ISSN (Online) 2287-3406
Journal of Life Science **2023** Vol. 33. No. 8. 663~679 DOI : https://doi.org/10.5352/JLS.2023.33.8.663

- Review -

Analysis of Research Trends in New Drug Development with Artificial Intelligence Using Text Mining

Jae Woo Nam[1] and Young Jun Kim[2*]

[1]*Department of Library and Information Science, Konkuk University, Chungju 27478, Korea*
[2]*Department of Medicinal Bioscience and Nanotechnology Research Center, Konkuk University, Chungju 27478, Korea*

Received July 12, 2023 /Revised July 30, 2023 /Accepted August 7, 2023

This review analyzes research trends related to new drug development using artificial intelligence from 2010 to 2022. This analysis organized the abstracts of 2,421 studies into a corpus, and words with high frequency and high connection centrality were extracted through preprocessing. The analysis revealed a similar word frequency trend between 2010 and 2019 to that between 2020 and 2022. In terms of the research method, many studies using machine learning were conducted from 2010 to 2020, and since 2021, research using deep learning has been increasing. Through these studies, we investigated the trends in research on artificial intelligence utilization by field and the strengths, problems, and challenges of related research. We found that since 2021, the application of artificial intelligence has been expanding, such as research using artificial intelligence for drug rearrangement, using computers to develop anticancer drugs, and applying artificial intelligence to clinical trials. This article briefly presents the prospects of new drug development research using artificial intelligence. If the reliability and safety of bio and medical data are ensured, and the development of the above artificial intelligence technology continues, it is judged that the direction of new drug development using artificial intelligence will proceed to personalized medicine and precision medicine, so we encourage efforts in that field.

Key words : AI, deep learning, machine learning, new drug development, text mining

서 론

인공지능(Artificial Intelligence)은 생물학 분야에서 점점 더 중요한 역할을 하며, 약물 개발부터 농업과 생체 대사 경로 개선에 이르기까지 다양한 응용 분야에서 활용될 수 있다. AI를 통해 정확한 진단과 비용 효율적인 치료, 생산성 향상 등을 기대할 수 있으며, 기계 학습(Machine Learning) 및 딥러닝(Deep Learning)을 활용한 프로그램으로 최상의 결과물을 얻을 수 있다. 이는 의학, 농업, 바이오 산업 등 다양한 생물학 분야에서의 혁신과 발전을 도모할 수 있다[149].

Artificial Intelligence (AI)와 Machine Learning (ML)의 발전 추세에 따라 신약 개발 또한 이를 활용하는 추세가 증가하고 있다[22, 60, 74, 77, 84, 109, 113, 123, 128, 134, 138, 147, 165, 168, 171, 172, 174, 177]. 약물의 개발 과정 중 타겟 검증(Target validation), 스크리닝(Target to hit), 선도물질 도출(Hit to Lead), 선도물질 최적화(Lead optimization) 등의 과정에서 AI와 ML을 활용하려는 노력이 논문의 주된 주제를 차지하고 있다. 이러한 노력의 일환으로 신물질의 타겟 검증을 위해 세포 투과가 가능한 후보 분자를 예측하는 모델을 개발하기도 하며[117], 약물의 재배치(repositioning and repurposing)를 위한 프로그램[166] 등이 개발되었다. 한편 분자 동역학 시뮬레이션을 결합한 AI 기반 약물 설계 프로그램 워크플로우를 개발[151]하거나 AI와 ML을 활용하여 임상적 응용 연구를 적용하기 위한 예측 프로그램까지 연구[177]되고 있다.

신약 개발에 있어 AI, ML 뿐만 아니라 딥러닝(Deep Learning, DL)을 이용한 연구도 박차를 가하고 있다. 백신과 항체를 개발하기 위한 항체의 구조 및 서열을 예측해주는 DL 프로그램 개발 연구와 재배치를 위한 연구[69, 118, 145], DL에 기반하여 활용한 분자 특성 및 화합물과 단백질 상호작용 예측 프로그램 개발 연구[19, 146, 166], 유전자 발현과 세포의 생존율 데이터를 활용하여 약물 작용 메커니즘과 DL 기반 파이프라인 구축 연구[46, 154], DL을 사용한 약물 스크리닝 연구[65, 115, 139, 144] 등이 이에 해당한다. DL은 알고리즘을 통해 대량의 데이터 스트림을 처리하는데 적합한 프로그램이기에 생물정보학

***Corresponding author**
Tel : +82-43-840-3569, Fax : +82-43-840-3929
E-mail : ykim@kku.ac.kr

데이터를 처리하는데 많이 활용되고 있다[20, 70, 102, 131, 145].

한편 생물학적 처리와 분석을 위해 컴퓨터 전공자들에게 키워드 유형과 개요를 제공해주는 노력들이 있으며[162], 천연 화합물과 같은 다양한 유망 화합물들로 약물을 설계하기 위한 화학정보학에서의 ML 활용 사례[23, 27, 28, 48, 54, 77, 79, 87, 108, 140, 143], 만성적 자가면역질환인 류마티스 관절염 환자들의 기존 치료 데이터를 DL 모델에 입력하여 조기 진단과 관리를 향상시키는 노력[13, 98, 161] 또한 보이고 있다.

신약 개발 연구에 있어 AI를 활용한 사례의 주된 연구 주제는 암이다. 이와 관련되어 진행되어지고 있는 연구 방식이 AI와 ML, DL 등을 활용하여 새롭게 접근하려는 시도들이 주된 연구의 흐름이며, 새로운 항암제를 개발한다는 의미에서 여전히 AI, ML, DL을 활용하여 신약으로 항암제를 개발하려는 연구가 많이 진행되고 있다[1, 7, 30, 34, 35, 41, 45, 78, 82, 91, 103, 104, 114, 141, 150, 155, 165]. 그리고 암치료에서의 약물 내성 문제를 해결하기 위해 ML과 구조 기반 가상 스크리닝 접근법을 사용하는 연구[3, 49, 72, 76, 102, 125, 132, 139], 방광암 오르가노이드 시스템을 활용하여 DL 활용 모델 개발[99], 암의 유발 인자 예측, 돌연변이 발생, 타겟 예측, 결합 부위 예측 등 종양학 관점에서의 AI를 활용하고자 하는 연구[6, 11, 24, 39, 40]가 주요하게 이루어지고 있다. 그리고 항암제 개발을 위해 AI 활용한 비소세포 폐암의 세포 돌연변이 식별을 목표로 설정한 연구[61, 133], 더 나아가 DL을 활용한 화합물 라이브러리의 예측 성능을 확인한 연구[12, 57, 62, 83] 등이 이루어지고 있다.

최근에 AI 관련 기술의 발달을 통해 신약 개발의 속도가 매우 빨라지고 있다. 이는 AI가 가지고 있는 장점들이 새롭게 각광을 받고 있기에 이에 대해 자세히 살펴볼 필요가 있다고 본다. 이에 본 리뷰 논문은 AI 기술을 이용한 신약개발 관련 연구 동향을 파악하기 위해 PubMed 문헌 데이터베이스를 검색하여 얻어진 논문들을 가지고 그 연구 동향에 대해 텍스트 마이닝(text mining) 기술을 사용하여 일차 분석하고 그 결과를 토대로 하여 AI 기술을 이용한 신약개발 관련 학문적 경향에 대해 정리하였다. 도출된 학문적 경향을 기반으로 하여 생물학자들의 인공지능 기술 적용을 추구하는데 도움이 되고자 AI 기술을 이용한 신약개발의 장점, 한계점, 도전과제, 그리고 발전 전망에 대해 기술하였다.

본 론

연구 동향 분석 방법

문헌 데이터 수집

본 논문은 '인공지능을 활용한 신약개발' 관련한 주제의 연구 동향을 파악하기 위해 PubMed 데이터베이스에서 문헌 데이터를 수집하였고, 텍스트 마이닝을 통해 수집된 데이터를 분석하였다. 문헌 데이터는 연구 명, 초록, 저자명, 수록 저널, 출판 연도 등 다양한 필드로 구성되어 있으나, 텍스트에서 의미를 추출하기 위해서는 다 문장 구조로 작성된 초록이 분석 대상 데이터로 가장 적합하였다. 따라서 각 문헌의 초록을 대상으로 역문서 빈도(Term Frequency-Inverse Document Frequency, TF-IDF) 분석을 수행해 중요 단어를 추출하였고, 토픽모델링(Topic Modeling) 분석을 통해 핵심 연구주제를 발견하고자 하였다. 분석 도구는 R (ver. 4.4.1)과 RStudio를 사용하였다. 데이터 수집을 위한 PubMed 검색식은 Table 1과 같다. 구체적으로 2010년부터 2022년까지 영어로 작성된 연구논문을 검색 대상으로 설정하였고, 검색 결과의 정확성을 위해 통일된 주제어 기반의 MeSH 검색을 수행하였다.

이러한 검색을 통해 얻어진 문헌은 총 2,456건이 검색되었으며, 초록이 없는 35개 문헌을 제외하여 최종 2,421건의 문헌을 로우 데이터로 구축하였다. 기간별 연구 추세를 살펴보면, 2010년부터 2017년까지 매년 100건 미만의 연구가 수행되었고, 2018년부터 100건 이상의 연구가 수행되었다. 특히 2021년부터는 매년 500건 이상의 연구가 수행되어 이 시기부터 연구가 급격하게 증가한 것을 알 수 있었다. 2021~2022년까지의 연구는 총 1,207건으로 2010~2020년 동안 이루어진 1,214건의 연구 량과 비슷한 추세를 나타내고 있다(Fig. 1). 따라서 본 연구에서는 2010~2020년까지를 1 구간으로 설정하고, 2021~2022년을 2 구간으로 설정하여 최신 연구 동향과 이전의 연구 동향의 차이를 살펴보고자 하였다.

데이터 전처리

위의 문헌 검색을 통해 얻어진 결과를 분석하는 과정은 Fig. 2에 제시되었다. 코퍼스(corpus)는 텍스트 마이닝을 위해 구축된 대량의 텍스트 데이터이다. 2010년부터 2022

Table 1. PubMed search method for data collection

("Drug Development"[Mesh:NoExp] OR "Drug Discovery"[Mesh] OR "Drug Repositioning"[Mesh] OR "Drug Development*"[TW] OR "Pharmaceutical Development*"[TW] OR "Drug Discovery"[TW] OR "Drug Design"[TW] OR "Drug Prospecting"[TW] OR "Drug Repurposing"[TW] OR "Drug Rescue"[TW]) AND ("Artificial Intelligence"[Majr] OR "Artificial Intelligence"[TI] OR "Machine Intelligence"[TI] OR "Machine Learning*"[TI] OR "Deep Learning*"[TI]) AND (2010/01/01:2022/12/31[Date - Publication])AND fha[Filter] AND english[lang]

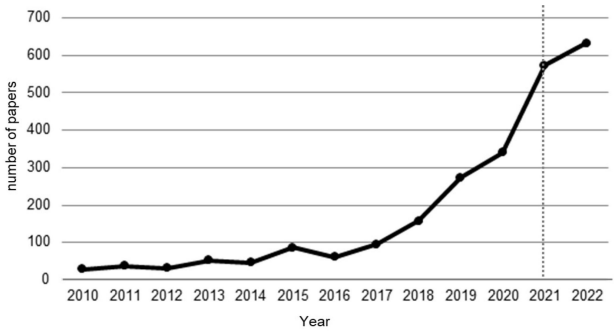

Fig. 1. Trend in the number of papers related to new drug development using artificial intelligence from 2010 to 2022.

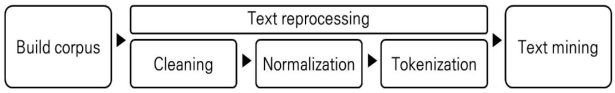

Fig. 2. Text mining research process.

년까지 수행된 2,447개의 연구논문 초록이 코퍼스로 구축되었고, 이들은 자연어 기반으로 작성되어 있으므로 데이터 분석을 위한 전처리 과정이 필요하다. 이에 다음 단계로 데이터 정제(cleaning)를 실시하였다. 데이터 정제는 모든 문자의 소문자 변경, 여백 제거, 특수문자 제거, 숫자 제거 등 다양한 정리 과정이 있지만, 본 연구에서는 고유명사와 전문용어, 약어를 보존하기 위해 문장 부호와 특수문자 제거 및 두 칸 이상의 여백을 한 칸으로 치환하는 수준의 정제 과정을 거쳤다. 또한, 관사, 전치사, 조사, 접속사 등 자주 사용되지만, 데이터 분석에 의미 없는 불용어(stop word)를 제거하였다. 본 연구에서는 R의 ‘tidytext’ 패키지에서 제공하는 1,149개의 불용어 목록을 적용해 일차적 처리를 한 후, ‘via', 'study', 'will' 등 데이터 분석과정에서 큰 의미가 없는 단어를 불용어로 지정해 추가 제거하였다.

