연합학습 기반 신약개발 가속화 프로젝트

2. 데이터 활용 신약개발 협력 체계구축

2024. 08.20

AI 신약팀, 대웅제약
신승우 (swshin017@daewoong.co.kr)

– 연구 개발: 기관 대웅제약

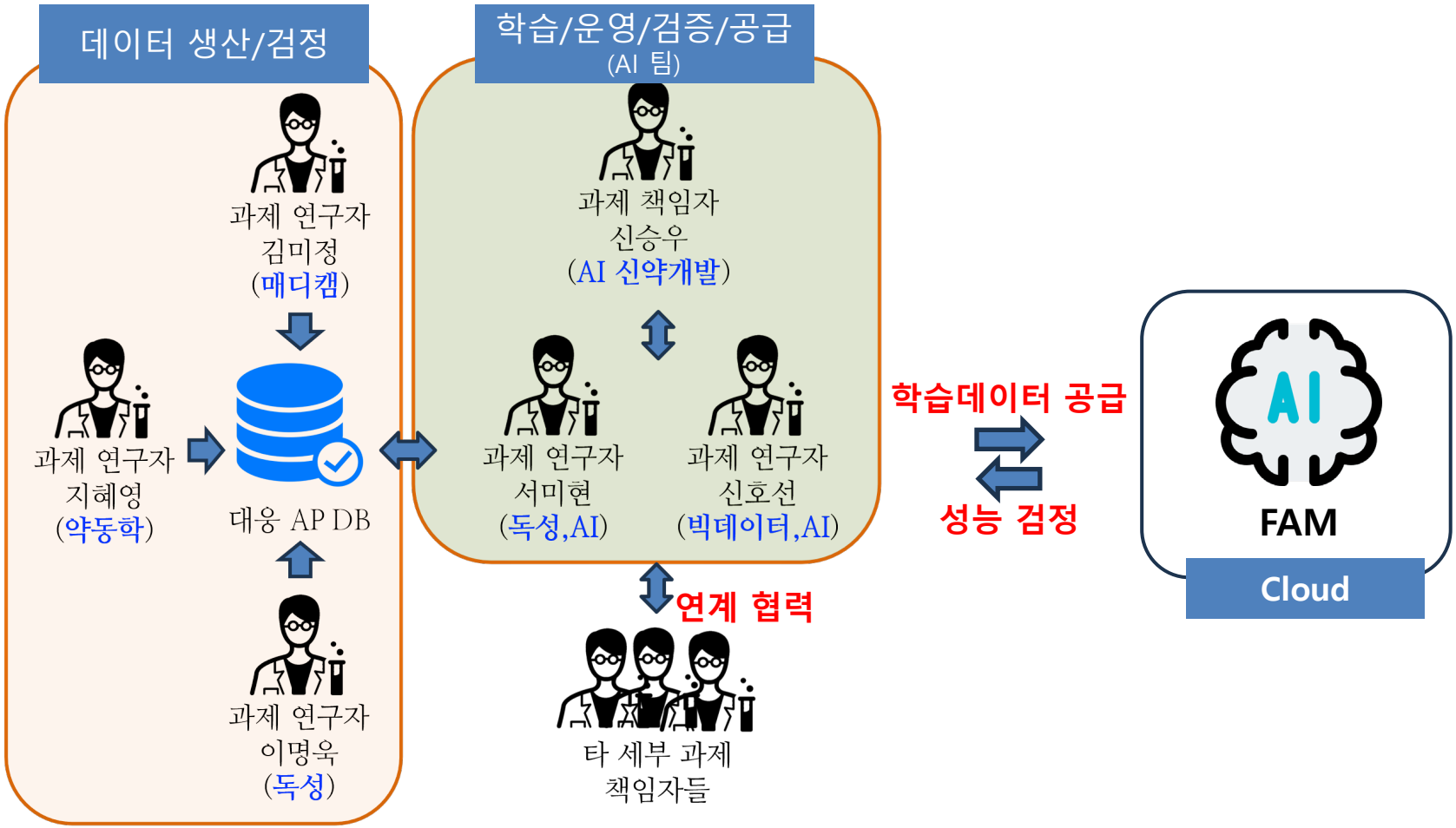
대웅제약의 ADME/T 데이터 보유 및 활용 현황

- 총 **8억 종의 화합물 정보를 대웅 제약 내 구축**
  → DAVID: Daewoong Advanced Virtual Database
- **AI 및 디지털 시스템 구축을 통한 화합물 정보 활용 및 가상 탐색에 적용**
- 본 과제에 필요한 ADME/T 데이터 총 **약 1만여종 보유 및 지속 확대 예정**
- 구축된 DB을 이용하여 **자체 ADME/T 예측 시스템 개발 및 활용 중**
  → **ADAPT**(Advanced Daewoong ADME/T Prediction Tool), **DW Discovery Portal**
- **지속적인 데이터 제공 가능** 및 **연합학습 모델 운용에 필요한 역량을 보유**

