JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 12: 1797~1808, Dec. 2021

신약개발에서의 AI 기술 활용 현황과 미래

정명희[1*] · 권원현[2]

Present Status and Future of AI-based Drug Discovery

Myunghee Jung[1*] · Wonhyun Kwon[2]

[1*]Professor, Department of Software Engineering, Anyang University, Anyang, 14028 Korea
[2]Professor, Department of Information, Electircal and Electronic Engineering, Anyang University, Anyang, 14028 Korea

요 약

4차 산업혁명을 주도하는 기술 중 가장 핵심적인 기술로 꼽히고 있는 인공지능은 다양한 분야에 접목되면서 우리 사회 전반에 걸쳐 패러다임의 전환을 가져오고 있다. 바이오 분야 역시 예외는 아니어서 컴퓨터, 전기·전자, 기계 등 타 학문과 융합되면서 방대한 데이터 기반의 AI 기술을 도입하고 있다. 신약개발에서 AI 기술 도입은 신약개발의 효율성을 개선하고 효능 및 품질 향상을 가져올 수 있다. 신약개발은 다학제 분야가 접목된 융합 분야이고 개발 과정 단계별로 결과의 불확실성이 존재하고 있어 실용적 수준의 신약 개발을 위해서는 화학, 생물학, 독성학, 약동학 등 전문 지식의 융합을 기반으로 하는 AI 기술 개발이 필요하다. 신약개발은 크게 주어진 질병에 대한 타겟 물질 발굴 및 검증, 히트 및 선도물질 발굴, 도출된 화합물에 대한 합성 가능성 및 효능 등에 대한 평가(Scoring)를 거쳐 최적의 신약 후보 물질을 발굴하고 마지막으로 전임상과 임상 과정의 단계를 거친다. 이때 AI 기술은 모든 단계에서 적용될 수 있고 단계마다 특화되어 적용될 수 있다. 본 논문에서는 신약개발을 위해 적용되고 있는 AI 기술 현황과 현재 기술의 한계를 살펴보고 향후 신약개발에서 AI 기술의 발전 방향을 고찰해 보고자 한다.

ABSTRACT

Artificial intelligence is considered one of the core technologies leading the 4th industrial revolution. It is adopted in various fields bringing about a huge paradigm shift throughout our society. The field of biotechnology is no exception. It is undergoing innovative development by converging with other disciplines such as computers, electricity, electronics, and so on. In drug discovery and development, big data-based AI technology has a great potential of improving the efficiency and quality of drug development, rapidly advancing to overcome the limitations in the existing drug development process. AI technology is to be specialized and developed for the purpose including clinical efficacy and safety-related end points based on the multidisciplinary knowledge such as biology, chemistry, toxicology, pharmacokinetics, etc. In this paper, we review the current status of AI technology applied for drug discovery and consider its limitations and future direction.

키워드 : 인공지능, 머신러닝, 딥러닝, 신약개발

Keywords : Artificial intelligence, Machine learning, Deep learning, Drug discover and development

Received 12 August 2021, **Revised** 25 September 2021, **Accepted** 29 October 2021

*** Corresponding Author** Myugnhee Jung (E-mail:mhjung@anyang.ac.kr, Tel:+82-31-467-0963)
Professor, Department of Software Engineering, Anyang University, Anyang, 14028 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.12.1797 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

인공지능은 4차 산업혁명을 주도하는 가장 핵심 기술 중의 하나로 사회 및 산업 전반에 걸쳐 패러다임의 변화를 가져오고 있다. 바이오 분야 역시 예외는 아니어서 컴퓨팅, 자동화 및 AI 기술 발전에 힘입어 바이오 혁신 기술들이 개발되고 있다. R&D 비용 증가와 효율성의 한계가 있던 신약개발 분야에서도 방대한 데이터와 인공지능을 활용해 효율성 개선과 효능 및 품질 향상을 위한 기술개발이 이루어지는 중이다[1-3].

전통적인 신약개발은 평균 약 10년의 기간이 소요되고, 성공률도 1/5,000~1/10,000으로 매우 낮은 고비용 과정이다. 신약개발은 관련된 연구논문들, 유전체 및 오믹스 정보, 화학유전체 등 방대한 전문지식을 바탕으로 화학 및 생물학의 통합 및 연계가 필요한 분야로 인공지능 기술을 적용하면 개발의 효율성을 높일 수 있다. 또한, 인공지능은 데이터에 내재한 패턴과 상호 연관성을 찾아낼 수 있는 기술로 기존 방법으로는 발굴하기 어려운 질병 타겟과 약물 재설계에 대한 새로운 가능성도 열어 준다[4]. 그동안 데이터 기반의 가상 탐색(Data-driven Virtual Screening, DDVS)에 관한 연구를 지속해 온 덕분에 현재는 방대한 화합물 생리활성 및 스크리닝 빅데이터들이 축적되어 있고, 빅데이터를 기반으로 분자 도킹, 약전모델링(pharmacophore modeling), 분자비교분석 등과 같이 약물 후보물질 발굴을 위해 머신러닝 및 딥러닝 기반의 기술이 다양하게 적용되고 있다[4-6].

머신러닝은 약물 설계를 위해 대규모 화합물 데이터베이스 마이닝에 중요한 도구로 패턴인식 알고리즘을 사용하여 저분자 물질의 실험적 관측치 간의 수학적 관계를 식별하여 새로운 화합물의 화학적, 생물학적, 물리적 특성을 예측한다. 이는 빅데이터 스크리닝을 통해 찾은 히트 화합물의 결합 친화도, 생물학적 반응 또는 물리화학적 특성 향상을 위한 화학적 구조의 최적화를 위해 사용되고 나아가 구조-활성의 관계를 정량적(수치적)으로 설명하는 모델인 QSAR(Quantitative structure-activity relationship)이나 QSPR(quantitative structure-property relationships) 모델링, 이를 통한 독성, 대사, 약물 간의 상호작용 및 발암성과 같은 약물의 물리화학적 특성을 이해하기 위해 도입되고 있다[7-9]. 머신러닝이 기존 물리적 모델보다 더 효율적이고 자원의 확장 없이도 더 많은 자료를 다룰 수 있는 장점은 있지만, 실험 데이터의 부재와 모델링에서 사용한 가정(assumption)으로 인해 결과의 불확실성도 존재한다. 따라서 실용적 수준의 맞춤형 약물 개발을 지원하는 기술이 되기 위해서는 특화되고 전문지능화된 모델링과 이를 위한 전문 데이터가 필요하다[10].

사실 신약개발의 성공은 임상의 성공을 의미한다. 따라서 이러한 관점에서 임상 시험 대상 화합물을 결정하는 것이 목표이기 때문에 단순히 데이터 기반의 화합물을 만드는 기술보다는 임상 효능 및 안전성의 관점에서 약물이 개발될 수 있도록 양질의 학습 데이터가 만들어져야 한다[10]. 따라서 AI 기술의 잠재력을 충분히 활용하기 위해서는 화학, 생물학, 독성학, 약동학(pharmacokinetics) 등 전문지식의 연계 및 융합이 필요하고 이러한 점이 실용화 관점에서 도전적이다.