다음 단계로 정규화(normalization)를 진행하였다. 영어 단어는 문법에 따라 명사형, 형용사형 등 다양한 품사적 특성과 과거형, 복수형 등 다양한 형태가 있다. 따라서 용언(用言)에서 접두사와 접미사를 분리하고 어근을 도출하여 다양한 형태의 단어들을 하나의 단어로 일반화시키는 어간 추출(stemming) 과정 필요하다. 이 과정을 정규화라고 하며, 본 연구에서는 ‘SnowballC’ 패키지의 stem Document 함수를 사용하여 이 과정을 수행하였다. 그리고 다음 단계로 토큰화(tokenization)를 실시하였다. 토큰화는 코퍼스로 구축된 대량의 텍스트 데이터를 한 개의 단어 또는 n개의 단어(문자열) 단위로 나누는 작업을 의미한다. 분리된 단어를 토큰(token)이라 하며, 일반적으로 구두점이나 공백을 기준으로 토큰을 분리한다. 본 연구에서는 데이터 정제와 정규화 과정을 마친 뒤, 공백 및 1개의 개별 단어를 기준으로 토큰화 하였다.

텍스트 마이닝

데이터 전처리를 마친 대량의 텍스트 데이터에서 정보를 추출하기 위해 텍스트 마이닝을 진행하였다. 텍스트 마이닝은 비정형 텍스트 데이터로부터 의미 있는 정보를 발견하기 위한 분석방법으로 다양한 방법론이 있지만, 본 연구에서는 TF-IDF 분석과 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 알고리즘에 기반한 토픽 모델링 기법을 사용하였다.

TF-IDF 분석은 기계학습 및 정보 검색 등에서 많이 사용하는 단어의 중요도 가중치를 분석하는 것으로, 여러 개의 문서 군이 주어질 때 특정 단어가 한 문서 내에서 얼마만큼 중요한 지를 나타내는 통계적 수치를 말한다[64]. 따라서 TF-IDF 분석은 텍스트 데이터에서 유용한 패턴을 발견하는데 핵심적인 역할을 하고 있으며[31, 32, 55, 66, 68, 97, 101, 120, 122, 164, 170], 다른 데이터 마이닝 기법과 연계를 통해 더 의미 있는 가치를 얻어낼 수 있다[158, 167].

토픽모델링은 구조화되지 않은 방대한 문헌집단에서 일정한 패턴을 발견하여 주제를 찾아 내기 위한 분석방법이다. 다양한 기법이 있지만, LDA는 숨겨진 문서 집단의 의미론적 구조를 이해하기 위한 확률론적 접근 방식이며, 기본 전제는 단어의 분포를 통해 문서의 잠재적 토픽을 나타낼 수 있다는 것이다[14]. 방대한 텍스트 데이터에서 토픽을 추출할 수 있으므로 학문의 동향과 지적 구조 연구 등을 위한 분석 방법으로 사용되고 있다.

실제로도 TF-IDF 분석과 LDA분석은 상호보완적으로 생의학 및 보건 관련 분야의 연구동향 분석과 지적 구조 분석에 많이 사용되고 있다. 이와 관련해 COVID-19 관련 연구 동향을 분석하여 세부 주제를 발견하고 시간의 흐름에 따른 주제의 변화를 분석한 연구[37], COVID-19 연구 동향을 분석하여 14개의 세부 주제를 발견한 연구[18], 트위터의 건강 관련 텍스트에 대한 토픽 모델링과 군집분석 결과를 평가한 연구[86], 트위터와 PubMed의 비만-건강습관 관련 텍스트에서 잠재적 주제를 분석한 연구[169], 의학교육 관련 연구들이 의학 지식과 실습 뿐만 아니라 사회과학 교육 이론과 관련된 주제를 포함하고 있다는 것을 밝힌 연구[64], 스마트 홈 헬스케어 분야 연구들의 잠재적 주제를 분석하여 지적 구조 및 학술 동향을 파악한 연구[67] 등의 다양한 노력들이 수행되었다.

연구 동향 분석 결과

TF-IDF 분석결과

전처리과정을 마친 2,421개의 텍스트 데이터를 대상으로 TF-IDF 분석을 수행하였다. 분석은 1 구간(2010~2020)과 2 구간(2021~2022) 별로 각각 진행하였고, 코퍼스에서

토큰화 된 단어들의 TF-IDF를 분석하였다. 그 결과는 Table 2와 같으며, TF-IDF값이 높을수록 구간 내 속해 있는 연구들 사이에서 상대적으로 중요한 핵심 단어이다. Table 2의 키워드를 주제별로 구분하면 Table 3과 같으며, 구간별로 핵심 키워드의 주제가 다르게 분포되어 있다. 1 구간(2010~2020)에는 '단백질 및 화학물질', '질병'과 관련된 키워드가 핵심 키워드였고, 신약개발과 관련된 AI 기술 관련 연구는 상대적으로 중요성이 낮았다. 그러나 2 구간(2021~2022)에는 신약개발-AI기술 관련 연구가 핵심 키워드로, '핵산'과 '암' 관련 연구의 중요도가 높아진 것을 알 수가 있다.

Topic Modeling 결과

LDA 기반 토픽모델링 분석을 수행하기 위해서는 우선 최적의 토픽 수를 결정하는 것이 선행되어야 한다. 본 연구에서는 기존 논문들(Deveaud 2014[36], Griffiths 2004 [50], CaoJuan 2009[17], Arun 2010[5])이 제시한 모델의 적합도 지수를 비교하여 가장 적합한 토픽 수를 산출하는 하이퍼 파라미터 튜닝(Hyper-parameter tuning) 방법을 선택하였다. Arun (▲), CaoJuan (●)의 지수는 토픽이 최소가

Table 2. TF-IDF analysis results

Rank	Period 1 word	tf_idf value	Period 2 word	tf_idf value
1	seizure	0.000109	aptamer	0.000121
2	radial	0.000085	miRNA	0.000112
3	DTO	0.000080	mRNA	0.000097
4	MLM	0.000080	degron	0.000087
5	DP7	0.000075	G12C	0.000083
6	efindsite	0.000075	3CL	0.000068
7	deamidation	0.000065	fentanyl	0.000068
8	Fabs	0.000065	HNN	0.000068
9	FXR	0.000065	BMI1	0.000058
10	NAD	0.000065	CT	0.000058
11	FS	0.000060	HR	0.000058
12	RFE	0.000060	ncRNA	0.000058
13	secrete	0.000060	QC	0.000058
14	drugscore	0.000055	viscosity	0.000058
15	CLint	0.000055	CETSA	0.000053
16	pediatric	0.000055	DeepDIL	0.000053
17	ALS	0.000050	DeepPurpose	0.000053
18	CPANN	0.000050	DLBCL	0.000053
19	HSVL	0.000050	hypergraph	0.000053
20	IVIVR	0.000050	KGs	0.53

DTO: diethyltoluamide, MLM: mouse liver microsome, DP7: an antibacterial peptide, Fabs: antigen binding fragment, FXR: bile acid receptor, NAD: nicotinamide adenine dinucleotide, FS: free systemic, RFE: recursive feature elimination, CLint: intrinsic clearance, ALS: amyotrophic lateral sclerosis, CPANN: ceramides in platelet-rich plasma and nanoparticles, HSVL: hypotension-severe visual loss, IVIVR: in vivo imaging of vascular reactivity, G12C: a k-ras mutation, 3CL: a protease, HNN: hybrid neural network., BMI1: a novel cancer target protein, CT: compute tomography, HR: human resources, QC: quantum computing, CESTA: cellular thermal shift assay, DeepDIL: deep learning-power, drug-induce liver injury, DeepPurpose: a deep learning based drug repurposing and virtual screening toolkit, DLBCL: diffuse large B cell lymphoma, KGs: knowledge graphs

Table 3. LDA analysis results

Research field	Period 1	Period 2
Protein and compound	DP7, FXR, MLM, DTO, NAD, CPANN, secreted, deamidation, Fabs	fentanyl
Nucleic acid		aptamer, message(messenger), ncRNA, miRNA
Disease	FS, CLint, ALS, HSVL	
Cancer		G12C, BMI1, degrons, DLBCL
AI tool	RFE, drug score, IVIVR	HNN, QC, CETSA, DeepDIL, DeepPurpose, hypergraph, KGs,
Others	pediatric	radial, viscosity, HR, CT

되는 지점이, Deveaud (◆), Griffiths (■)의 지수는 토픽이 최대가 되는 지점을 잠재적 토픽 수로 산출한다. 이를 위해 R의 'ldatuning' 패키지를 이용하여 모형의 적합도 지수를 산출하였으며, 그 결과는 Fig. 3과 같이 나타났다. 여기서 Arun (▲), CaoJuan (●), Griffiths (■)의 지수가 일치하는 지점을 기준으로 1 구간의 토픽 수를 10개로, 2 구간의 토픽 수를 7개로 설정하였다.

이후 R의 'topicmodels' 패키지를 이용해 토픽모델링 분석을 실시하였고, Table 4와 같이 1 구간의 각 토픽과 관련된 상위 10개의 키워드를 도출하였다. 각 토픽의 키워드들은 통계적 방법으로 추론된 상관관계의 키워드들이므로 이들을 조합하면 토픽의 이름을 유추할 수 있다. 이와 같은 방법으로 각 토픽의 이름을 명명했으며, 그 결과 1 구간은 주로 머신러닝 방법인 서포트 벡터 머신(Support Vector Machine, SVM)과, 생성적 적대 신경망(Generative Adversarial Networks, GAN), 인공신경망(Artificial Neural Network, ANN) 알고리즘을 이용한 HIV치료제 및 신약개발 관련 연구가 진행되어 온 것을 알 수 있다. 그리고 합성곱 신경망(Convolutional Neural Networks, CNN)을 이용한 의료영상분석에 대한 연구가 수행되었으며, AI를 이

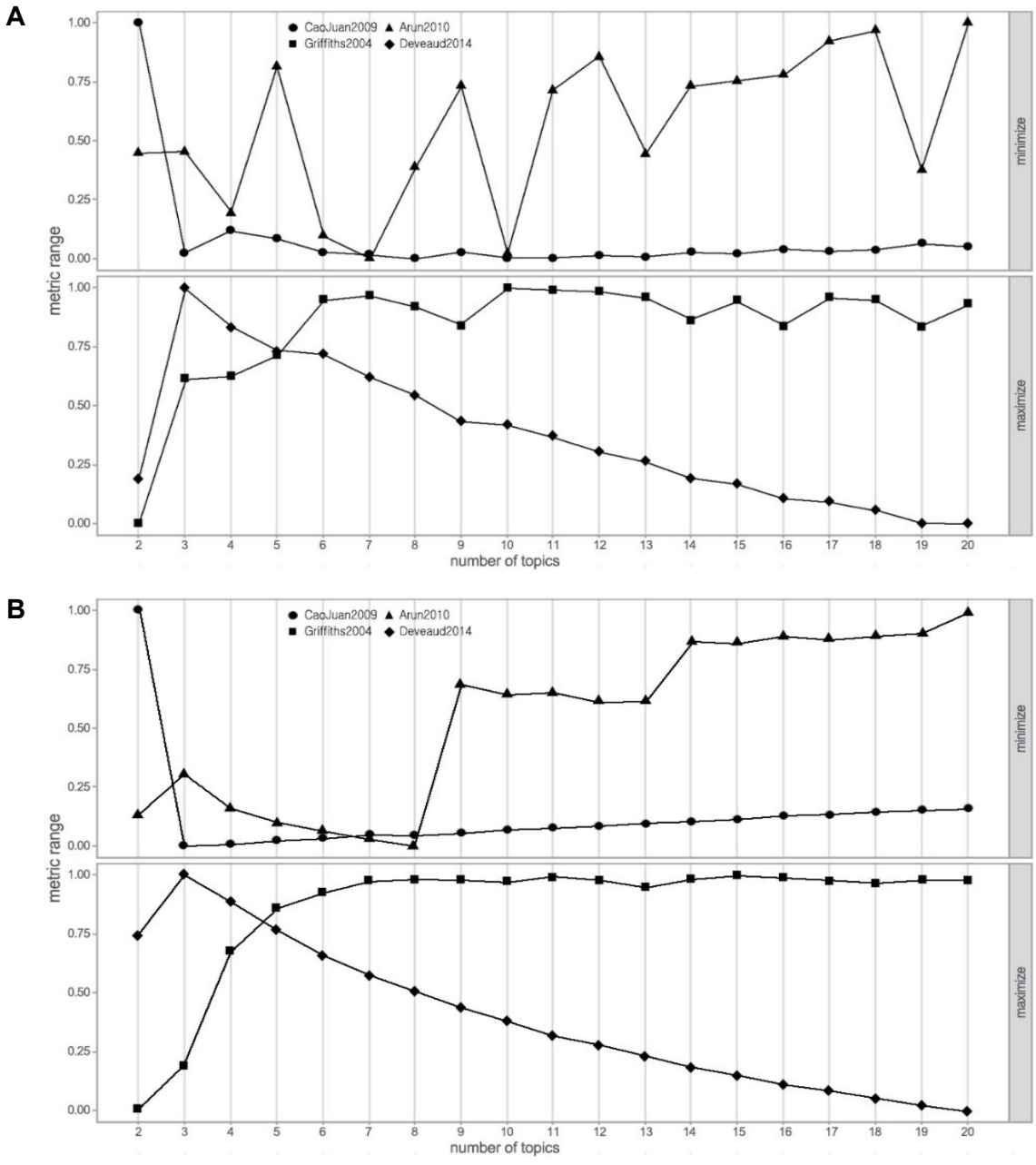

Fig. 3. Determination of the optional number of topics in LDA analysis in period 1 (A) and 2 (B).