– 연구자 소개

◦ 과제 수행을 위한 각 분야 전문가들로 구성

→ **FAM관련 운영 3명**(AI신약 개발, 독성 AI, 빅테이터 AI), **데이터 생산/검정 3명** (약동학, 독성, 매디캠)



– 연구 개발 목표

1. 연합학습 기반 플랫폼(FDD: Federated Drug Discovery)을 위한 데이터를 **수집·가공하는 연구 개발을 지원**

2. FDD 플랫폼에서 실행되는 ADMET 및 PK 파라미터 예측 AI 솔루션(FAM: Federated ADMET Model)의 **운영 및 성능 개선을 위한 연구 지원**

_ 핵심 연구 내용

1. 타 세부 과제와의 연계 협력
   → 학습 데이터 포맷정의, 전 처리 방안 등

2. FAM 태스크 정의 및 데이터 파악
   → 필요한 항목 정의, 비공개 학습 데이터 준비

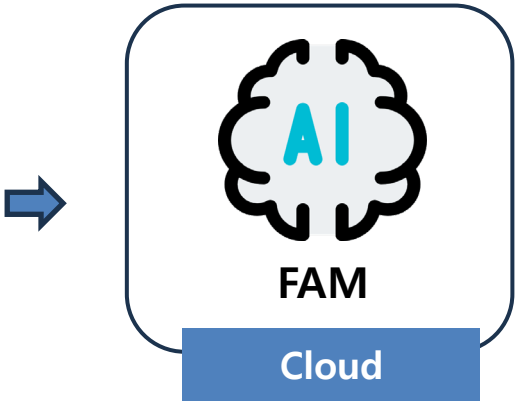
3. 연합학습 플랫폼 데이터 공급
   → 대웅 제약 AP DB을 이용한 데이터 공급, 고통 CRO을 이용한 추가적 데이터 공급

4. 연합학습 참여 및 학습 모델 활용성 검증
   → ADMET 및 PK 데이터를 이용하여 FAM의 예측 결과를 분석 및 검증
   → 연구기관과 주기적인 회의를 통해서 FAM을 검증

대웅 제약 핵심 기술: ADME/T 데이터 보유 및 모델

◦ ADME/T 데이터 총 약 1만 여종 보유 및 지속 확대 예정 (1단계)

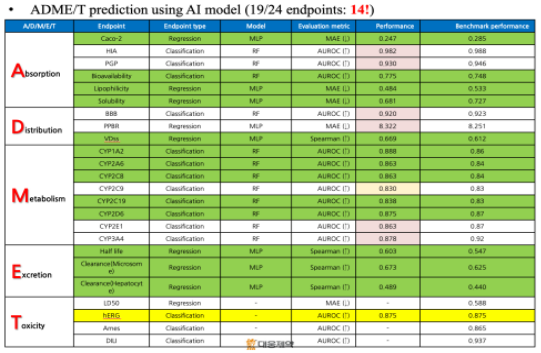
항목	대분류	중분류	소분류(예시)	데이터 보유량
ADME	약물 흡수	물성	● 용해도 예측 (Water, 인공장액, 인공위액, Intrinsic) ● 이온화 상수(pKa) 예측 ● 친유성, 분배계수 예측 (ClogP, logD):	1481
		투과도	● 소장 상피세포 투과도 예측 (Caco-2, MDCK) ● 인공지질막 투과도 예측(PAMPA) ● 뇌장벽투과도 예측(BBB)(Mouse) ● 투과도 예측	99
	약물 분포	분포 용적	● Plasma Protein Binding ● 분포용적 값 예측	113
		수송체 약물 분포 영향력	● OAT family(OATP1B1, OATP1B3, OAT1) 결합친화도 및 기질성 예측 ● OCT 전사인자 or OCT TF(OCT1, OCT2) 결합친화도 및 기질성 예측 ● BCRP 결합친화도 및 기질성 예측 ● BBB Transporter(P-gp 등) 결합친화도 및 기질성 예측	
	약물 대사 및 배설	버퍼 안정성	● Plasma Stability (mouse, human, rat, dog)	395
		대사 안정성	● Metabolic Stability (mouse, human, rat, dog) monkey, minipig ● Microsomal (mouse, human, rat, dog) ● Hepatocyte Stability (mouse, human, rat, dog)	820
독성	Toxicity	심장 독성	● hERG 채널 저해 예측	488
		간 독성	● CYP450 저해 예측(CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP3A4,. 등)	908
		발암성	● Ames Test 결과 예측 ● 발암성 예측 ● 항원성 독성 시험	
		생식, 내분비 독성	● 생식독성 예측 ● 유전 독성(돌연변이, 세포염색체 등)	
		세포 독성	● 정상세포 독성	
약효	약효	인산화효소 (단순 약효)	● 약효 측정(JAK1,EGFR, 등)	4300
		GPCR	● GPCR 선택성 예측	
약동학	약동학	약동학 파라미터	● AUC : Area under plasma concentration-time curve ● Cmax : Maximum plasma concentration ● Tmax : Time to maximum plasma concentration ● T1/2 : Elimination half-life ● CL : Clearance ● V : Volume of distribution ● F : Bioavailability	411