신약개발은 크게 질병에 대한 타겟 단백질 발굴, 히트(hit) 및 선도(lead)물질 발굴, 합성 가능성 및 효능, 독성 등에 대한 평가(Scoring)를 거쳐 최적의 후보물질을 발굴하고 마지막으로 전임상과 임상의 단계를 거치며 맞춤형 약물로 개발된다[3, 5, 9]. 이때 AI 기술은 신약개발의 거의 모든 단계에서 목적에 맞게 적용될 수 있다[9]. 본 논문에서는 신약개발을 위해 적용되고 있는 인공지능의 머신러닝 및 딥러닝 알고리즘과 기술적용의 현재 한계를 살펴보고 향후 신약개발에서 AI 기술의 미래 발전 방향을 고찰해 보고자 한다.

Ⅱ. 본 론

2.1. 신약개발과 컴퓨팅 기술

신약개발은 화학 및 생물학 등 관련 분야에 최신 컴퓨팅 기술이 융합되면서 그 복잡성이 증가하였고 일련의 순차적 선형과정이라기보다는 대용량처리 화합물 및 스크리닝, 계산 모델링 및 문헌 정보와 같은 다양한 결과를 활용한 피드백 기반의 과정으로 질적인 결과를 얻기 위해 반복과 최적화 과정을 거치는 특징이 있다. 컴퓨터를 사용한 신약 설계(CADD computer-aided drug design)는 1950년대부터 시작되어 현재는 빅데이터 분석을 통한 새로운 타겟 발굴, 전체 신호전달경로 모델링을 통한 더 효과적인 타겟 발굴, 잠재적인 리간드 분석을 통한 후보물질 발굴 및 최적화, 독성 분석 등 다양한 부분에 컴퓨팅 기술이 적용되고 있다[7-9].

2.2. 화학 정보학에 대한 이해

신약개발에 인공지능을 적용하려면 화학 정보학에서 다루는 화합물에 사용되는 데이터 형식과 유형 등에 대한 이해가 필요하다[11]. 화학 정보학은 화학 데이터의 저장, 처리 및 분석을 위한 컴퓨터 응용을 연구하는 분야로 30년 넘게 화학적 표현, 화학적 설명자 분석, 라이브러리 설계, QSAR 분석 및 CADD와 같은 주제를 다루어 왔다[12]. 이러한 과정에서 데이터 처리를 위한 여러 화학 데이터 형식이 제안되었는데 화합물의 2D 또는 3D 좌표를 저장하기 위해 SDF(Structure Data Format), MDL(Molfile) 및 구조 데이터를 저장하는 PDB(Protein Data Bank) 등이 대표적인 형식이다. 또한, 제한된 컴퓨터 용량으로 대규모 화합물 데이터베이스를 다루기 위해 화학적 라인 표기법도 도입되었는데 가장 많이 사용되는 형식이 SMILES(simplified molecular-input line-entry system) 형식이다. SMILES는 1986년에 David Weininger가 처음 개발한 방식으로 간단한 ASCII 코드를 사용하여 복합 구조를 저장하는 표현법인데 화합물을 찾거나 유사한 구조를 가진 화합물을 검색하는 데 유용하다. 더욱 복잡한 분자의 경우는 화학구조의 일관된 라벨링과 정렬을 위해 Morgan 알고리즘을 적용한 표준 SMILES가 사용되고 있다[13].

또한, 화학 물질의 구조가 생물학적 활동에 영향을 미치는지 이해하려면 화학 그래프 이론에 대해 이해해야 한다. '분자 그래프' 또는 '구조 그래프'라고도 하는 화학 그래프는 정렬된 쌍 G = (V, E)로 표시하며 V는 정점(원자), E는 V를 연결하는 선(bond)을 뜻한다. 화학 그래프는 화학구조를 결합 인접 행렬 또는 토폴로지 거리 행렬을 사용하여 원자 연결성을 나타내고 생물학적 현상을 모델링하는데 필요한 여러 토폴로지 인덱스의 계산을 지원한다[11].

화학적 표현자(chemical descriptor)는 화합물의 구조 정보를 수학적으로 해석해서 도출한 수치적 값으로 분자 데이터 마이닝과 화합물 다양성 분석 및 활성 예측 등에 사용된다. 화학적 표현자는 0차원(0D)에서부터 4차원(4D)까지 다양한 표현자가 있는데 3D 표현자는 구조적 변화에 가장 민감하고 '스캐폴드 홉(scaffold hops)'을 식별할 수 있다고 알려져 있다. 이러한 화학구조 특징을 계산하기 위해 PaDEL-Descriptor와 같은 다양한 프로그램들이 개발되어 사용되고 있다[11, 12, 14].

화학적 지문(chemical fingerprint)은 화합물에 대한 정보를 나타내는 고차원 벡터로 화학적 분석 및 유사성 기반 가상 스크리닝에 사용된다. 사전에 하위구조(substructure) 키들을 정해 놓고 화합물에서 이 하위구조 키의 존재 여부를 0과 1로 나타내면 화합물을 일정한 길이의 바이너리 지문(0과 1)으로 나타낼 수 있다. 많이 사용되는 예로 ECFP(extended-connectivity fingerprints)가 있는데 이는 분자 특성화, 유사성 검색 및 구조-활성 모델링을 위해 설계된 원형 토폴로지 지문이다 [15]

머신러닝을 적용하여 화학정보를 분석하는 과정의 첫 단계로 화합물 데이터베이스에서 얻은 화합물에 대해 화학적 특징을 추출(chemical feature extraction)하는데 화합물은 하위구조(substructure fragment)나 기타 화학적 표현자로 표현된다. 다음은 유사성 비교를 위해 특정 하위구조의 유무를 이용해 화학적 지문을 생성(chemical fingerprint creation)하고 이를 기반으로 알려진 화합물의 화학적 특징을 사용하여 머신러닝 모형을 훈련한 뒤 이를 통해 화합물 특성을 예측한다[11-13]. 이 과정은 그림 1과 같이 요약될 수 있다.

Fig. 1 Chemoinformatics analysis framework using machine learning

2.3. 신약개발에서의 AI 기술의 활용

신약개발은 연구논문이나 특허 자료, 화합물의 구조 및 효능 관련된 빅데이터, 의료데이터, 임상데이터 등 방대한 자료 분석을 기반으로 화합물 활성 및 효능 극대화에서 독성 및 부작용 최소화에 이르기까지 다양한 요인을 동시에 최적화해야 하는 까다로운 과정이다. 이러한 과정에서 특정 부분을 자동화하면 무작위성과 오류가 줄어들어 약물 개발의 효율성이 크게 향상될 수 있고 방대한 자료의 지능적 탐색 및 패턴인식도 가능하다. 인공지능 기술은 신약개발 모든 단계에서 유용하게 적용될 수 있는데 자료에 대한 단순한 분석 및 자동화를 넘어 자료로부터의 학습을 통해 표면적으로는 보이지 않은 내재한 현상과 패턴에 대한 통찰을 주기 때문에 차세대 기술로 주목받고 있다[1-4].