Table 4. Main words and topic in period 1

	Main words	Topic
1	covid19, sars, cov, hemolytic, coronavirus, TCM, address, pandemic, vaccine, rapid	A study on the role of oriental medicine in Corona virus and epidemic response
2	drug, predict, model, learning, method, data, base, machine, protein, develop	Machine learning-based protein-drug interaction model development research
3	rank, SVM, pocket, sample, react, vector, search, database, shape	A study on sample classification and ranking using SVM
4	SVM, residue, semantic, QSAR, annotate, DDI, vector, pair, kinase	A study on SVM-based semantic QSAR model for DDI prediction
5	SVM, GA, hot, seizure, substrate, DTO, spots, ACT, ANN	Prediction of seizure-inducing effects of drugs listed in DTO using SVM
6	AD, ANN, constituent, ATP, generative, novo, composition, FXR, formula	AD treatment drug candidate generation study using ANN
7	SFs, HIV, noise, anti, genes, ontology, ADME, ADR, SVR, SVM	Prediction of genes for ADME/ADR of HIV using SVM
8	disorder, simulations, potent, bank, parallel, documents, electrostatic, launched, reproduce, LogP	Potent drug discovery research using molecular dynamics simulations
9	reposition, line, MS, HIV, react, metabolism, DP7, GPCR, apoptosis, Bcl	A study of new HIV therapies targeting DP7
10	deep, AI, intelligence, challenge, attention, CNN, patient, generative, novo, generation	A study on the challenges of medical image analysis using CNN

TCM: traditional Chinese medicine, SVM: support vector machine, QSAR: quantitative structure–activity relationship, DDI: drug-drug interactions, DTO: diethyltoluamide, GA: genetic algorithm, ACT-SVM: a prediction model of protein-protein interaction, ANN: artificial neural network, AD: Alzheimer's disease, SFs: scoring functions, ADRs: adverse drug reactions or events, SVR: support vector regression, DP7: an antibacterial peptide, CNN: convolutional neural networks

용해 코로나 전염병에 대응하는 동양의학의 역할과 AI와 분자 시뮬레이션을 통한 신약개발 연구들이 주요 주제로 자리 잡고 있었다.

동일한 방법으로 2 구간 토픽의 주요 키워드를 도출하고 토픽 이름을 명명하였다(Table 5). 이러한 토픽 이름 추출을 통해 2 구간의 연구 추이를 살펴본 결과 AI를 이용한 코로나 바이러스 관련 연구는 동양의학에서 방법을 찾아보려는 1구간 연구보다 구체적이고 임상 수준의 실질적 연구로 발전한 것을 알 수 있었다. 신기술-신약개발과 관련된 연구는 머신 러닝을 이용한 방법이 동일한 추세로 연구되었으나 알고리즘적 측면에서 서포트 벡터 머신(SVM)과, 생성적 적대 신경망(GAN), 합성곱 신경망(CNN) 관련된 연구가 줄었으며, 인공신경망(Artificial Neural Network, ANN), 그래프신경망(Graph Neural Network, GNN), 그라디언트 부스팅 알고리즘(Gradient Boosting Algorithm, GBM) 등을 이용한 연구가 주요 주제로 부상하였다. 그 외, AI와 CADD (Computer-Aid Drug Design)를 이용한 신약개발 관련 연구와 AI를 활용해 단백질의 특성을 이해하기 위한 연구들이 주로 수행되었다.

연구 동향과 전망

앞에서 제시된 문헌 분석 결과를 기반으로 하여 AI를 활용한 신약개발 관련 분야에서 최근 연구 동향을 살펴보면 aptamer 등의 핵산 관련 연구, 신약 재배치 관련 연구, 새로운DL 기술 적용한 연구, 항암제 개발 연구 등이 주요하게 이루어지고 있는 것으로 나타났다. 이러한 결과를 기반으로 하여 AI를 활용한 연구동향과 전망을 간략하게 정리하여 본다. 계산 기반의 접근법은 약물 개발에서 시간과 비용을 절감하고 효율성을 높일 수 있으며, 최근 많은 새로운 약물의 승인에 기여하고 있다[149]. AI를 활용한 신약개발의 응용 분야는 약물 디자인 및 모델링, 약물 스크리닝, 부작용 예측, 임상 시험 디자인 및 최적화 분야 등으로 나누어 볼 수 있다[149]. 이에 대해 좀 더 자세히 살펴보아 관련 연구 동향을 파악하고자 하며 이를 기반으로 연구 전망에 대해 제시하고자 한다. 한편 이러한 연구의 과정에서 사용되어지는 관련 tool들에 대해 생물학 연구자들에게 참고가 되기 위해 Table 6에 정리하여 제시하였다.

신약개발에 AI를 활용하기 위해서는 먼저 화학정보학에 대한 이해가 매우 필요하다. 화학정보학에서 사용되는 많은 데이터의 형식인 SDF (Structure Data Format), MDL (Molfile), PDB (Protein Data Bank), 그리고 SMILES (Simplified Molecular-Input Line-Entry System) 등을 활용하는 화학적 표현자와 화학적 지문을 구축하여 신약을 스크리

Table 5. Main words and topic in period 2

	Main words	Topic
1	drug, model, predict, learning, method, base, develop, molecule, protein, data	Drug development research using machine learning
2	Covid19, DILI, SFs, virus, ANN, Pl3K, formulation, publish, contact, dimension	DILI risk assessment study of COVID-19 treatment using ANN
3	CPI, effect, DTI, toxic, abnormal, root, highlighting, fragment, normal, PCA	A study on the utilization of DTI and CPI for new drug development
4	aptamer, phosphorylation, CADD, fluorescent, alert, expedite, severity, rational, solvent, recommendation	Development of aptamer-based fluorescence sensor using CADD and research on rapid phosphorylation detection method
5	similar, neighbor, graph, node, topology, QSP, synthetic, GNN, DTA, covalent	A study of a new method to predict drug-target affinity using GNN
6	physical, proteome, effort, future, wet, correct, experts, expertise, gained, bound	Efforts to understand the physical properties of proteomes
7	background, typical, microbial, mining, GBM, graph, multi, manual, chapter, view	A study on the use of GBM for molecular graph data mining

DILI: drug-induce liver injury, SFs: score functions, ANN: artificial neural network, CPI: compound-protein interaction, DTI: drug-target interaction, CADD: computer-aid drug design, QSP: quantitative system pharmacology, GNN: graph neural network, DTA: drug-target affinity, GBM: gradient boosting algorithm

Table 6. Artificial intelligence tools for new drug development*

Research field	Tools
Homology modeling prediction	MODELLER, Swiss model, Phyre & Phyre2, 3D-JIGSAW, HHpred, RaptorX, ESyPred3D, MOE, Yasara, FoldX, BhageerathH
Threading modeling prediction	Muster, GenTHREADER, I-TASSER, DescFold
Ab initio modeling prediction	QUARK, Rosetta/Robetta, I-TASSER, CABS-FOLD, EVfold
Binding site prediction	FPOCKET, SURFNET, Q-SITEFINDER, DoGSite, Scorer server, CASTp, BiteNet, Metapocket, DEPTH, LISE, MSpocket, Epock, TRAnsient Pockets in Proteins, POVME, POOL, MetalDetector
Stochastic search algorithms	AutoDock, Gold, PRO_LEADS, EADock, LigandFit, ICM, Molegro Virtual Docker, CDocker, GlamDock, PLANTS, MolDock
Systematic search algorithms	eHiTS, FRED, Surflex－Dock, DOCK, GLIDE EUDOC, FlexX, Hammerhead, Flog, SLIDE, ADAM
Docking	AutoDock, AutoDock Vina, BetaDock, Blaster, DARWIN, DOCK, DockVision, DOLINA, EADock, FlexX, FlexAID, FLIPDock, GalaxyPepDock, GEMDOCK, Glide, GOLD, GPCRautomodel, idTarget, LeDock, LightDock, MedusaDock 2.0, MOLS 2.0, SwissDock
Graphical display	UCSF Chimera, JSmol, Jmol, RasMol, BALL, Phymol, VMD
Molecular dynamics simulation	AMBER, CHARMM, OPLS, GROMOS, Coarse grained
Small molecule database	DrugBank, PubChem, BindingDB, BindingMOAD, ChEMBdb, ChemSpider
QSAR	OECD QSAR Toolbox, CORAL, PharmQSAR, AutoQSAR, GUSAR
Pharmacophore modeling	MolSign, LigandScout, Catalyst, CASTSlight2, PharmMapper, Pharmer, Phase, ZincPharmer
AI	Alphafold, Chemputer, Chemical VAE, DeltaVina, Hit Dexter, InnerOuterRNN, JunctionTree VAE, NNScore, ORGANIC, Open Drug Discovery Toolkit, PPB2, QML, REINVENT, XenoSite, SMARTCyp, DIA-NN
DL	Tensorflow, Pytorch, Scikit learn, MXNet, Gluon, Deep Docking, Deep Chem, DeepTox, DeepNeural Net QSAR, PotentialNet, Conv_qsar_fast, Neural graph fingerprint, PADME, Tox_(R)CNN

*This table is adapted from a published review [149].

닝 하는 모델링 연구가 이루어지고 있다[25, 90, 93, 100, 121]. 이러한 과정 속에서 화합물의 정보를 고도화하기 위한 DNA-Encoded Chemical Libraries (DELs) 관련 연구가 최근에 많이 진행되고 있다[126]. 그리고 다양한 관점과 화학 및 생물학적 구조의 이론적 프레임워크를 제공하기 위한 여러가지 모델들이 개발되어지고 있다[8, 95, 96, 129, 149, 156].

DELs과 ML을 결합하여 약물 개발에서 출발 화합물을 식별하고 예측 독성을 개선하는데 유용한 정보를 제공할 수 있다[43, 81, 96, 100, 148, 153, 159, 163, 173]. 충분한 데이터가 있으면 ML 모델은 광범위한 화학 공간에서 정확한 예측을 할 수 있으며, 다른 프로젝트에서도 재사용하고 확장할 수 있을 것이다. 이는 약물 개발 초기에 안전한 측면에서 비용을 절감하면서 유용한 독성 예측 모델을 제공할 수 있으리라 본다.

한편 약물 발견을 가속화하기 위한 multimodal deep generative model들을 통해 적대적 생성 모델의 잠재 공간에서 새로운 분자를 탐색하고 개선하는 방법을 제시하고 있다[51]. 이러한 모델들은 새로운 치료제 개발에 활용될 수 있는 고품질의 분자 구조를 설계할 수 있는 능력을 보여주고 있다. 이러한 노력을 통해 AI는 기존 논문, 특허 정보와 유전체, 단백질체 정보 등을 기반으로 하여 타겟 단백질을 발굴 및 검증을 쉽게 하고 기존 약물의 활성도 향상시키는데 활용되어지고 있다[2, 10, 21, 25, 33, 51, 58, 72, 89, 134, 136, 142, 152].

