittää syvemmin). IFN kouluttamista varten luotiin 100 miljoonaa epäonnistunutta reaktiota ja 10 miljoonaa testaamista varten.

3N-MCTS:ässä IFN ja EPN on yhdistetty toimimaan yhdessä. Tutkittaessa puun tilaa S_i (selitä Si vaihe) jokainen molekyyli syötetään EPN:älle ja se tulostaa, mitkä reaktiot voivat muodostaa annetun yhdisteen ja näin ollen myös mitkä lähtöaineet voivat muodostaa annetun yhdisteen. Nämä reaktiot syötetään IFN:älle, joka suodattaa valituista reaktioista toteutettavissa olevat. Tämän jälkeen algoritmissa iteroidaan neljää vaihetta, jotka muodostavat lopullisen puun.

(1) Ensimmäisessä vaiheessa algoritmi valitsee seuraavan lupaavimman tilan puusta kunnes puun lehti on saavutettu. Jos lehdessä käydään ensimmäisen kerran valinta vaiheen aikana, niin lehti arvostellaan simuloimalla hakualgoritmia d askelta eteenpäin samalla muodostaen synteesi polkua (rollout). Jos lehdessä käydään useamman kuin yhden kerran valinta vaiheen aikana, niin mahdolliset reaktiot, jotka muodostavat lehden, tutkitaan ja lisätään lehden lapsiksi (expansion)

(2) Toisessa vaiheessa lupaavien tilojen lapset tutkitaan. Tällöin etsitään lupaavimmat reaktiot, jotka muodostavat kyseisessä tilassa olevan yhdisteen.

(3) Kolmannessa vaiheessa tarkistetaan lehden tila. Jos lehti on 'todistetusti toimiva', niin algoritmi palauttaa luvun suuremman kuin yksi, jolloin lehteä suositellaan käytettävän synteesipolussa. Muussa tapauksessa lehdelle suoritetaan rollout, jolloin RPN antaa rekursiivisesti uusia reaktioita niin kauan, kunnes lehti on pilkottu lähtöaineisiin tai kunnes suurin sallittu syvyys d on saavutettu.

(4) Viimeisessä vaiheessa lehtien arvot päivitetään. Jos lähtöaineet löydetään rolloutin

aikana, niin lehti saa palkonnoksi arvon 1. Jos Kaikkia lähtöaineita ei löydetty, niin lehdelle annetaan osittainen palkinto. Jos yhtään lähtöainetta ei löytynyt, niin lehti saa arvon -1.

Saatta kuietnkin olla, että synteesi polkua ei voida luoda. Joko synteesin polun tutkimiseen menee liian kauan aikaa tai synteesi polku sisältää liian monta vaihetta yhdisteen syntetisoimiseen.

3.2.2 Expert knowledge aided neural networks

- expert knowledge aided neural networks [9]

Miten koneoppimista hyödynnetään tällä hetkellä lääkkeiden syntetisoinnissa? [**10.1145/3219819.3219** 8, 9]

4 Tulevaisuuden koneoppimisen mallit ja kehitys

Miten lääkkeiden kehitys tulee hyötymään tulevaisuuden koneoppimisesta? [2]

Lähteet

[1] A. F. de Almeida, R. Moreira ja T. Rodrigues. ”Synthetic organic chemistry driven by artificial intelligence”. eng. *Nature reviews. Chemistry* 3.10 (2019), s. 589–604. ISSN: 2397-3358.

[2] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev ja A. Walsh. ”Machine learning for molecular and materials science”. eng. *Nature (London)* 559.7715 (2018), s. 547–555. ISSN: 0028-0836.

[3] C. W. Coley, W. H. Green ja K. F. Jensen. ”Machine Learning in Computer-Aided Synthesis Planning”. eng. *Accounts of chemical research* 51.5 (2018), s. 1281–1289. ISSN: 0001-4842.

[4] S. Ekins, A. C. Puhl, K. M. Zorn, T. R. Lane, D. P. Russo, J. J. Klein, A. J. Hickey ja A. M. Clark. ”Exploiting machine learning for end-to-end drug discovery and development”. eng. *Nature materials* 18.5 (2019), s. 435–441. ISSN: 1476-1122.

[5] S. Harel ja K. Radinsky. ”Accelerating Prototype-Based Drug Discovery Using Conditional Diversity Networks”. Teoksessa: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’18. London, United Kingdom: Association for Computing Machinery, 2018, s. 331–339. ISBN: 9781450355520. DOI: 10.1145/3219819.3219882. URL: https://doi-org.libproxy.helsinki.fi/10.1145/3219819.3219882.

[6] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper ja A. Zhavoronkov. ”druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico”. eng. *Molecular pharmaceutics* 14.9 (2017), s. 3098–3104. ISSN: 1543-8384.

[7] M. Popova, O. Isayev ja A. Tropsha. ”Deep reinforcement learning for de novo drug design”. eng. *Science advances* 4.7 (2018), eaap7885–eaap7885. ISSN: 2375-2548.

[8] M. H. S. Segler, M. Preuss ja M. P. Waller. ”Planning chemical syntheses with deep neural networks and symbolic AI”. eng. *Nature (London)* 555.7698 (2018), s. 604–610. ISSN: 0028-0836.

[9] B. Shin, S. Park, J. Bak ja J. C. Ho. "Controlled Molecule Generator for Optimizing Multiple Chemical Properties". Teoksessa: *Proceedings of the Conference on Health, Inference, and Learning*. CHIL '21. Virtual Event, USA: Association for Computing Machinery, 2021, s. 146–153. ISBN: 9781450383592. DOI: 10.1145/3450439.3451879. URL: https://doi.org/10.1145/3450439.3451879.

[10] C. Sotriffer, R. Mannhold, H. Kubinyi ja G. Folkers. *Virtual Screening: Principles, Challenges, and Practical Guidelines*. eng. Vol. 48. Methods and principles in medicinal chemistry. Weinheim: John Wiley ja Sons, Incorporated, 2011. ISBN: 9783527326365.

[11] R. K. Srivastava, K. Greff ja J. Schmidhuber. "Training Very Deep Networks". *CoRR* abs/1507.06228 (2015). arXiv: 1507.06228. URL: http://arxiv.org/abs/1507.06228.