신약개발에 사용되는 AI 기술은 머신러닝(ML: machine learning), 세부적으로는 딥러닝(DL: deep learing) 알고리즘이 주로 사용된다[7, 16, 17]. 머신러닝의 대표적인

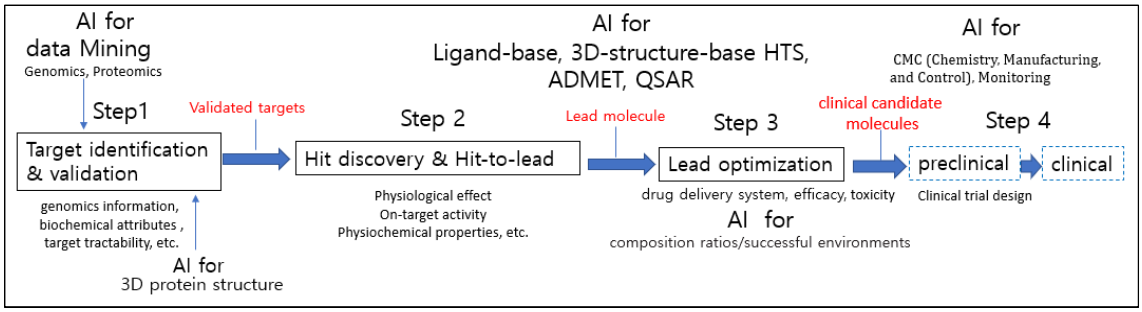

Fig. 2 Drug development and artificial intelligence: utilization and applications of AI in the drug development process

적용 예로는 약물의 새로운 용도 발견, 약물-단백질 상호작용 예측, 약물 효능 발견, 바이오마커 확인, 분자의 생물학적 활성 최적화 등이 있다[8, 16]. 신약개발 과정에서 인공지능 기술의 활용은 그림 2와 같이 요약될 수 있다.

2.3.1. 타겟 단백질 발굴

AI는 질병에 대한 타겟 발굴 및 검증을 효율적으로 할 수 있는 기술로 사용된다. 논문 특허 데이터 화합물의 구조 및 효능 관련 자료, 유전체, 프로테오믹스 등 유전 및 단백질 관련 자료 등을 학습하여 타겟 발굴 및 검증을 용이하게 하고 기존 약물의 재활성도 향상할 수 있다[18-23].

2.3.2. 단백질 3차원 구조 규명

현재 신약개발에서 가장 어려운 부분이 단백질의 3차원 구조를 밝히는 것이다. 신약개발에서 신약후보 물질로 도출된 화합물과 타겟 단백질의 구조는 열쇠와 자물쇠의 관계와 같다. 아무리 이론적, 실험적으로 뒷받침되는 후보물질을 발견해도 타겟 단백질 및 활성 부위의 구조를 모른다면 신약의 유효성을 확보하기 어렵다. 그동안 단백질 3차원 구조는 주로 X선 결정학(X-ray crystallography)과 핵자기 공명 분광법 (NMR)을 사용하였는데 최근 구글의 딥마인드가 단백질 3차원 구조 규명에 AI 기술을 활용한 '알파폴더'를 발표하면서 인공지능이 단백질 구조 유추의 새로운 방법론이 될 수 있음을 보여주었다. 또한, 약물 설계 과정에서 단백질 변형은 중요한 과정 중의 하나인데 AI 기술은 이러한 3차원 약물 설계에도 사용될 수 있다[24, 25].

2.3.3. 히트 및 리드 물질 발굴

약물 후보 발굴을 위해서는 처리량이 많은 화합물 및 단편 스크리닝, 컴퓨터 모델링 기술 및 문헌을 비롯한 다양한 정보가 활용된다. 일반적으로 약물이 적용될 단백질 타겟이 선정되면 대규모 화합물 라이브러리 중에서 대상 타겟에 대해 활성을 보이는 약물 후보를 선별하는 스크리닝 단계를 거친다. 이 단계는 많은 시간과 자원을 필요로 하므로 이를 효율적으로 수행하기 위해 타겟과 화합물의 구조 모델링 기반의 분자 도킹기술을 사용한 가상탐색(Virtual Screening) 기법을 수행한다. 그동안 화합물에 대해 다양한 생리활성 데이터가 축적되었고 화합물-타겟 DB도 구축되어 공개되었다. 많은 도킹 알고리즘이 개발되어 사용되고 있지만, 최근에는 AI 기술을 적용하여 활성 화합물 스크리닝의 효율성을 높이고 있다. 대규모 화합물 활성 데이터는 딥러닝을 포함한 머신러닝 모델의 학습 데이터로 사용되며 도킹결과를 심층신경망의 입력으로 사용하여 일반적인 도킹보다 우수한 예측 결과를 보여주는 연구도 있다[9, 11, 20, 21, 26]. 이렇듯 능동 학습 알고리즘을 통해 주어진 질병의 타겟에 대해 유망한 활성을 가진 리드물질을 발굴한다.

2.3.4. 후보 약물 최적화

유효 물질을 도출하고, 도출된 화합물에 대한 합성 가능성 및 효능 등에 대한 평가(Scoring)를 거쳐 최종후보 물질을 도출하여 임상 시험을 위한 단계에서도 AI 기술이 활용된다. 리드 화합물 최적화 과정을 통해 약물 경로예측을 포함한 효과적인 약물 전달 시스템이 개발될 때 성공적인 약물전달체계 구축을 위해서는 약물 화합물 성분의 구성뿐 아니라 생체 적용 환경도 고려해야 한다. 이러한 과정에서 AI는 독성 예측, 전임상 시험 설계 및 실험 결과 분석의 정확성을 높이는 데도 사용된다[16, 17, 27].

2.3.5. 임상 시험

최종 약물 후보는 임상 시험을 거쳐 상용 의약품으로 승인받게 된다. 성공적인 임상 시험을 위해 가장 핵심적인 부분은 적절한 임상 대상자를 선별하는 일이다. 이러한 임상 시험 설계 과정에서도 최적화를 위해 AI 기술을 사용할 수 있고 이를 통해 분석결과의 정확성을 높여 임상 시험의 효율성과 신뢰성을 향상할 수 있다. 또한, 신약이 상용화된 후 약물 부작용이나 관련 정보 분석 등 약물 모니터링 과정에서도 AI를 사용하여 약물 감시 모니터링의 효율성을 높일 수 있다[7-9].

Ⅲ. 신약개발에 사용되는 인공지능 기술

3.1. 머신러닝

머신러닝은 지도학습(supervised learning), 비지도(unsupervised learning)학습 및 강화(reinforcement)학습으로 분류된다. 지도학습은 알려진 레이블(목표값)이 있는 훈련 데이터를 학습하여 이를 기반으로 새로운 데이터의 레이블을 결정하고, 비지도학습은 레이블이 없는 자료의 내재한 패턴을 인식한다. 일반적으로 고차원 데이터는 패턴인식 전에 자료의 정보를 많이 잃지 않으면서 간소화하는 차원 축소를 시키는데 저차원 공간에서 더 효율적이고 패턴인식도 용이하게 해준다. 지도와 비지도 학습이 혼합된 반지도(semisupervised)학습 또는 전이(transductive)학습은 적은 양의 레이블이 있는 데이터와 레이블이 없는 데이터를 학습하는 방법으로 크기가 작고 불균형한 데이터에서 학습 정확도를 높이는 방법으로 사용된다. 지도학습은 입력-출력 학습 데이터를 기반으로 예측 모델을 개발하는 분류(classification)와 회귀(regression) 방법이 있는데 분류방법은 질병 진단에, 회귀 방법은 약물 효능 및 ADMET 예측에 주로 사용된다. 가장 많이 사용되는 지도학습 알고리즘은 회귀 분석, k-최근접 이웃(kNN), 베이지안 확률 학습, SVM, 랜덤 포레스트 및 신경망이다 [11].