단백질 3차원 구조 규명 연구는 신약개발에 있어 가장 어려운 부분 중에 하나이었다. 신약개발에 있어 후보물질을 발굴하기 위해서는 단백질 3차원 구조를 규명하는 것은 매우 중요한 부분이다. 그런데 최근 구글이 단백질의 3차원 구조 규명에 AI를 활용한 알파폴드를 발표하였다. 알파폴드는 생물학적 데이터와 AI를 결합한 강력한 알고리즘이며, 다양한 분야에서 혁신적인 응용 가능성을 가지고 있다. 이러한 구조 유추의 방법은 약물설계에 있어서 핵심적인 기술로 자리잡고 있다[111, 112].

분자 도킹은 약물 발견 초기에 중요한 역할을 하는데, 스코어링 기능과 ML의 발전으로 단백질-리간드 상호작용을 예측하고 약물 스크리닝부터 최적화까지 이루어지고 있다[75, 94, 110, 119, 132, 176]. Computer-Aided Drug Design (CADD)은 AI, ML, DL과 결합하여 약물 개발 과정을 가속화하고 비용을 절감하는 중요한 역할을 한다[8, 95, 149, 156]. 이 접근 방식은 생물학적 데이터 처리와 *in silico* 도구를 통해 약물 후보물질을 탐색하고 발견 과정을 지원한다. 최근 연구에서는 AI 기계 학습 알고리즘이 새로운 약물 유사 분자를 자동으로 생성할 수 있는 능력을 보여주며[136], 이는 약물 발견과정을 혁신하고 탐색을 효율적으로 만들어 준다. 다양한 모델 프레임워크와 입력 형식의 제안으로 AI 알고리즘의 성능을 향상시키는 노력이 이루어지고 있다[74, 92].

기계 학습(ML)은 독성, 흡수, 약물간 상호작용, 발암성, 분포 등 많은 약물의 물리화학적 특성을 효과적으로 모델링할 수 있는 Quantitative Structure Activity Relationship (QSAR) 기법을 포함하여 많은 잠재력을 갖고 있다[44, 47, 56, 63, 105, 116, 119]. 이러한 예측 도구와 모델은 최근 좋은 정확도를 보여주었으며, 더 많은 관련 입력 데이터, 매개변수 및 적절한 알고리즘의 사용을 통해 정확도를 더욱 향상시킬 수 있다. 이러한 과정에서 AI는 신약개발의 다양한 예측의 정확성을 높이는데 사용되고 있다. 이에 ML 알고리즘을 사용하여 수백만 개의 잠재적 조합 중에서 새로운 상호작용 약물을 식별하는 데 활용되고 있다[59, 71, 88, 107, 130].

AI를 활용한 약물 다중 상호작용 예측은 다중 약물의 조합과 안전성에 중요한 역할을 하는데, 약물-약물, 약물-음식, 약물-미생물 상호작용을 체계적으로 분석하고 모델링함으로써 실용적이고 안전한 약물 사용에 기여할 수 있다. 특히 약물-약물 상호작용을 예측하는 데에는 P450 대사 효소, 약물 유사성, 약물 타겟 기반의 ML 모델이 주로 사용된다[15, 29, 80, 85, 96, 127, 160]. 약물의 대사 반응은 중독성, 부작용, 효능 저하 등에 영향을 미치며, 병용 약물 요법에서는 약물 대사 상호작용이 중요하다. ML 알고리즘들은 대사체 예측과 약물 간 상호작용 예측에 사용되며, 이는 약물 개발과 연구에 효과적인 도구로 활용될 수 있다. DL과 계산 약물 개발 분야의 발전은 이러한 연구에서 유용한 결과를 도출하고 있다[9, 42, 52, 102, 105, 144, 151].

약물 재배치는 승인된 또는 연구 중인 약물의 새로운 치료 용도를 찾기 위한 전략으로, 컴퓨터 기반의 다양한 방법과 특히 DL이 타겟 단백질 발굴과 약물 재배치에 적용되어 효율성과 성공률을 향상시키고 있다[16, 38, 52, 73, 124, 166]. AI와 ML을 활용한 약물 재배치의 응용은 다양하게 이루어지고 있으며, 약물 재활용을 통한 혁신을 가속화하기 위한 다양한 경로를 제공하고 있다. 이러한 노력이 제약회사들의 새로운 바이오 의약 시장 개발과 성장에 기여하고 있다[4].

AI 기반의 DL 기술은 약물 개발의 모든 단계에서 활용되며, 약물-표적 상호작용, 약물-약물 유사성 상호작용, 약물 감수성 및 반응성, 약물 부작용 예측 등 다양한 응용 분야에 적용된다. 제약 회사들을 위해 개발된 ML 모델이 여러 종류로 제공되어져 있고, 신약 개발에 효과적인 의사 결정을 돕는 맞춤형 약물 추천과 성공 확률 예측을 수행하고 있다. 콘텐츠 기반 필터링과 다양한 접근법을 결합하여 COVID-19 백신 개발 기업들의 성공 확률을 예측하고, 모델의 점수가 높을수록 임상 단계 진행이 많아진다는 입증 결과를 보여주고 있다. 이 모델들은 과학적 및 산업적으로 신약 개발의 합리적인 의사 결정을 지원하

고 있다[4].

AI을 통한 신약개발의 장점은 신속하고 효율적인 검색 및 분석, 대규모 데이터 처리 능력, 비용과 시간 절감, 개인 맞춤형 치료 개발 가능성 등으로 볼 수 있다[4]. AI는 대규모 데이터를 활용하여 약물 스크리닝을 수행하고, 새로운 약물을 디자인하는데 도움을 주고 있다. 기존의 방법에 비해 정확한 예측을 통한 유망한 후보 물질을 찾아낼 수 있게 하고 있다. 그리고 AI는 약물의 부작용을 예측하는 모델을 구축함을 통해 이를 사전에 식별하고 최소화하는데 도움을 줄 수 있을 것으로 본다.

한편 AI 기술은 의료 이미지, 유전자 데이터, 환자의 건강 정보 등을 기반으로 질병을 진단하고 예측하는 데 사용될 수 있다. 이를 통해 초기 진단과 조기 예방이 가능해지며, 개인 맞춤형 치료 계획을 수립할 수 있으리라 본다. AI는 환자의 유전자 정보, 의료 기록, 생활 양식 등을 분석하여 개인에게 맞춤형 치료 계획을 제시할 수 있다. 이를 통해 개인의 특성과 상황을 고려한 최적의 치료 방법을 찾을 수 있을 것이다. 그리고 AI는 임상시험의 디자인 및 집단 분류를 최적화할 수 있다. 적은 수의 환자로도 효과적인 결과를 얻을 수 있으며, 시험 기간을 단축시키고 비용을 절감할 수 있으리라 본다. AI 기술은 신약 개발 과정의 각 단계를 가속화할 수 있다. 이를 통해 개발 비용과 시간을 절감하고, 빠르게 안전하고 효과적인 치료제를 시장에 도입할 수 있을 것으로 예상한다.

하지만 AI를 통한 신약개발은 아직 초기 단계이기에 몇가지의 문제점이 존재한다. 관련 의료 데이터의 부족, 모델의 신뢰성, 규제적인 측면 등이 문제가 될 수 있으리라 본다. 따라서 신약개발에 AI를 적용할 때에는 이러한 문제를 고려하며 신중한 접근이 필요하다. 예를 들어 단백질의 구조 예측 분야에서 현재 개발되어진 AI 기반 단백질 3차원 구조를 결정하는 측면에 있어 아직 일부 막단백질 구조 예측은 제한이 있고, 구조의 다양한 ensembles을 제공하지 않으며, allosteric 신약 등의 작동 메커니즘을 밝혀낼 수 없다는 단점을 가지고 있다.

AI를 통한 신약개발은 잠재력이 매우 높지만, 여전히 AI 통한 신약개발에 있어 해결해야 할 도전 과제가 존재한다. 이를 살펴보면, 데이터 품질과 양의 문제 해결, 해석 가능성과 신뢰성 향상, 윤리적 고려 사항에 대한 해결, 규제와 법적 측면의 해결 과제 등이 있을 수 있다. 신약개발을 위한 고성능 AI 모델을 구축하기 위해서는 대량의 고품질 바이오와 의료 데이터가 필요하다. 하지만 양질의 데이터는 현재 부족한 현실이다. 이에 대한 해결이 AI를 활용한 신약개발을 가속화하기 위해 필요로 하는 분야이다. 그리고 데이터의 신뢰성과 안전성에 대한 보증이 이루어져야 하리라 본다. 한편 의약품 개발은 매우 복잡한 과정으로 이에 대한 해석이 복잡하게 이루어지기에 이에 대한 해석을 인간 전문가의 지식과 함께 이루어져야 하리라 본다. 그리고 AI를 통해 개발되어진 약에 대한 안전성 검증, 데이터의 개인 정보 보호 등의 윤리적인 측면에 대한 엄중한 고려가 필요하다고 보며, 이에 대한 법적 측면에 대한 논의와 규제도입에 대한 검토가 필요하다고 본다.

AI를 통한 신약개발의 미래 연구는 강화학습을 활용한 약물 디자인 추구, 통합적인 데이터 분석 증가, DL 활용 모델 개발 가속화, 개인맞춤형 치료에 적용 사례 확대, 개인정보 보호 강화 기술 개발 등을 예상할 수 있다. 약물 디자인 분야에서 강화학습을 적용하여 약물 분자 구조를 최적화하고 타겟 약물을 개발하는 연구가 수행되어질 것으로 예상한다. 다양한 유형의 바이오와 의료 데이터가 생성되어지고 있어 이에 대한 종합적인 분석을 수행하고 이해를 도모하는 연구가 증가되리라 본다. DL을 활용한 생성 모델은 다양한 분야에서 좋은 성과를 보여주고 있어 신약개발 분야에도 이의 적용이 가속화되리라 예측한다. 향후 여러가지의 개인 정보를 기초로 하여 AI를 활용한 개인 맞춤형 치료를 추구하는 연구가 증대되어질 것으로 예상된다. 마지막으로 개인정보 보호가 매우 중요해지고 있어 이의 보호기술 발전을 통한 신약개발을 추구하려는 노력이 많이 이루어질 것으로 본다.

결 론

본 논문은 학제간 연구동향 분석방법론을 제시하고, 생화학 관련 연구동향을 리뷰하여 연구 흐름과 새로운 연구 방향을 제시하기 위한 목적을 갖고 있다. 따라서 본고는 TF-IDF분석결과와 LDA분석결과를 각각 제시하여 AI를 활용한 신약개발 연구의 일반적인 흐름을 파악하고자 했다. 이 리뷰의 인공지능 활용 신약개발 연구 동향 파악을 위한 텍스트 마이닝 절차는 세 가지 주요 단계로 구성된다. 먼저 분석할 인공지능 활용 신약개발 연구 관련 대규모 텍스트 데이터를 코퍼스로 얻었다. 둘째, 이 코퍼스를 가지고 데이터 정리, 정규화, 토큰화 등의 텍스트 전처리를 통해 데이터를 분석에 적합하도록 정리하였다. 셋째, 전처리된 데이터를 이용하여 TF-IDF 분석과 주제 모델링 분석을 수행하였다. 2021년 이전(1기)과 2021년 이후(2기)로 나누어 토픽 모델링을 수행하였다. 각 기간별 적정 토픽 수를 산정하기 위해 앞서의 연구자들이 제안한 방법을 사용하여 적정 토픽 수를 산정하였다.

이러한 인공지능 활용 신약개발의 연구 동향에 대한 분석과 정리 과정을 통해 본고에서 제시할 수 있는 시사점은 다음과 같이 정리할 수 있다. 첫번째로 2010년부터 2022년까지 '인공지능을 이용한 신약 개발 연구' 관련 논문의 양적 추이는 2017년까지는 년 100건 미만이었으나 2021년부터 연간 500건 이상으로 급증하고 있다는 점이다. 최근의 인공지능 관련 컴퓨터 기술의 발전이 다양한 방향에서 이루어지고 있어 AI 활용 신약개발 연구도 매우

급격하게 증가되는 것으로 본다. 두번째 시사점은 1 구간의 경우 대체로 ML을 활용한 신약개발 연구가 주요하게 이루어졌으나 2구간의 경우에 DL을 활용한 신약개발 연구가 많아지고 있는 것이다. 이 또한 인공지능 관련 컴퓨터 기술의 발전이 미치는 영향으로 볼 수 있다. 세번째 시사점은 이전의 전통적인 신약개발 연구 방법 이외에 약물의 재배치 연구, 항암제 개발 연구, 임상실험에 인공지능 적용 연구 등과 같이 인공지능 적용 영역이 많이 확대되고 있다는 점이다. 이는 관련 신약개발 관련 많는 바이오와 의료 데이터의 축적을 통해 인공지능의 적용 영역이 증가하고 있는 것으로 판단할 수 있다. 네번째 시사점은 AI 활용한 미래 신약개발 연구는 좀 더 많아지고 신뢰가 보장된 바이오와 의료 데이터의 축적과 인공지능 기술의 발전이 가속화가 함께 이루어진다면 종합적인 데이터 분석이 가능해져 좀 더 신뢰성 높은 인공지능 활용 신약이 개발되어질 것으로 예측된다. 마지막으로 위와 같은 기술적 진보가 담보된다면, 인공지능 기술을 활용한 개인 맞춤형 의료와 정밀 의료로 발전되어지는 것을 바라볼 수 있으리라 본다.

한편 본 리뷰에서 수행한 연구 방법은 인공지능 활용 신약개발 연구동향을 일반적인 관점에서 도출하기는 하나, 앞으로 좀 더 심도 있는 의미를 도출하기 위해TF-IDF를 통해 핵심 키워드를 파악하고, 이 키워드를 중심으로 문헌을 재 검색하여 토픽모델링을 수행하면 두 분석기법을 연동한 정확한 분석이 이루어질 것으로 본다. 향후 좀 더 심도 있는 분석 결과를 도출하기 위한 다양한 접근의 노력을 기대해 본다.

감사의 글

본 과제(논문)는 2023년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다(2021RIS-001).

The Conflict of Interest Statement

The authors declare that they have no conflicts of interest with the contents of this article.

References

1. Abdelbasset, W. K., Elsayed, S. H., Alshehri, S., Huwaimel, B., Alobaida, A., Alsubaiyel, A. M., Alqahtani, A. A., El Hamd, M. A., Venkatesan, K., AboRas, K. M. and Abourehab, M. A. S. 2022. Development of gbrt model as a novel and robust mathematical model to predict and optimize the solubility of decitabine as an anti-cancer drug. *Molecules* **27**, 5676.
2. Abubaker Bagabir, S., Ibrahim, N. K., Abubaker Bagabir, H. and Hashem Ateeq, R. 2022. Covid-19 and artificial intelligence: Genome sequencing, drug development and vaccine discovery. *J. Infect. Public Health* **15**, 289-296.
3. Ahmed, F., Kang, I. S., Kim, K. H., Asif, A., Rahim, C. S. A., Samantasinghar, A., Memon, F. H. and Choi, K. H. 2023. Drug repurposing for viral cancers: A paradigm of machine learning, deep learning, and virtual screening-based approaches. *J. Med. Virol.* **95**, e28693.
4. An, Q., Rahman, S., Zhou, J. and Kang, J. J. 2023. A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges. *Sensors (Basel)* **23**, 4178.
5. Arun, R., Suresh, V., Madhavan, C. E. V. and Murty, M. N. 2010. On finding the natural number of topics with latent dirichlet allocation: Some observations. *Lect. Notes Artif. Int.* **6118**, 391-402.
6. Badwan, B. A., Liaropoulos, G., Kyrodimos, E., Skaltsas, D., Tsirigos, A. and Gorgoulis, V. G. 2023. Machine learning approaches to predict drug efficacy and toxicity in oncology. *Cell Rep. Methods* **3**, 100413.
7. Bao, L. 2005. Identifying genes related to chemosensitivity using support vector machine. *Methods Mol. Med.* **111**, 233-240.
8. Bao, L., Wang, Z., Wu, Z., Luo, H., Yu, J., Kang, Y., Cao, D. and Hou, T. 2023. Kinome-wide polypharmacology profiling of small molecules by multi-task graph isomorphism network approach. *Acta Pharm. Sin. B* **13**, 54-67.
9. Baskin, II, Winkler, D. and Tetko, I. V. 2016. A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discov.* **11**, 785-795.
10. Baylon, J. L., Cilfone, N. A., Gulcher, J. R. and Chittenden, T. W. 2019. Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. *J. Chem. Inf. Model* **59**, 673-688.
11. Bhalla, S. and Lagana, A. 2022. Artificial intelligence for precision oncology. *Adv. Exp. Med. Biol.* **1361**, 249-268.
12. Bian, Y. and Xie, X. Q. 2022. Artificial intelligent deep learning molecular generative modeling of scaffold-focused and cannabinoid cb2 target-specific small-molecule sublibraries. *Cells* **11**, 915.
13. Bird, A., Oakden-Rayner, L., McMaster, C., Smith, L. A., Zeng, M., Wechalekar, M. D., Ray, S., Proudman, S. and Palmer, L. J. 2022. Artificial intelligence and the future of radiographic scoring in rheumatoid arthritis: A viewpoint. *Arthritis Res. Ther.* **24**, 268.
14. Blei, D., Carin, L. and Dunson, D. 2010. Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. *IEEE Signal Process. Mag.* **27**, 55-65.
15. Burton, J., Ijjaali, I., Petitet, F., Michel, A. and Vercauteren, D. P. 2009. Virtual screening for cytochromes p450: Successes of machine learning filters. *Comb. Chem. High Throughput Screen.* **12**, 369-382.
16. Canizares-Carmenate, Y., Mena-Ulecia, K., MacLeod Carey, D., Perera-Sardina, Y., Hernandez-Rodriguez, E. W.,

Marrero-Ponce, Y., Torrens, F. and Castillo-Garit, J. A. 2022. Machine learning approach to discovery of small molecules with potential inhibitory action against vasoactive metalloproteases. *Mol. Divers.* **26**, 1383-1397.
17. Cao, J., Xia, T., Li, J. T., Zhang, Y. D. and Tang, S. 2009. A density-based method for adaptive lda model selection. *Neurocomputing* **72**, 1775-1781.
18. Cao, Q., Cheng, X. and Liao, S. Y. 2023. A comparison study of topic modeling based literature analysis by using full texts and abstracts of scientific articles: A case of covid-19 research. *Libr. Hi Tech.* **41**, 543-569.
19. Carter, R., Luchini, A., Liotta, L. and Haymond, A. 2019. Next generation techniques for determination of protein-protein interactions: Beyond the crystal structure. *Curr. Pathobiol. Rep.* **7**, 61-71.
20. Caruso, F. P., Scala, G., Cerulo, L. and Ceccarelli, M. 2021. A review of covid-19 biomarkers and drug targets: resources and tools. *Brief. Bioinform.* **22**, 701-713.
21. Cavalla, D. and Crichton, G. 2023. Drug repurposing: Known knowns to unknown unknowns - network analysis of the repurposome. *Drug Discov. Today* **28**, 103639.
22. Ceccarelli, F., Natalucci, F., Picciariello, L., Ciancarella, C., Dolcini, G., Gattamelata, A., Alessandri, C. and Conti, F. 2023. Application of machine learning models in systemic lupus erythematosus. *Int. J. Mol. Sci.* **24**, 4514.
23. Chang, S. S., Huang, H. J. and Chen, C. Y. 2011. Two birds with one stone? Possible dual-targeting h1n1 inhibitors from traditional chinese medicine. *PLoS Comput. Biol.* **7**, e1002315.
24. Chang, W. T., Liu, C. F., Feng, Y. H., Liao, C. T., Wang, J. J., Chen, Z. C., Lee, H. C. and Shih, J. Y. 2022. An artificial intelligence approach for predicting cardiotoxicity in breast cancer patients receiving anthracycline. *Arch. Toxicol.* **96**, 2731-2737.
25. Chen, B., Garmire, L., Calvisi, D. F., Chua, M. S., Kelley, R. K. and Chen, X. 2020. Harnessing big 'omics' data and ai for drug discovery in hepatocellular carcinoma. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 238-251.
26. Chen, H., Kogej, T. and Engkvist, O. 2018. Cheminformatics in drug discovery, an industrial perspective. *Mol. Inform.* **37**, e1800041.
27. Chen, J. Q., Chen, H. Y., Dai, W. J., Lv, Q. J. and Chen, C. Y. 2019. Artificial intelligence approach to find lead compounds for treating tumors. *J. Phys. Chem. Lett.* **10**, 4382-4400.
28. Chen, Z., Zhao, M., You, L., Zheng, R., Jiang, Y., Zhang, X., Qiu, R., Sun, Y., Pan, H., He, T., Wei, X., Chen, Z., Zhao, C. and Shang, H. 2022. Developing an artificial intelligence method for screening hepatotoxic compounds in traditional chinese medicine and western medicine combination. *Chin. Med.* **17**, 58.
29. Cheng, F. and Zhao, Z. 2014. Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *J. Am. Med. Inform. Assoc.* **21**, e278-286.
30. Coker, E. A., Stewart, A., Ozer, B., Minchom, A., Pickard, L., Ruddle, R., Carreira, S., Popat, S., O'Brien, M., Raynaud, F., de Bono, J., Al-Lazikani, B. and Banerji, U. 2022. Individualized prediction of drug response and rational combination therapy in nsclc using artificial intelligence-enabled studies of acute phosphoproteomic changes. *Mol. Cancer Ther.* **21**, 1020-1029.
31. Cong, Y., Chan, Y. B., Phillips, C. A., Langston, M. A. and Ragan, M. A. 2017. Robust inference of genetic exchange communities from microbial genomes using tf-idf. *Front. Microbiol.* **8**, 21.
32. Cong, Y., Chan, Y. B. and Ragan, M. A. 2016. Exploring lateral genetic transfer among microbial genomes using tf-idf. *Sci. Rep.* **6**, 29319.
33. Cova, T., Vitorino, C., Ferreira, M., Nunes, S., Rondon-Villarreal, P. and Pais, A. 2022. Artificial intelligence and quantum computing as the next pharma disruptors. *Methods Mol. Biol.* **2390**, 321-347.
34. Cui, Q., Lu, S., Ni, B., Zeng, X., Tan, Y., Chen, Y. D. and Zhao, H. 2020. Improved prediction of aqueous solubility of novel compounds by going deeper with deep learning. *Front. Oncol.* **10**, 121.
35. Das, S., Babu, A., Medha, T., Ramanathan, G., Mukherjee, A. G., Wanjari, U. R., Murali, R., Kannampuzha, S., Gopalakrishnan, A. V., Renu, K., Sinha, D. and George Priya Doss, C. 2023. Molecular mechanisms augmenting resistance to current therapies in clinics among cervical cancer patients. *Med. Oncol.* **40**, 149.
36. Deveaud, R., SanJuan, E. and Bellot, P. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique* **17**, 61-84.
37. Dornick, C., Kumar, A., Seidenberger, S., Seidle, E. and Mukherjee, P. 2021. Analysis of patterns and trends in covid-19 research. *Procedia Comput. Sci.* **185**, 302-310.
38. El-Behery, H., Attia, A. F., El-Fishawy, N. and Torkey, H. 2022. An ensemble-based drug-target interaction prediction approach using multiple feature information with data balancing. *J. Biol. Eng.* **16**, 21.
39. Elemento, O., Leslie, C., Lundin, J. and Tourassi, G. 2021. Artificial intelligence in cancer research, diagnosis and therapy. *Nat. Rev. Cancer* **21**, 747-752.
40. Elkhader, J. and Elemento, O. 2022. Artificial intelligence in oncology: From bench to clinic. *Semin. Cancer Biol.* **84**, 113-128.
41. Fan, K., Cheng, L. and Li, L. 2021. Artificial intelligence and machine learning methods in predicting anti-cancer drug combination effects. *Brief Bioinform.* **22**, bbab271.
42. Feng, H., Gao, K., Chen, D., Shen, L., Robison, A. J., Ellsworth, E. and Wei, G. W. 2022. Machine learning analysis of cocaine addiction informed by dat, sert, and net-based interactome networks. *J. Chem. Theory Comput.* **18**, 2703-2719.
43. Galati, S., Di Stefano, M., Martinelli, E., Macchia, M., Martinelli, A., Poli, G. and Tuccinardi, T. 2022. Venompred: A machine learning based platform for molecular toxicity predictions. *Int. J. Mol. Sci.* **23**, 2105.
44. Gaurav, A., Agrawal, N., Al-Nema, M. and Gautam, V.