비지도 학습은 레이블이 없는 입력 데이터만을 기반으로 하는 클러스터링(clustering) 및 특징 찾기(feature-finding) 방법인데 특징 찾기 방법은 질병에 대한 타겟 발견에 주로 사용된다. 화학 정보학에서 클러스터링 알고리즘은 중요하다. 화합물의 구조가 유사하면 유사한 생물학적 활성을 갖기 때문에 클러스터링 알고리즘은 유사한 화합물을 하나의 클러스터로 그룹화하여 원하는 화합물들을 찾는다. 주로 사용되는 곳은 화합물 선택, 가상 라이브러리 생성, HTS(고처리량 스크리닝), QSAR(정량적 구조-활성 관계), ADMET (Absorption, Distribution, Metabolism, Elimination and Toxicity) 예측이다. K-Means, 이등분 K-Means 및 Ward 클러스터링 알고리즘이 많이 사용되는 클러스터링 알고리즘이다. 또한, MCL(Markov Clustering algorithm)과 같은 그래프 기반의 클러스터링은 단백질-단백질 상호작용과 같이 상호 연관된 네트워크 구조에서 관련성이 없는 노이즈 정보를 제거하며 조밀하게 연결된 영역을 감지하는 분할기법이다. 강화학습은 주로 주어진 환경에서의 의사결정 및 성능 극대화를 위해 사용되는 방법으로 새로운 약물 설계와 실험 설계 등에 주로 이용된다[8, 11, 28, 29].

신약개발과정에서 사용되는 데이터는 머신러닝 알고리즘의 성능에 영향을 미치는 매우 중요한 요소이다. 고품질 데이터와 잘 정의된 학습 데이터는 신약개발 성공에 매우 중요하며 정밀의약의 경우 그 의존도는 더욱 높다. 지난 20년 동안 고처리량(high-throughput) 스크리닝 및 시퀀싱, 온라인 다중 오믹 데이터베이스, 머신러닝 알고리즘을 통해 약물 개발에 필요한 데이터 생성, 수집 및 유지 관리 환경이 갖추어져 왔고 데이터 분석력도 크게 향상되었다. 현재 많은 머신러닝 기술과 통합 데이터베이스 사용을 지원되는 소프트웨어 도구가 개발 및 공개되어 신약개발의 모든 단계에서 유용하게 사용되고 있다[8, 16, 28]. 신약개발에 사용되고 있는 주요 머신러닝 기술과 응용사례는 다음 표 1과 같다.

3.2. 딥러닝 (Deep Learning Methods)

머신러닝의 한 분야로 현재 거의 모든 기술 분야에서 최첨단 방법론으로 떠오르고 있는 딥러닝은 신약개발 분야에서도 타겟 및 리드 물질 발견이나 약물 활성 예측 등에 사용되는 대표적인 기술 중의 하나가 되고 있다 [38-43].

딥러닝 알고리즘은 다층의 비선형 처리 장치를 통해 데이터에 포함된 고도의 추상화를 모델링하는 인공신경망(ANN-artifical neural net)을 이용한 학습방법이다. 신경망의 각 층은 기본 처리 단위인 뉴런으로 구성되어 있는데 다층의 뉴런들은 상호 연결된 네트워크를 형성하면 복잡한 연산을 해결할 수 있는 막대한 컴퓨팅 성능

Table. 1 Machine learning algorithm and application for drug discovery

	method	algorithm application for drug discovery
Supervised Learning	random forest	An ensemble method that outputs (majority voting rules) class (classification) or average prediction (regression analysis) from multiple decision trees
		Feature selections, classifiers, and regression in drug discovery. Prediction of Ligand-protein affinity in virtual screening[30, 31]
	naive bayes	Estimation of the probability that given data will be assigned to a specific label based on the prior probabilities and Bayes' rule
		Drug activity prediction: Establishing prior probabilities for a set of biologically active compounds based on the ratio of active and inactive chemical substructures. Estimating drug activity of query structures using prior probability distributions[11, 32, 33]
	k-nearest neighbor	Instance-based learning classified according to the majority rule(similar properties) among the k nearest neighbors of a given object
		(ligand-based) virtual screening, chemical similarity search. prediction of bio activities using a chemical similarity metric as a measure the distance between compounds [8, 11]
	Support vector machine	Low-dimensional hyperplane identification that maximizes data separation using a non-linear kernel. The hyperplane is fit to maximize the margin between support vectors
		QSAR, drug-target interaction, identify a number of structural features of the derivatives, distinguish between active and inactive compounds, ranking compounds in databases[8, 34]
	regression analysis	Relational model between dependent and independent variables. Linear regression for continuous data and logistic regression for categorical data
		QSAR techniques: Normalization for predictive accuracy of QSAR models, dimensionality reduction, solve collinearity and data dimensionality problems using genetic algorithm, determination of the drug-ligand relationship[11, 31]
Unsupervised Learning	clustering	k-means clustering: Minimize the distance to the center within the group and classify the data into k groups Hierarchical clustering: A classification method that builds a hierachy of clusters by agglomerative clustering or divisive clustering (splitting a large cluster to smaller ones) Self Organizing Map (SOM) - vector-based clustering technique suitable for large-scale, high-dimensional feature data Markov Clustering (MCL)- Algorithm to partition the interaction graph by detecting densely connected regions within the interaction graph. Genome-scale data for protein interactions are presented as large networks or graphs of hundreds or thousands of proteins interconnected. Proteins tend to function as groups or complexes, so it is important to reliably identify protein complexes in these graph (network) structures[35]
		primary screen, virtual screening: selection of desirable compounds within large chemical libraries, virtual library creation, QSAR, ADMET prediction Discovery of effective drug repositioning candidates through heterogeneous data integration using multi-clustering and similarity analysis between drugs[11,36,37]
	Principal component analysis	transform a set of correlated features to new independent variables called principal components
		QSAR: used to break down protein structures into modules or to identify specific pathways in a metabolic network. A new emerging 'network pharmacology' interprets the observed pharmacology as a relational paradigm of interactions from the concept of an ideal biological profile of drug molecules and PCA provides a way to predict and quantitatively summarize the most influential factors in these network concepts[38]
Reinforecement	Algorithm that trains a defined agent to select the one with the highest reward among the actions that can be selected in the current state. Reinforcement learning is not learning from data, but iteratively learning behaviors that lead to rewards through experience.	
	Structure-based drug design: designing new drugs of molecules with desired properties In structure-based design, it is difficult to effectively control the properties of the generated molecules, but reinforcement learning can be used to generate molecules with desired properties. Two models are used: the generative model acts as the “reinforcement learning agent” and the property prediction model acts as the “critic” which is responsible for assigning the reward or punishment. The critic assigns a prize and punishment by a set reward function and induces the reinforcement learning agent to have the maximum value[8, 16].	

을 갖게 된다.

신경망은 입력, 은닉 및 출력층으로 구성되는데 이 층들은 다양한 방식으로 연결될 수 있고, 네트워크를 통해 전파되는 입력의 최종 결과는 모든 층의 연결 패턴과 뉴런 활성화 함수 및 뉴런 간의 연결 가중치에 따라 달라진다. 따라서 아키텍처(연결 패턴)와 활성화 기능은 수행할 작업에 대한 네트워크 성능과 최종 결과에 영향을 미치는 중요한 역할을 한다.

은닉층의 수가 많은 경우(보통 수백 층)를 심층신경망(DNN-deep neural net)이라고 하고, DNN은 학습하고 문제를 해결할 수 있는 기능이 있어 이러한 학습을 딥러닝이라고 한다. 각 은닉층이 이전 계층에서 더 많은 기능을 추출하고 고유한 추상 표현을 생성하기 때문에 은닉층의 수가 많아질수록 네트워크는 더 깊어진다. 따라서 복잡한 기능을 해결하려면 숨겨진 정보를 학습할 수 있도록 많은 레이어(층)를 삽입해야 하고 보통 대규모 데이터를 다루므로 수행시간이 오래 걸려 DNN을 사용하기 위해서는 높은 처리 능력, 컴퓨팅 속도, 대용량 데이터베이스 및 병렬 처리가 가능한 소프트웨어 사용환경이 필요하다.

컴퓨터 성능 향상에 힘입은 딥러닝은 신경망 아키텍처의 유연성을 이용해 많은 양의 자료 학습을 기반으로 주어진 질병에 대한 타겟 발굴 및 타겟에 활성을 나타내는 약물 후보물질을 식별할 수 있고 머신러닝 기술과 비교하여 생물학적 활성, ADMET 특성 및 물리화학적 매개변수 예측 등에서 성능이 더 우수하다고 평가되고 있다[42]. 이외에도 딥러닝 알고리즘은 독성 및 안정성 예측과 임상 시험 등 여러 단계에서 도입되고 있는데 2012년에 제약회사 Merck가 약물 특성 및 활동 예측에 DL을 적용하여 성능의 우수성을 보여준 것을 시작으로 독성 영향 예측, 전사프로필(transcriptional profile)에 기반한 약물의 치료 범주 분류, 화학 데이터 모델링 등 신약개발 전 과정에서 적용되고 있다[11]. 최근에는 딥러닝이 새로운 분자 설계 및 분자 특성 추출에 적용되어 원하는 특성을 가진 새로운 분자를 만들어 내는 데도 활용되고 있다.

현재 효율적인 DNN 설계 및 구현을 위해 Keras 및 Gluon과 같은 API(Application Programming Interfaces)와 다양한 오픈 소스 DL 프레임워크가 공개되어 있다[8, 44, 45]. 또한, 전 세계 연구자들이 관련 라이브러리에 대규모로 참여하여 수십억 개의 매개변수를 가진 DNN을 훈련하고 있다.

이러한 개발환경을 잘 활용하면 수행하고자 하는 목표에 적합한 아키텍처를 갖는 DL 신경망을 구축할 수 있다. 가장 많이 사용되는 DNN은 컨벌루션 신경망(CNN), 반복 신경망(RNN), 심층신뢰 신경망(DBN) 등이 있는데 대표적인 DL 방법을 간략히 살펴보면 표2와 같다.

Table. 2 Representative deep learning algorithm for drug discovery

	Algorithm
CNN	It typically goes through a series of convolution, activation, pooling/subsampling and fully connected layers to produce an output: 1) Convolutional layer - the most important part of a CNN, consisting of a set of filters called a kernel. Extracting features from a lower level to a higher level with the goal of extracting various features of the input data 2) A pooling or subsampling layer often immediately follows a convolution layer in CNN. It downsamples the output of a convolution layer along both the spatial dimensions of height and width. 3) The last layer of a CNN is usually a fully connected layer, which is a multilayer perceptron (MLP) that computes the weighted sum of all features in the previous layer using an activation function to classify it into the final target class[46]
RNN	RNN has the capability to learn sequences. The weights are shared across all steps and neurons. It has a characteristic that the connection between nodes has a cyclic structure, which creates an internal state of the network that can display dynamic temporal behavior. Unlike feedforward networks, RNNs contain a feedback element that can feed back the signal of one layer to the previous layer, and have an internal memory, which is used to store long-term information. This memory allows for sequence and contextual awareness, which is advantageous for processing data with characteristics that change over time, such as text, protein sequences, and time series data[47].
DBN	DBN has unidirectional connection and is used in both supervised and unsupervised ML.It is a generative graphical model composed of multiple layers of latent variables ("hidden units") , with connections between the layers but not between units within each layer. The hidden layers of each sub-network serves as visible layer for the next layer. The visible layer receives data and the hidden layer acts as a feature detector to extract features. High Accuracy Drug-Target Protein Interaction Prediction Method based on DBN is proposed using the characteristic vector extracted from drugs and proteins[43, 45].

약물-타겟 스코어링(scoring)은 구조 기반의 약물 설계 파이프라인의 핵심 단계로 적절한 결합 위치를 선택해서 약물-타겟의 결합 친화도를 예측하면서 가상 스크

리닝의 성공 가능성을 평가한다. CNN기반 스코어링은 약물-타겟 복합체에 대해 결합 포즈/친화성 예측과 활성/비활성 감지 측면에서 이전의 방법보다 더 탁월한 성능을 보여주었다 [46].

Ragoza et al.의 CNN 모델은 3D 약물 표적 구조에 대해 학습할 때 올바른 결합 포즈와 잘못된 결합 포즈를 구별할 수 있고 CNN 알고리즘 기반의 스코어링이 바인딩 모양과 친화도를 모두 예측하는 측면에서 Autodock Vina보다 훨씬 더 나은 성능을 보임을 입증했다 [47]. Wallach et al.은 신약개발 응용 분야에서 저분자의 생체활성 예측을 위한 심층 CNN인 AtomNet을 구축했고 [48], 단백질을 3D 이미지로 취급하며 단백질 구조 데이터를 훈련하는 심층 CNN인 DeepSite [49], 3D 그래프 CNN 모델을 사용하여 리간드-단백질 결합 친화도를 예측하는 KDEEP도 개발되었다[50].

Ⅳ. 현재 기술의 한계와 미래 발전 방향

4.1. 단백질 3차 구조에 대한 이해

단백질은 DAN, RNA와 더불어 생명현상의 핵심 물질이다. 생체 내에서 단백질의 기능은 그 구조와 연관되어 있어서 단백질의 3차원 구조 규명은 생명과학 분야의 중요한 기초연구 분야이다. 단백질은 수천 개의 아미노산이 각각 고유한 방식으로 접히면서 입체적인 구조가 형성되는데 단백질 하나의 입체 구조와 접히는 과정을 온전히 이해하는 것은 사실상 매우 어려운 문제이다. 단백질 구조 예측이란 아미노산 서열에만 기반하여 단백질이 채택할 3차원 구조를 예측하는 것인데 이 문제는 지난 50년 이상 동안 도전적 과제였다. 단백질 구조 결정은 오랜 기간의 노력이 필요하므로 신약개발에 병목 현상을 만들어 왔다.