2022. Computational approaches in the discovery and development of therapeutic and prophylactic agents for viral diseases. *Curr. Top. Med. Chem.* **22**, 2190-2206.
45. Gerdes, H., Casado, P., Dokal, A., Hijazi, M., Akhtar, N., Osuntola, R., Rajeeve, V., Fitzgibbon, J., Travers, J., Britton, D., Khorsandi, S. and Cutillas, P. R. 2021. Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. *Nat. Commun.* **12**, 1850.
46. Gimeno, M., Sada Del Real, K. and Rubio, A. 2023. Precision oncology: A review to assess interpretability in several explainable methods. *Brief. Bioinform.* **24**, bbad200.
47. Goller, A. H., Kuhnke, L., Ter Laak, A., Meier, K. and Hillisch, A. 2022. Machine learning applied to the modeling of pharmacological and admet endpoints. *Methods Mol. Biol.* **2390**, 61-101.
48. Gong, J. N., Zhao, L., Chen, G., Chen, X., Chen, Z. D. and Chen, C. Y. 2021. A novel artificial intelligence protocol to investigate potential leads for diabetes mellitus. *Mol. Divers* **25**, 1375-1393.
49. Gorostiola Gonzalez, M., Janssen, A. P. A., IJzerman, A. P., Heitman, L. H. and van Westen, G. J. P. 2022. Oncological drug discovery: Ai meets structure-based computational research. *Drug Discov. Today* **27**, 1661-1670.
50. Griffiths, T. L. and Steyvers, M. 2004. Finding scientific topics. *Proc. Natl. Acad. Sci. USA.* **101 Suppl 1**, 5228-5235.
51. Grisoni, F. and Schneider, G. 2019. De novo molecular design with generative long short-term memory. *Chimia (Aarau)* **73**, 1006-1011.
52. Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K. and Kumar, P. 2021. Artificial intelligence to deep learning: Machine intelligence approach for drug discovery. *Mol. Divers* **25**, 1315-1360.
53. Gupta, R. R. 2022. Application of artificial intelligence and machine learning in drug discovery. *Methods Mol. Biol.* **2390**, 113-124.
54. He, X., Zhao, L., Zhong, W., Chen, H. Y., Shan, X., Tang, N. and Chen, C. Y. 2020. Insight into potent leads for alzheimer's disease by using several artificial intelligence algorithms. *Biomed. Pharmacother.* **129**, 110360.
55. Heikel, E. and Espinosa-Leal, L. 2022. Indoor scene recognition via object detection and tf-idf. *J. Imaging* **8**, 209.
56. Hermansyah, O., Bustamam, A. and Yanuar, A. 2021. Virtual screening of dipeptidyl peptidase-4 inhibitors using quantitative structure-activity relationship-based artificial intelligence and molecular docking of hit compounds. *Comput. Biol. Chem.* **95**, 107597.
57. Hu, F., Wang, L., Hu, Y., Wang, D., Wang, W., Jiang, J., Li, N. and Yin, P. 2021. A novel framework integrating ai model and enzymological experiments promotes identification of sars-cov-2 3cl protease inhibitors and activity-based probe. *Brief. Bioinform.* **22**, bbab301.
58. Hulsen, T. 2022. Literature analysis of artificial intelligence in biomedicine. *Ann. Transl. Med.* **10**, 1284.
59. Hung, T. N. K., Le, N. Q. K., Le, N. H., Van Tuan, L., Nguyen, T. P., Thi, C. and Kang, J. H. 2022. An ai-based prediction model for drug-drug interactions in osteoporosis and paget's diseases from smiles. *Mol. Inform.* **41**, e2100 264.
60. Iftikhar, S., Karim, A. M., Karim, A. M., Karim, M. A., Aslam, M., Rubab, F., Malik, S. K., Kwon, J. E., Hussain, I., Azhar, E. I., Kang, S. C. and Yasir, M. 2023. Prediction and interpretation of antibiotic-resistance genes occurrence at recreational beaches using machine learning models. *J. Environ. Manage.* **328**, 116969.
61. Ishii, S., Takamatsu, M., Ninomiya, H., Inamura, K., Horai, T., Iyoda, A., Honma, N., Hoshi, R., Sugiyama, Y., Yanagitani, N., Mun, M., Abe, H., Mikami, T. and Takeuchi, K. 2022. Machine learning-based gene alteration prediction model for primary lung cancer using cytologic images. *Cancer Cytopathol.* **130**, 812-823.
62. Jamal, S. and Scaria, V. 2013. Cheminformatic models based on machine learning for pyruvate kinase inhibitors of leishmania mexicana. *BMC Bioinformatics* **14**, 329.
63. Jayaprakash, V., Saravanan, T., Ravindran, K., Prabha, T., Selvaraj, J., Jayapalan, S., Chaitanya, M. and Sivakumar, T. 2023. Relevance of machine learning to predict the inhibitory activity of small thiazole chemicals on estrogen receptor. *Curr. Comput. Aided Drug Des.* **19**, 37-50.
64. Ji, Y. A., Nam, S. J., Kim, H. G., Lee, J. and Lee, S. K. 2018. Research topics and trends in medical education by social network analysis. *BMC Med. Educ.* **18**, 222.
65. Jiang, J., Ouyang, D. and Williams, R. O. 3rd. 2023. Predicting glass-forming ability of pharmaceutical compounds by using machine learning technologies. *AAPS PharmSciTech* **24**, 103.
66. Kang, H., Yu, Z. and Gong, Y. 2017. Initializing and growing a database of health information technology (hit) events by using tf-idf and biterm topic modeling. *AMIA Annu. Symp. Proc.* **2017**, 1024-1033.
67. Kang, H. J., Han, J. and Kwon, G. H. 2021. Determining the intellectual structure and academic trends of smart home health care research: Coword and topic analyses. *J. Med. Internet Res.* **23**, e19625.
68. Karim, M., Saad Missen, M. M., Umer, M., Fida, A., Eshmawi, A. A., Mohamed, A. and Ashraf, I. 2022. Comprehension of polarity of articles by citation sentiment analysis using tf-idf and ml classifiers. *PeerJ Comput. Sci.* **8**, e1107.
69. Kaushal, K., Sarma, P., Rana, S. V., Medhi, B. and Naithani, M. 2022. Emerging role of artificial intelligence in therapeutics for covid-19: A systematic review. *J. Biomol. Struct. Dyn.* **40**, 4750-4765.
70. Kaushik, A. C., Li, M., Mehmood, A., Dai, X. and Wei, D. Q. 2021. Acps: An accurate bioinformatics tool for precision-based anti-cancer peptide generation via omics data. *Chem. Biol. Drug Des.* **97**, 372-382.
71. Kha, Q. H., Le, V. H., Hung, T. N. K., Nguyen, N. T. K. and Le, N. Q. K. 2023. Development and validation of an explainable machine learning-based prediction model for drug-food interactions from chemical structures. *Sensors(Basel)* **23**, 3962.
72. Kimani, S. W., Owen, J., Green, S. R., Li, F., Li, Y., Dong,

A., Brown, P. J., Ackloo, S., Kuter, D., Yang, C., MacAskill, M., MacKinnon, S. S., Arrowsmith, C. H., Schapira, M., Shahani, V. and Halabelian, L. 2023. Discovery of a novel dcaf1 ligand using a drug-target interaction prediction model: Generalizing machine learning to new drug targets. *J. Chem. Inf. Model* **63**, 4070-4078.
73. Koromina, M., Pandi, M. T. and Patrinos, G. P. 2019. Rethinking drug repositioning and development with artificial intelligence, machine learning, and omics. *OMICS* **23**, 539-548.
74. Koutroumpa, N. M., Papavasileiou, K. D., Papadiamantis, A. G., Melagraki, G. and Afantitis, A. 2023. A systematic review of deep learning methodologies used in the drug discovery process with emphasis on *in vivo* validation. *Int. J. Mol. Sci.* **24**, 6573.
75. Kumar, R., Yadav, G., Kuddus, M., Ashraf, G. M. and Singh, R. 2023. Unlocking the microbial studies through computational approaches: How far have we reached? *Environ. Sci. Pollut. Res. Int.* **30**, 48929-48947.
76. Kumar, S. A., Ananda Kumar, T. D., Beeraka, N. M., Pujar, G. V., Singh, M., Narayana Akshatha, H. S. and Bhagyalalitha, M. 2022. Machine learning and deep learning in data-driven decision making of drug discovery and challenges in high-quality data acquisition in the pharmaceutical industry. *Future Med. Chem.* **14**, 245-270.
77. Li, D., Hu, J., Zhang, L., Li, L., Yin, Q., Shi, J., Guo, H., Zhang, Y. and Zhuang, P. 2022. Deep learning and machine intelligence: New computational modeling techniques for discovery of the combination rules and pharmacodynamic characteristics of traditional chinese medicine. *Eur. J. Pharmacol.* **933**, 175260.
78. Li, G., Lin, P., Wang, K., Gu, C. C. and Kusari, S. 2022. Artificial intelligence-guided discovery of anticancer lead compounds from plants and associated microorganisms. *Trends Cancer* **8**, 65-80.
79. Li, J. Y., Chen, H. Y., Dai, W. J., Lv, Q. J. and Chen, C. Y. 2019. Artificial intelligence approach to investigate the longevity drug. *J. Phys. Chem. Lett.* **10**, 4947-4961.
80. Li, L., Xiong, Y., Zhang, Z. Y., Guo, Q., Xu, Q., Liow, H. H., Zhang, Y. H. and Wei, D. Q. 2015. Improved feature-based prediction of snps in human cytochrome p450 enzymes. *Interdiscip. Sci.* **7**, 65-77.
81. Li, X., Cheng, W., Yang, S., Liang, F., Wang, H., Feng, Y. and Wang, Y. 2022. Establishment of a 13 genes-based molecular prediction score model to discriminate the neurotoxic potential of food relevant-chemicals. *Toxicol. Lett.* **355**, 1-18.
82. Liang, G., Fan, W., Luo, H. and Zhu, X. 2020. The emerging roles of artificial intelligence in cancer drug development and precision therapy. *Biomed. Pharmacother.* **128**, 110255.
83. Lien, S. T., Lin, T. E., Hsieh, J. H., Sung, T. Y., Chen, J. H. and Hsu, K. C. 2023. Establishment of extensive artificial intelligence models for kinase inhibitor prediction: Identification of novel pdgfrb inhibitors. *Comput. Biol. Med.* **156**, 106722.
84. Lin, Z., Cheng, Y. T. and Cheung, B. M. Y. 2023. Machine learning algorithms identify hypokalaemia risk in people with hypertension in the united states national health and nutrition examination survey 1999-2018. *Ann. Med.* **55**, 2209336.
85. Liu, Y., Lim, H. and Xie, L. 2022. Exploration of chemical space with partial labeled noisy student self-training and self-supervised graph embedding. *BMC Bioinformatics* **23**, 158.
86. Lossio-Ventura, J. A., Gonzales, S., Morzan, J., Alatrista-Salas, H., Hernandez-Boussard, T. and Bian, J. 2021. Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artif. Intell. Med.* **117**, 102096.
87. Lu, W. W., Chen, X., Ni, J. L., Cai, W. J., Zhu, S. L., Fei, A. H. and Wang, X. S. 2021. Study on the medication rule of traditional chinese medicine in the treatment of acute pancreatitis based on machine learning technology. *Ann. Palliat. Med.* **10**, 10616-10625.
88. Maassen, O., Fritsch, S., Palm, J., Deffge, S., Kunze, J., Marx, G., Riedel, M., Schuppert, A. and Bickenbach, J. 2021. Future medical artificial intelligence application requirements and expectations of physicians in german university hospitals: Web-based survey. *J. Med. Internet Res.* **23**, e26646.
89. Malone, B., Simovski, B., Moline, C., Cheng, J., Gheorghe, M., Fontenelle, H., Vardaxis, I., Tennoe, S., Malmberg, J. A., Stratford, R. and Clancy, T. 2020. Artificial intelligence predicts the immunogenic landscape of sars-cov-2 leading to universal blueprints for vaccine designs. *Sci. Rep.* **10**, 22375.
90. Marechal, E. 2008. Chemogenomics: A discipline at the crossroad of high throughput technologies, biomarker research, combinatorial chemistry, genomics, cheminformatics, bioinformatics and artificial intelligence. *Comb. Chem. High Throughput Screen* **11**, 583-586.
91. Martin, R. and Yu, K. 2006. Assessing performance of prediction rules in machine learning. *Pharmacogenomics* **7**, 543-550.
92. Martinelli, D. D. 2022. Generative machine learning for de novo drug discovery: A systematic review. *Comput. Biol. Med.* **145**, 105403.
93. McDonagh, J. L., Nath, N., De Ferrari, L., van Mourik, T. and Mitchell, J. B. 2014. Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules. *J. Chem. Inf. Model* **54**, 844-856.
94. McNair, D. 2023. Artificial intelligence and machine learning for lead-to-candidate decision-making and beyond. *Annu. Rev. Pharmacol. Toxicol.* **63**, 77-97.
95. Medina-Franco, J. L., Martinez-Mayorga, K., Fernandez-de Gortari, E., Kirchmair, J. and Bajorath, J. 2021. Rationality over fashion and hype in drug design. *F1000Res.* **10**, Chem Inf Sci-397.
96. Mishra, N. K. 2011. Computational modeling of p450s for toxicity prediction. *Expert Opin. Drug Metab. Toxicol.* **7**,

1211-1231.
97. Mohammed, M. and Omar, N. 2020. Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec. *PLoS One* **15**, e0230 442.
98. Momtazmanesh, S., Nowroozi, A. and Rezaei, N. 2022. Artificial intelligence in rheumatoid arthritis: Current status and future perspectives: A state-of-the-art review. *Rheumatol. Ther.* **9**, 1249-1304.
99. Morales Pantoja, I. E., Smirnova, L., Muotri, A. R., Wahlin, K. J., Kahn, J., Boyd, J. L., Gracias, D. H., Harris, T. D., Cohen-Karni, T., Caffo, B. S., Szalay, A. S., Han, F., Zack, D. J., Etienne-Cummings, R., Akwaboah, A., Romero, J. C., Alam El Din, D. M., Plotkin, J. D., Paulhamus, B. L., Johnson, E. C., Gilbert, F., Curley, J. L., Cappiello, B., Schwamborn, J. C., Hill, E. J., Roach, P., Tornero, D., Krall, C., Parri, R., Sille, F., Levchenko, A., Jabbour, R. E., Kagan, B. J., Berlinicke, C. A., Huang, Q., Maertens, A., Herrmann, K., Tsaioun, K., Dastgheyb, R., Habela, C. W., Vogelstein, J. T. and Hartung, T. 2023. First organoid intelligence (oi) workshop to form an oi community. *Front. Artif. Intell.* **6**, 1116870.
100. Moshawih, S., Goh, H. P., Kifli, N., Idris, A. C., Yassin, H., Kotra, V., Goh, K. W., Liew, K. B. and Ming, L. C. 2022. Synergy between machine learning and natural products cheminformatics: Application to the lead discovery of anthraquinone derivatives. *Chem. Biol. Drug Des.* **100**, 185-217.
101. Moussa, M. and Mandoiu, II. 2018. Single cell rna-seq data clustering using tf-idf based methods. *BMC Genomics* **19**, 569.
102. Nag, S., Baidya, A. T. K., Mandal, A., Mathew, A. T., Das, B., Devi, B. and Kumar, R. 2022. Deep learning tools for advancing drug discovery and development. *3 Biotech.* **12**, 110.
103. Nagarajan, N., Yapp, E. K. Y., Le, N. Q. K., Kamaraj, B., Al-Subaie, A. M. and Yeh, H. Y. 2019. Application of computational biology and artificial intelligence technologies in cancer precision drug discovery. *Biomed. Res. Int.* **2019**, 8427042.
104. Najmi, M., Ayari, M. A., Sadeghsalehi, H., Vaferi, B., Khandakar, A., Chowdhury, M. E. H., Rahman, T. and Jawhar, Z. H. 2022. Estimating the dissolution of anticancer drugs in supercritical carbon dioxide with a stacked machine learning model. *Pharmaceutics* **14**, 1632.
105. Nayarisseri, A., Khandelwal, R., Madhavi, M., Selvaraj, C., Panwar, U., Sharma, K., Hussain, T. and Singh, S. K. 2020. Shape-based machine learning models for the potential novel covid-19 protease inhibitors assisted by molecular dynamics simulation. *Curr. Top. Med. Chem.* **20**, 2146-2167.
106. Nayarisseri, A., Khandelwal, R., Tanwar, P., Madhavi, M., Sharma, D., Thakur, G., Speck-Planche, A. and Singh, S. K. 2021. Artificial intelligence, big data and machine learning approaches in precision medicine & drug discovery. *Curr. Drug Targets* **22**, 631-655.
107. Nigam, A. K., Ojha, A. A., Li, J. G., Shi, D., Bhatnagar, V., Nigam, K. B., Abagyan, R. and Nigam, S. K. 2021. Molecular properties of drugs handled by kidney oats and liver oatps revealed by chemoinformatics and machine learning: Implications for kidney and liver disease. *Pharmaceutics* **13**, 1720.
108. Niu, Q., Li, H., Tong, L., Liu, S., Zong, W., Zhang, S., Tian, S., Wang, J., Liu, J., Li, B., Wang, Z. and Zhang, H. 2023. Tcmfp: A novel herbal formula prediction method based on network target's score integrated with semi-supervised learning genetic algorithms. *Brief Bioinform.* **24**, bbad102.
109. Noorain, L., Nguyen, V., Kim, H. W. and Nguyen, L. T. B. 2023. A machine learning approach for plga nanoparticles in antiviral drug delivery. *Pharmaceutics* **15**, 495.
110. Nowak, D., Bachorz, R. A. and Hoffmann, M. 2023. Neural networks in the design of molecules with affinity to selected protein domains. *Int. J. Mol. Sci.* **24**, 1762.
111. Nussinov, R., Zhang, M., Liu, Y. and Jang, H. 2023. Alphafold, allosteric, and orthosteric drug discovery: Ways forward. *Drug Discov. Today* **28**, 103551.
112. Overduin, M., Kervin, T. A., Klarenbach, Z., Adra, T. R. C. and Bhat, R. K. 2023. Comprehensive classification of proteins based on structures that engage lipids by composel. *Biophys. Chem.* **295**, 106971.
113. Ozcelik, R., van Tilborg, D., Jimenez-Luna, J. and Grisoni, F. 2023. Structure-based drug discovery with deep learning. *Chembiochem* **24**, e202200776.
114. Pandiyan, S. and Wang, L. 2022. A comprehensive review on recent approaches for cancer drug discovery associated with artificial intelligence. *Comput. Biol. Med.* **150**, 106140.
115. Pirzada, R. H., Ahmad, B., Qayyum, N. and Choi, S. 2023. Modeling structure-activity relationships with machine learning to identify gsk3-targeted small molecules as potential covid-19 therapeutics. *Front. Endocrinol. (Lausanne)* **14**, 1084327.
116. Popa, S. L., Pop, C., Dita, M. O., Brata, V. D., Bolchis, R., Czako, Z., Saadani, M. M., Ismaiel, A., Dumitrascu, D. I., Grad, S., David, L., Cismaru, G. and Padureanu, A. M. 2022. Deep learning and antibiotic resistance. *Antibiotics(Basel)* **11**, 1674.
117. Poweleit, E. A., Vinks, A. A. and Mizuno, T. 2023. Artificial intelligence and machine learning approaches to facilitate therapeutic drug management and model-informed precision dosing. *Ther. Drug Monit.* **45**, 143-150.
118. Prabakaran, P., Rao, S. P. and Wendt, M. 2021. Animal immunization merges with innovative technologies: A new paradigm shift in antibody discovery. *MAbs.* **13**, 1924347.
119. Priya, S., Tripathi, G., Singh, D. B., Jain, P. and Kumar, A. 2022. Machine learning approaches and their applications in drug discovery and design. *Chem. Biol. Drug Des.* **100**, 136-153.
120. Purpura, A., Giorgianni, D., Orru, G., Melis, G. and Sar-

tori, G. 2022. Identifying single-item faked responses in personality tests: A new tf-idf-based method. *PLoS One* **17**, e0272970.
121. Qiu, H. Y., Clausen, R. P., He, Y. and Zhu, H. L. 2021. Artificial intelligence and cheminformatics-guided modern privileged scaffold research. *Curr. Top. Med. Chem.* **21**, 2593-2608.
122. Ranjan, A., Fernandez-Baca, D., Tripathi, S. and Deepak, A. 2022. An ensemble tf-idf based approach to protein function prediction via sequence segmentation. *IEEE/ ACM Trans. Comput. Biol. Bioinform.* **19**, 2685-2696.
123. Ranson, J. M., Bucholc, M., Lyall, D., Newby, D., Winchester, L., Oxtoby, N. P., Veldsman, M., Rittman, T., Marzi, S., Skene, N., Al Khleifat, A., Foote, I. F., Orgeta, V., Kormilitzin, A., Lourida, I. and Llewellyn, D. J. 2023. Harnessing the potential of machine learning and artificial intelligence for dementia research. *Brain Inform.* **10**, 6.
124. Rema, J., Novais, F. and Telles-Correia, D. 2022. Precision psychiatry: Machine learning as a tool to find new pharmacological targets. *Curr. Top. Med. Chem.* **22**, 1261-1269.
125. Sahu, A., Mishra, J. and Kushwaha, N. 2022. Artificial intelligence (ai) in drugs and pharmaceuticals. *Comb. Chem. High Throughput Screen* **25**, 1818-1837.
126. Sanawar, R., Sahayasheela, V. J., Sarath, P. and Dan, V. M. 2023. Discoidin domain receptor 1 inhibitors: Advances and future directions for novel therapeutics with aid of DNA encoded library screens and artificial intelligence. *Mini Rev. Med. Chem.* **23**, 1507-1513.
127. Sasahara, K., Shibata, M., Sasabe, H., Suzuki, T., Takeuchi, K., Umehara, K. and Kashiyama, E. 2021. Predicting drug metabolism and pharmacokinetics features of in-house compounds by a hybrid machine-learning model. *Drug Metab. Pharmacokinet.* **39**, 100395.
128. Selvaraj, C., Chandra, I. and Singh, S. K. 2022. Artificial intelligence and machine learning approaches for drug design: Challenges and opportunities for the pharmaceutical industries. *Mol. Divers.* **26**, 1893-1913.
129. Serafim, M. S. M., Gertrudes, J. C., Costa, D. M. A., Oliveira, P. R., Maltarollo, V. G. and Honorio, K. M. 2021. Knowing and combating the enemy: A brief review on sars-cov-2 and computational approaches applied to the discovery of drug candidates. *Biosci. Rep.* **41**, BSR 20202616.
130. Seyedtabib, M. and Kamyari, N. 2023. Predicting polypharmacy in half a million adults in the iranian population: Comparison of machine learning algorithms. *BMC Med. Inform. Decis. Mak.* **23**, 84.
131. Sharma, P., Dahiya, S., Kaur, P. and Kapil, A. 2023. Computational biology: Role and scope in taming antimicrobial resistance. *Indian J. Med. Microbiol.* **41**, 33-38.
132. Shimazaki, T. and Tachikawa, M. 2022. Collaborative approach between explainable artificial intelligence and simplified chemical interactions to explore active ligands for cyclin-dependent kinase 2. *ACS Omega* **7**, 10372-10381.
133. Sicular, S., Alpaslan, M., Ortega, F. A., Keathley, N., Venkatesh, S., Jones, R. M. and Lindsey, R. V. 2022. Reevaluation of missed lung cancer with artificial intelligence. *Respir. Med. Case Rep.* **39**, 101733.
134. Singh, A. K., Ling, J. and Malviya, R. 2023. Prediction of cancer treatment using advancements in machine learning. *Recent Pat. Anticancer Drug Discov.* **18**, 364-378.
135. Singh, V., Shrivastava, S., Kumar Singh, S., Kumar, A. and Saxena, S. 2022. Accelerating the discovery of antifungal peptides using deep temporal convolutional networks. *Brief Bioinform.* **23**, bbac008.
136. Singla, R., Aggarwal, S., Bindra, J., Garg, A. and Singla, A. 2022. Developing clinical decision support system using machine learning methods for type 2 diabetes drug management. *Indian J. Endocrinol. Metab.* **26**, 44-49.
137. Singla, R. K., Joon, S., Sinha, B., Kamal, M. A., Simal-Gandara, J., Xiao, J. and Shen, B. 2023. Current trends in natural products for the treatment and management of dementia: Computational to clinical studies. *Neurosci. Biobehav. Rev.* **147**, 105106.
138. Srisongkram, T. and Weerapreeyakul, N. 2022. Drug repurposing against kras mutant g12c: A machine learning, molecular docking, and molecular dynamics study. *Int. J. Mol. Sci.* **24**, 669.
139. Srivathsa, A. V., Sadashivappa, N. M., Hegde, A. K., Radha, S., Mahesh, A. R., Ammunje, D. N., Sen, D., Theivendren, P., Govindaraj, S., Kunjiappan, S. and Pavadai, P. 2023. A review on artificial intelligence approaches and rational approaches in drug discovery. *Curr. Pharm. Des.* **29**, 1180-1192.
140. Tang, K., Zhu, R., Li, Y. and Cao, Z. 2011. Discrimination of approved drugs from experimental drugs by learning methods. *BMC Bioinformatics* **12**, 157.
141. Tanoli, Z., Vaha-Koskela, M. and Aittokallio, T. 2021. Artificial intelligence, machine learning, and drug repurposing in cancer. *Expert Opin. Drug Discov.* **16**, 977-989.
142. Terranova, N., Venkatakrishnan, K. and Benincosa, L. J. 2021. Application of machine learning in translational medicine: Current status and future opportunities. *AAPS J.* **23**, 74.
143. Tian, S., Wang, J., Li, Y., Xu, X. and Hou, T. 2012. Drug-likeness analysis of traditional chinese medicines: Prediction of drug-likeness using machine learning approaches. *Mol. Pharm.* **9**, 2875-2886.
144. Tran, T. T. V., Tayara, H. and Chong, K. T. 2023. Recent studies of artificial intelligence on in silico drug distribution prediction. *Int. J. Mol. Sci.* **24**, 1815.
145. Tripathi, A., Misra, K., Dhanuka, R. and Singh, J. P. 2023. Artificial intelligence in accelerating drug discovery and development. *Recent Pat. Biotechnol.* **17**, 9-23.
146. Trisciuzzi, D., Villoutreix, B. O., Siragusa, L., Baroni, M., Cruciani, G. and Nicolotti, O. 2023. Targeting protein-protein interactions with low molecular weight and