최근에는 구글 자회사의 딥마인드가 딥러닝 기술을 이용해 단백질의 3차원 구조를 예측하는 '알파폴드'(AlphaFold)를 발표하였는데 후속인 알파폴드2는 CASP14에서 출제된 110개의 단백질 구조 예측 문제 중에서 90% 이상 적중하여 기술의 우수성을 보이며 향후 단백질 3차원 구조 규명에 새로운 가능성을 열었다 [24, 25]. 딥러닝은 단백질 시퀀스와 다중서열정렬 (MSA) 특징을 입력으로 받아 컨볼루션 신경망(CNN)을 거쳐 잔기쌍들의 거리(Amino Acid distances)와 비틀림(Chemical bond angles) 분포를 출력한다. 이 두 출력물을 이용해 단백질의 구조를 잘 표현할 수 있는 에너지 함수를 만들고 경사 하강법으로 에너지가 최소화될 때까지 반복해서 에너지 함수를 최적화시키면서 최종적으로 주어진 단백질의 3차원 구조를 예측하는 방식이다.

인공지능 방법의 핵심은 학습 데이터와 인공지능 모델의 파라미터 최적화인데 현재 단백질 구조 데이터베이스인 PDB(Protein Data Bank)에는 인공지능 학습 데이터로 활용될 수 있는 15만개 이상의 단백질 구조가 밝혀져 있다. 알파폴드는 PDB의 단백질 구조를 폴딩 기본 단위로 나눈 CATH 도메인 데이터베이스를 사용하였고 유사도 필터링을 통해 대표적인 29,400개의 단백질 도메인에 대해 학습하였다[12]. 알파폴드의 결과는 단백질의 3차원 구조 모델링에 새로운 가능성을 보여주었지만 아직은 정확한 알고리즘이 알려지지 않아 CASP 자료 이외의 일반적인 단백질 시퀀스에 적용될 수는 없는 상황이다. 최근 딥마인드는 알파폴드를 새롭게 재설계한 버전을 통해 유사 구조가 알려지지 않은 경우에도 원자단위 정확도로 단백질 구조를 규칙적으로 예측할 수 있는 최초의 계산 방법을 제공한다고 발표하였다[25]. 알파폴드는 다중 시퀀스 정렬을 활용하여 단백질 구조에 대한 물리적, 생물학적 지식을 딥러닝 알고리즘 설계에 통합하는 새로운 기계학습 방법이 될 것이라 기대를 받고 있다. 앞으로 알파폴드에 대한 많은 연구가 이루어질 것이고 이를 통해 단백질 3차 구조에 대한 정확한 계산 방법으로서 실용적 도구가 된다면 신약개발을 비롯한 여러 바이오 분야에 혁신을 가져올 것이라 기대되고 있다.

4.2. 학문 분야 융합

신약개발에서 화학과 생물학의 융합은 필수적인데 주목할 점은 생물학은 화학보다 계산적으로 훨씬 더 처리되기 어렵다는 것이다. 예를 들어 화학적 측면에서 수용체에 대한 리간드의 친화도를 결정하는 열역학의 기본 원리는 잘 정의되어 모형화할 수 있지만, 수용체의 구조적 변화나 평형상태, 신호전달과 같은 생물학적 원리는 복잡하여 이해하기 어렵고, 유전자 발현의 변화나 단백질 변형과 같은 일이 발생하면 더욱 복잡해진다. 또한, 화학적 정보는 상호 연관성을 고려하지 않는 10^{63} 정도의 대규모 저분자 화합물을 다루고 화학적 반응이나 열역학과 같은 기본 원리에 대한 규명이 잘 되어 있다.

반면에 생물학은 2300 유전자와 10^{13}개의 세포, 더욱 복잡한 상호관계에 관한 정보를 다루고 분자들과 질병과의 직·간접적인 연관성, 생물학적 시스템의 세포 변화 및 가소성과 이질성 등 정확한 메커니즘을 정의하기 어렵고 알려진 지식도 상대적으로 제한적이다[51]. 이러한 생물학적 특성은 실제 AI 기술을 적용한 신약개발 과정에서 어려운 장벽이 되고 있다.

AI는 그동안 축적된 많은 데이터를 기반으로 학습을 통한 화학적인 모델링에 초점을 두고 발전했는데 실제 약물 적용의 관점에서 보면 약물은 유한개의 매개변수들로 정의하기 어려운 훨씬 더 복잡한 생물학적 시스템에 의해 작용하므로 효능 안전성에 대해 불확실성이 커질 수밖에 없다[10, 51]. 따라서 의미 있는 정량적 변수와 레이블보다는 실제로 어떤 변수가 중요한지 결정하고 이를 실험적으로 어떻게 정의하는지 생물학적 기반에서도 AI 기술개발이 충분히 연구되어야만 한다. 이런 이유로 제약산업에서 AI 기술에 대한 투자가 기대만큼 활성화되지 않고 있다.

그동안 기술적 측면에서 화학적 시스템에 대한 이해가 크게 발전해 왔고 조합 화학의 등장과 스크리닝 기술의 발전 등 더 많은 자료를 생산 및 탐색, 처리할 수 있는 기술은 크게 향상되어 양적 성공을 거둔 것은 확실하다. 그러나 탐색에서 신약 발견으로 이어지기 위해서는 생물학적 시스템의 프로파일링을 통해 생물학적 관계에 대한 이해가 반영된 질적 정보와 자료가 필요하다. 따라서 AI 기술이 신약 개발과 승인에 새로운 흐름을 바꿀 수 있는 정도로 발전하기 위해서는 화학과 생물학, 독성학 및 약동학 등의 광범위한 전문분야의 협력과 융합이 더 이루어져야 한다[4, 5].

4.3. AI 기술적용을 위한 충분한 질적, 양적 데이터의 필요성

신약개발의 모형은 일반적으로 두 가지 유형으로 구분된다. 하나는 실제 존재하거나 아니면 가상 스크리닝을 통해 후보 화합물을 추출하는 작업과 같이 대용량 자료와 예측 측정치에 기반한 모형인데 여기에는 타겟 단백질에 대한 용해도, logD나 생물 활성도와 같은 정성적 특징들에 대한 모형도 포함되어야 한다. 두 번째는 작용원리나 이유에 대한 추론 모형이다. 이런 내용과 방법은 일반화되어 모든 개발과정에서 일괄 적용될 수 없고 개발 경우마다 특화되어 최적화하는 방향으로 진행되어야 한다.

따라서 신약개발과정에서 AI 기술은 일반적으로 적용되지 않으며 모든 경우 상당한 수작업이 필요하다. 효능 및 안전성을 포함한 안전한 치료 지수 달성을 위한 올바른 투여량/PK 관련 사항 등을 포함하여 임상으로 진행될 화합물이 도출되어야 한다. 왜냐하면, 예측 결과나 타당성에서의 아주 작은 차이라도 추후 임상 결과에서 10배, 심지어 100배와 같이 매우 큰 약 효용성의 차이를 가져올 수 있기 때문이다 [51]. 이는 화합물 선택을 위한 예측 모델이 매우 중요하다는 의미이다. 모형 자체가 질적으로 우수하지만, 생체 내에서는 예상과 다른 결과를 가져오는 모형이라면 AI 기술을 이용해 신약개발 속도를 향상한다는 점에서는 한계가 있다. 이러한 이유로 AI 기술을 이용한 예측 결과를 사용하는데 아직은 한계가 있다.