short peptide modulators: Insights on disease pathways and starting points for drug discovery. *Expert Opin. Drug Discov.* **18**, 737-752.
147. Tseng, Y. J., Chuang, P. J. and Appell, M. 2023. When machine learning and deep learning come to the big data in food chemistry. *ACS Omega* **8**, 15854-15864.
148. Uesawa, Y. 2020. [Ai-based qsar modeling for prediction of active compounds in mie/aop]. *Yakugaku Zasshi* **140**, 499-505.
149. Vemula, D., Jayasurya, P., Sushmitha, V., Kumar, Y. N. and Bhandari, V. 2023. Cadd, ai and ml in drug discovery: A comprehensive review. *Eur. J. Pharm. Sci.* **181**, 106324.
150. Veselkov, K., Gonzalez, G., Aljifri, S., Galea, D., Mirnezami, R., Youssef, J., Bronstein, M. and Laponogov, I. 2019. Hyperfoods: Machine intelligent mapping of cancer-beating molecules in foods. *Sci. Rep.* **9**, 9237.
151. Vidovic, T., Dakhovnik, A., Hrabovskyi, O., MacArthur, M. R. and Ewald, C. Y. 2023. Ai-predicted mtor inhibitor reduces cancer cell proliferation and extends the lifespan of c. Elegans. *Int. J. Mol. Sci.* **24**, 7850.
152. Villalobos-Alva, J., Ochoa-Toledo, L., Villalobos-Alva, M. J., Aliseda, A., Perez-Escamirosa, F., Altamirano-Bustamante, N. F., Ochoa-Fernandez, F., Zamora-Solis, R., Villalobos-Alva, S., Revilla-Monsalve, C., Kemper-Valverde, N. and Altamirano-Bustamante, M. M. 2022. Protein science meets artificial intelligence: A systematic review and a biochemical meta-analysis of an inter-field. *Front. Bioeng. Biotechnol.* **10**, 788300.
153. Vishnoi, S., Matre, H., Garg, P. and Pandey, S. K. 2020. Artificial intelligence and machine learning for protein toxicity prediction using proteomics data. *Chem. Biol. Drug Des.* **96**, 902-920.
154. Vo, D., Ghosh, P. and Sahoo, D. 2023. Artificial intelligence-guided discovery of gastric cancer continuum. *Gastric Cancer* **26**, 286-297.
155. Wang, L., Song, Y., Wang, H., Zhang, X., Wang, M., He, J., Li, S., Zhang, L., Li, K. and Cao, L. 2023. Advances of artificial intelligence in anti-cancer drug design: A review of the past decade. *Pharmaceuticals (Basel)* **16**, 253.
156. Wang, M., Wang, J., Weng, G., Kang, Y., Pan, P., Li, D., Deng, Y., Li, H., Hsieh, C. Y. and Hou, T. 2022. Remode: A deep learning-based web server for target-specific drug design. *J. Cheminform.* **14**, 84.
157. Wang, M., Zhou, X., King, R. W. and Wong, S. T. 2007. Context based mixture model for cell phase identification in automated fluorescence microscopy. *BMC Bioinformatics* **8**, 32.
158. Wang, Y. Y. and Acero, A. 2007. Maximum entropy model parameterization with TF*IDF weighted vector space model. *2007 Ieee Workshop on Automatic Speech Recognition and Understanding*. December 9-13. Kyoto, Japan. Vols **1** and **2**, 213-218.
159. Wu, Y. and Wang, G. 2018. Machine learning based toxicity prediction: From chemical structural description to transcriptome analysis. *Int. J. Mol. Sci.* **19**, 2358.
160. Wu, Z., Lei, T., Shen, C., Wang, Z., Cao, D. and Hou, T. 2019. Admet evaluation in drug discovery. 19. Reliable prediction of human cytochrome p450 inhibition using artificial intelligence approaches. *J. Chem. Inf. Model.* **59**, 4587-4601.
161. Xing, G., Liang, L., Deng, C., Hua, Y., Chen, X., Yang, Y., Liu, H., Lu, T., Chen, Y. and Zhang, Y. 2020. Activity prediction of small molecule inhibitors for anti-rheumatoid arthritis targets based on artificial intelligence. *ACS Comb. Sci.* **22**, 873-886.
162. Xu, D., Liu, B., Wang, J. and Zhang, Z. 2022. Bibliometric analysis of artificial intelligence for biotechnology and applied microbiology: Exploring research hotspots and frontiers. *Front. Bioeng. Biotechnol.* **10**, 998298.
163. Xu, R. and Wang, Q. 2014. Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. *J. Biomed. Inform.* **51**, 191-199.
164. Xu, S., Leng, Y., Feng, G., Zhang, C. and Chen, M. 2023. A gene pathway enrichment method based on improved tf-idf algorithm. *Biochem. Biophys. Rep.* **34**, 101421.
165. Yang, F., Darsey, J. A., Ghosh, A., Li, H. Y., Yang, M. Q. and Wang, S. 2022. Artificial intelligence and cancer drug development. *Recent Pat. Anticancer Drug Discov.* **17**, 2-8.
166. Yang, J., Li, Z., Wu, W. K. K., Yu, S., Xu, Z., Chu, Q. and Zhang, Q. 2022. Deep learning identifies explainable reasoning paths of mechanism of action for drug repurposing from multilayer biological network. *Brief Bioinform.* **23**, bbac469.
167. Yang, X. D. and Jia, B. 2010. Vector space model based on lucene index and tf-idf weighting algorithm. *Proceedings of 2010 Asia-Pacific Youth Conference on Communication.* August 7-8, Kunming, China. Vols **1** and **2**, 20-23.
168. Ye, J., Li, A., Zheng, H., Yang, B. and Lu, Y. 2023. Machine learning advances in predicting peptide/protein-protein interactions based on sequence information for lead peptides discovery. *Adv. Biol.(Weinh)* **7**, e2200 232.
169. Yeruva, V. K., Junaid, S. and Lee, Y. 2019. Contextual word embeddings and topic modeling in healthy dieting and obesity. *J. Healthc. Inform. Res.* **3**, 159-183.
170. Yuan, H., Tang, Y., Sun, W. and Liu, L. 2020. A detection method for android application security based on tf-idf and machine learning. *PLoS One* **15**, e0238694.
171. Yuba, M. and Iwasaki, K. 2023. Performance evaluation methods for improvements at post-market of artificial intelligence/machine learning-based computer-aided detection/diagnosis/triage in the united states. *PLOS Digit. Health* **2**, e0000209.
172. Zaizar-Fregoso, S. A., Lara-Esqueda, A., Hernandez-Suarez, C. M., Delgado-Enciso, J., Garcia-Nevares, A., Canseco-Avila, L. M., Guzman-Esquivel, J., Rodriguez-Sanchez, I. P., Martinez-Fierro, M. L., Ceja-Espiritu, G.,

Ochoa-Diaz-Lopez, H., Espinoza-Gomez, F., Sanchez-Diaz, I. and Delgado-Enciso, I. 2023. Using artificial intelligence to develop a multivariate model with a machine learning model to predict complications in mexican diabetic patients without arterial hypertension (national nested case-control study): Metformin and elevated normal blood pressure are risk factors, and obesity is protective. *J. Diabetes Res.* **2023**, 8898958.

173. Zhang, C., Cheng, F., Li, W., Liu, G., Lee, P. W. and Tang, Y. 2016. In silico prediction of drug induced liver toxicity using substructure pattern recognition method. *Mol. Inform.* **35**, 136-144.
174. Zhang, H., Guo, J., Li, H. and Guan, Y. 2022. Machine learning for artemisinin resistance in malaria treatment across *in vivo-in vitro* platforms. *iScience* **25**, 103910.
175. Zhang, N., Zhang, H., Liu, Z., Dai, Z., Wu, W., Zhou, R., Li, S., Wang, Z., Liang, X., Wen, J., Zhang, X., Zhang, B., Ouyang, S., Zhang, J., Luo, P., Li, X. and Cheng, Q. 2023. An artificial intelligence network-guided signature for predicting outcome and immunotherapy response in lung adenocarcinoma patients based on 26 machine learning algorithms. *Cell Prolif.* **56**, e13409.
176. Zhang, X., Shen, C., Guo, X., Wang, Z., Weng, G., Ye, Q., Wang, G., He, Q., Yang, B., Cao, D. and Hou, T. 2021. Asfp (artificial intelligence based scoring function platform): A web server for the development of customized scoring functions. *J. Cheminform.* **13**, 6.
177. Zulqarnain, F., Rhoads, S. F. and Syed, S. 2023. Machine and deep learning in inflammatory bowel disease. *Curr. Opin. Gastroenterol.* **39**, 294-300.

초록 : 텍스트 마이닝을 이용한 인공지능 활용 신약 개발 연구 동향 분석

남재우[1] · 김영준[2]*
([1]건국대학교 문헌정보학과, [2]건국대학교 바이오의약학과)

본 리뷰 논문은 2010년부터 2022년까지의 인공지능을 활용한 신약개발 관련 연구동향을 분석하여 정리하였다. 이러한 분석을 통해 2,421개 연구의 초록을 코퍼스로 구성하고, 전처리를 거쳐 빈도가 높고 연결중심성이 높은 단어를 추출하였다. 분석 결과 2010-201년과 2020-2022년 단어빈도 추이는 비슷한 것으로 구분되어 나타났다. 연구 방법으로는 2010년부터 2020년까지 머신 러닝을 활용한 연구가 많이 진행되었고, 2021년부터는 딥러닝을 활용한 연구가 증가하고 있다. 이러한 연구를 통해 이루어지고 있는 인공지능 활용 연구 동향에 대해 분야별로 살펴보고 관련 연구의 장점, 문제점, 도전과제 등을 살펴보았다. 파악되어진 연구 동향은 2021년 이후로 약물의 재배치를 인공지능 활용 연구, 항암제 개발을 위한 컴퓨터 활용 연구, 임상시험에 인공지능 적용 연구 등과 같이 인공지능 적용 분야가 확대되고 있다는 점이다. 이러한 과정을 통해 향후 이루어질 것으로 예상되는 인공지능 활용 신약개발 연구의 전망에 대해 간략히 제시하였다. 위의 인공지능 기술 발전과 함께 바이오와 의료데이터의 신뢰성과 안전성이 확보되어진다면 인공지능 활용 신약개발의 방향이 개인 맞춤형 의료와 정밀의료 분야로 진행되어질 것으로 판단하기에 이에 대한 지속적인 노력이 필요하리라 본다.