또한, AI 기술은 사용하는 학습 자료에 의존하기 때문에 자료의 질이 중요하며 이 자료에는 그동안 축적된 대용량의 자료뿐 아니라 관련한 생물학적 기초연구 결과들도 충분히 포함되어야 한다. 계산적으로 생물의 복잡함을 100% 구현할 수 없으므로 신약개발에서 AI 기술을 더 효율적으로 사용하기 위해서는 앞으로 학습자료에 대한 더 많은 연구가 필요하다.

머신러닝과 특히 딥러닝은 일반적으로 훈련을 위한 대규모 데이터 세트가 필요하다. 인간 뇌의 학습능력처럼 소량의 데이터로 학습하는 방법에 관한 연구도 현재 많은 관심을 받고 있다. 예를 들어 원샷 학습(one-shot learning), 이를 변형한 매칭 네트워크와 같은 방법은 제한된 데이터를 사용하는 모델들로 소수 데이터 학습에 관한 의미 있는 결과를 보여주었고, 양적 자료를 확보하기 어려운 생물학적 자료에 대한 대안으로 더 연구가 필요한 분야이다[52].

Ⅴ. 결 론

컴퓨팅, 자동화 및 AI 기술 발전의 가속화는 신약개발에도 새로운 혁신을 가져오고 있다. 질병에 대한 타겟 단백질 도출에서부터 임상적으로 활용될 수 있는 약물을 도출하는 신약개발 과정은 고비용의 길고 어려운 과정이다. 인공지능 기술은 약물 설계 및 신약개발 모든 과정에서 유용하게 사용되며 효율성을 크게 향상할 수

있는데 특히, 머신러닝 및 딥러닝은 CADD의 여러 분야에서 이미 그 유용성이 입증되었다.

현재 국내외 대학과 제약 혹은 신약개발 회사들이 많은 공동연구 프로젝트를 수행하고 있고 이미 신약개발을 위한 AI 기술사업화에 성공한 기업들도 등장하며 솔루션 업체들을 중심으로 AI 기반 신약기술시장이 빠르게 열리고 있다. 신약개발 단계별로 필요한 기술이 상이하여 단계별로 핵심요소 기술의 최적화가 필요하며 개발하려고 하는 신약에 따라 특화 및 최적화된 기술로 개발되어야 한다.

사실 효능 및 안전성과 관련해서 화학적, 생물학적 및 생리학적 데이터를 생성하고 라벨링하는 것은 쉽지 않다. AI 기술의 성공 여부는 학습을 위한 올바른 형식의 올바른 데이터에 달려있어 신약개발 프로세스의 의미있는 변화를 가져오려면 학습을 위한 자료가 중요하다. 현재까지 축적된 데이터가 제공하는 정보의 한계를 넘어서 생체 내 안전성 및 효능과 관련된 정보도 제공할 수 있는 데이터가 확보되어야 AI 기술이 한 단계 더 발전해 갈 수 있을 것이다. 또한, 신약개발에서 단백질 구조는 매우 중요한데 알파폴드 발표로 기술적인 측면에서 AI 기술이 단백질 폴딩 문제해결을 위한 새로운 가능성을 보여주었는데 AI 기술의 성공적 적용을 위해서는 앞으로 더 많은 단백질서열 데이터와 구조정보가 축적되어야 하고 이와 관련된 기술의 발전도 필요하다.

AI 기술은 데이터 학습을 통해 주어진 질문에 답을 주는 방식으로 질병에 대한 이론이나 가설 아래 관련 데이터를 생성하고 이 자료로부터 우리가 원하는 답을 찾는 방식이다. 이 과정에서 잘못된 생각에서 나온 목표나 부적합한 표현과 잘못된 데이터는 기술적으로는 수정될 수 없으므로 AI 기술을 제대로 활용하기 위한 데이터 생성 기반이 잘 확립되어 있어야 한다. 이 과정에서 생물학적 메카니즘에 대한 이해가 필수적이며 생물학적 관점에서 생체 내에서 성공적으로 작동하는 최종 약물을 만들어 내는 노력이 필요하다.

인공지능 기술과 접목된 신약개발이 직접 임상의 성공으로 이어지지 못한다고 할지라도 이 과정에서 질병에 관한 더 많은 정보와 정확한 진단, 더 나은 바이오마커, 유전적 발견 및 병태생리학 지식의 향상 등 개발과정에서도 많은 발전을 가져올 수 있다. 이러한 결과는 궁극적으로 더 나은 의약품 생산에 기여할 것이므로 앞으로 화학과 생물학이 IT 기술 분야와 잘 융합된다면 인공지능 기술은 신약개발에 새로운 혁신의 동력이 될 것이다. 향후 AI의 효율적 활용은 임상 실패를 최소화하며 빠르고 비용이 절감된 새로운 신약개발 프로세스의 시대를 열어가게 될 것으로 기대하고 있다.

REFERENCES

[1] K. Mak and M. Pichika, “Artificial intelligence in drug development: present status and future prospects,” *Drug Discovery Today*, vol. 24, no. 3, pp. 773-780, Mar. 2019.

[2] N. Fleming, “How artificial intelligence is changing drug discovery”, *Nature*, vol. 557, pp. 55-57, May. 2018.

[3] G. Hessler and K. Baringhaus, “Artificial Intelligence in Drug Design”, *Molecules*, vol. 23, no. 10, pp. 2520, 2018.

[4] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, “The rise of deep learning in drug discovery,” *Drug Discovery Today*, vol. 23, no. 6, pp. 1241-1250, Jun. 2018.

[5] I. M. Kapetanovic, “Computer-aided drug discovery and development (CADDD): in-silico-chemico-biological approach,” *Chem. Biol. Interact*, vol. 171, pp. 165-176, 2008.

[6] S. P. Leelananda and S. Lindert, “Computational methods in drug discovery,” *Beilstein J. Org. Chem.*, vol. 12, pp. 2694-2718, 2016.

[7] L. Zhang, J. Tan, D. Han, and H. Zhu, “From machine learning to deep learning: progress in machine intelligence for rational drug discovery,” *Drug Discovery Today*, vol. 22, no. 11, pp. 1680-1685, Nov. 2017.

[8] L. Patel, T. Shukla, X. Huang, D. Ussery, and S. Wang, “Machine Learning Methods in Drug Discovery,” *Molecules*, vol. 25, pp. 5277, 2020.

[9] S. Woo, “Drug Discovery Enhanced by Artificial Intelligence,” *Biomedical Journal of Scientific & Technical Research*, vol. 12, no. 1, Dec. 2018.

[10] A. Bender and I. Cortés-Ciriano, “Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet,” *Drug Discovery Today*, vol. 26, no. 2, pp. 511-524, 2021.

[11] Y. Lo, S. E. Rensi, W. Torng, and R. B. Altman, “Machine learning in chemoinformatics and drug discovery,” *Drug Discovery Today*, vol. 23, no. 8, pp. 1538-1546, 2018.

[12] K. T. Nho and S. J. Lee, “chemoinformatics for drug discovery,” *Journal of Scientific & Technological Knowledge Infrastructure*, no. 3, pp. 68-75, 2000.

[13] A. Varnek and I. Baskin, “Machine learning methods for

property prediction in chemoinformatics: Quo Vadis?," *Journal of Chemical Information and Modeling*, vol. 52, pp. 1413-1437, 2012.

[14] E. J. Bjerrum, "SMILES enumeration as data augmentation for neural network modeling of molecules," *arXiv: 1703.07076*, 2017.

[15] D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742-754, 2010.

[16] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao, "Applications of machine learning in drug discovery and development," *Nature Review of Drug Discovery*, vol. 18, no. 6, pp. 463-477, Jun. 2019.

[17] E. Gawehn, J. Hiss, and G. Schneider, "Deep Learning in Drug Discovery," *Molecular Informatics*, vol. 35, no. 1, pp. 3-14, Jan. 2016.

[18] Y. Yang, S. J. Adelstein, and A. I. Kassis, "Target discovery from data mining approaches," *Drug Discovery. Today*, vol. 14, pp. 147-154. 2009.

[19] T. Katsila, G. A. Spyroulias, G. P. Patrinos, and M. Matsoukasa, "Computational approaches in target identification and drug discovery," *Computational and Structural Biotechnology Journal*, vol. 14, pp. 177-184, 2016.

[20] T. Huang, H. Mi, C. Lin, L. Zhao, L. Zhong, F. Liu, G., A. Lu, and Z. Bian, "MOST: most-similar ligand based approach to target prediction," *BMC Bioinformatics*, vol. 18, no. 1, 2017.

[21] T. Rodrigues and G. J. Bernardes, "Machine learning for target discovery in drug development," *Current Opinion in Chemical Biology*, vol. 56, pp. 16-22, 2020.

[22] D. Gao, Q. Chen, Y. Zeng, M. Jiang, and Y. Zhang, "Application of Machine Learning on Drug Target Discovery," *Current Drug Metabolism*, 2020.

[23] R. Chen, X. Liu, S. Jin, J. Lin, and J. Liu, "Machine Learning for Drug-Target Interaction Prediction," *Molecules*, vol. 23, no. 9:2208, 2018

[24] G. Y. Joo, "AlphaFold: Ai-based protein 3D structure," Technical Report, Apr. 2019.

[25] AlphaFold [Internet]. Available: https://deepmind.com/blog/alphafold/.

[26] J. Pereira, E. Caffarena, C. Dos Santos, and C. N. Boosting, "Docking-Based Virtual Screening with Deep Learning," *J. Journal of Chemical Information and Modeling*, vol. 56, pp. 2495-2506, 2016.

[27] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. Koes, "Protein–ligand scoring with convolutional neural networks," *Journal of Chemical Information and Modeling*, vol. 57, pp. 942-957, 2017.

[28] C. Réda, E. Kaufmann, and A. Delahaye-Duriezade, "Machine learning applications in drug development," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 241-252, 2020.

[29] M. Elbadawi, S. Gaisford, and A. W. Basit, "Advanced machine-learning techniques in drug discovery," *Drug Discovery Today*, vol. 26, no. 3, pp. 769-777, Mar. 2021.

[30] V. Svetnik, A. Liaw, C. Tong, J. Culberson, R. Sheridan, and B. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, pp. 1947-1958, 2003.

[31] K. Lee, M. Lee, and D. Kim, "Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server," *BMC Bioinformatics*, vol. 18, no. 567, 2017.

[32] H. Sun, "A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing," *Journal of Medical Chemistry*, vol. 48, pp. 4031-4039, 2005.

[33] C. Ratanamahatana and D. Gunopulos, "Feature selection for the naive bayesian classifier using decision trees," *Applied Artificial Intelligence*, vol. 17, no. 5-6, 2003.

[34] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273-297, 1995.

[35] J. D. MacCuish and N. E. MacCuish, "Clustering in Bioinformatics and Drug Discovery," *CRC Press*, 2019.

[36] G. Mohamed, M. Hamdy, and B. Ashraf, "Clustering of chemical data sets for drug discovery," *2014 9th International Conference on Informatics and Systems*, Dec. 2014.

[37] A. Giuliani, "The application of principal component analysis to drug discovery and biomedical data," *Drug Discovery Today*, vol. 22, no. 7, pp. 1069-1076, 2017.

[38] A. Rifaioglu, H. Atas, M. Martin, R. Cetin-Atalay, V. Atalay, and T. Do˘gan, "Recent applications of deep learning and machine intelligence on in silico drug discovery, methods, tools and databases," *Brief Bioinformatics*, vol. 20, pp. 1878-1912, 2019.

[39] I. Baskin, D. Winkler, and I. Tetko, "A renaissance of neural networks in drug discovery," *Expert Opin. Drug Discov.*, vol. 11, pp. 785-795, 2016.

[40] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1-55, Jan. 2009.

[41] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning,"

Nature, vol. 521, pp. 436-444, 2015.

[42] I. Goldberg, "Deep Learning in Drug Discovery and Medicine; Scratching the Surface," *Molecules*, vol. 23,no. 9:2384, 2018.

[43] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, "The rise of deep learning in drug discovery," *Drug Discovery Today*, vol. 23, no. 6, pp. 1241-1250, Jun. 2018.

[44] K. Manish, N. Abhigyan, T. P. Singh, A. S. Ethayathulla, and P. Kaur, "Evolving scenario of big data and Artifcial Intelligence (AI) in drug discovery," *Molecular Diversity*, vol. 25, pp. 1439-1460, 2021.

[45] A. Arabi, "Artificial intelligence in drug design: algorithms, applications, challenges and ethics," *Future Drug Discovery*, vol. 3, no. 2, 2021.

[46] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Unsupervised learning of hierarchical representations with convolutional deep belief networks," *Communications of the ACM*, vol. 54, pp. 95-103, 2011.

[47] J. Yasonik, "Multiobjective de novo drug design with recurrent neural networks and nondominated sorting," *Journal of Cheminformatics*, vol. 12, no. 14, 2020.

[48] I. Wallach, M. Dzamba, and A. Heifets, "AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery," *CoRR*, 2015.

[49] DeepSite [Internet]. Available: https://playmolecule.org/deepsite/.

[50] J. Jiménez, M. Škalič, G. Martínez-Rosell, and G. De Fabritiis, "K DEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks," *Journal of Chemical Information and Modeling*, vol. 58, no. 2, pp. 287-296, 2018.

[51] A. Bender and I. Cortes-Ciriano, "Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 2: a discussion of chemical and biological data," *Drug Discovery Today*, vol. 26, no. 4, pp. 1040-1052, Apr. 2021.

[52] R. C. Mohs and N. H. Greig, "Drug discovery and development: role of basic biological research," *Alzheimer's & Dementia*, vol. 3, no. 4, pp. 651-657, 2017.

정명희(Myunghee Jung)

1991년: Univ. of Texas, 석사
1997년: Univ. of Texas, 박사
1997-1998: 삼성 SDS 선임연구원
1998-현재: 안양대학교 소프트웨어공학과 교수
※관심분야 : 영상처리, 알고리즘

권원현(Wonhyun Kwon)

1985년: 연세대학교, 석사
1990년: 연세대학교, 박사
1984-1994: 삼성전자 선임연구원
1994-현재: 안양대학교 정보전기전자공학과 교수
※관심분야 : 무선통신, 전파전파