◦ ADPAT의 실행 화면과 성능

→ 24개 ADMET 및 PK 항목 중 **14개 항목이 다른 예측 프로그램 보다 더 높게 또는 동등(TDC site기준)**

→ **ADME/T 모델 개발에 풍부한 경험을 가지고 있음**



1. ADPAT 예측 성능 결과

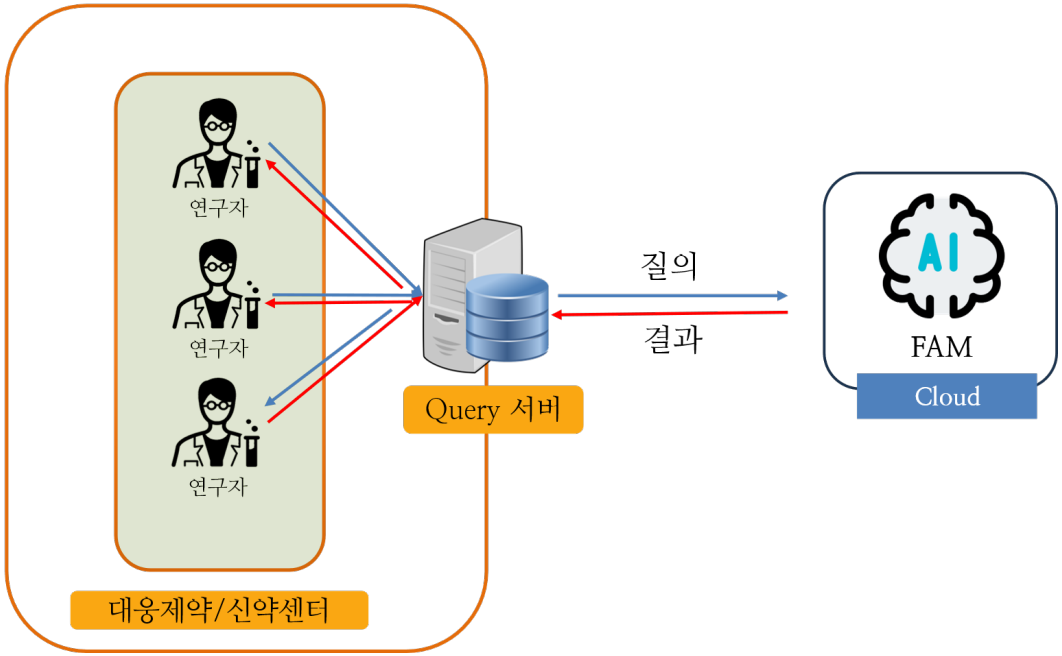
– 추진 전략

추진내용	추진 일정												결과물
	1	2	3	4	5	6	7	8	9	10	11	12	
1차 년도													
기관내 보유 데이터 현황 조사							■	■	■				화합물 현황서 및 DB (문서1건, DB1건)
데이터 포맷 및 표준화 작업							■	■					표준화 문서 (1건)
FAM 초기 학습 데이터 설정									■	■	■		공개 데이터셋 (1건)
전 처리 도구 개발 지원									■	■	■		전처리 플랫폼 (1개)
FAM의 항목 정의 및 태스크 정의							■	■	■	■			항목 정의서 및 태스크 정의서 (2건 이상)
FAM에게 초기 학습할 데이터 공급											■	■	공급 데이터 셋 (1건)
2차~ 3차년도													
FAM에게 학습할 데이터 공급			■			■			■			■	4건
CRO 활용한 데이터 확보					■			■				■	확보 데이터(2건)
학습 모델 활용성 모니터링 및 검증			■			■			■			■	성능 모니터링 결과 분석서 (4건)
4차~ 5차 년도													
학습 모델 활용성 모니터링 및 검증		■			■			■			■		성능 모니터링 결과 분석서 (4건)
CRO 활용한 데이터 확보					■			■					확보 데이터(2건)
FAM 활용성 분석 및 검증		■			■			■			■		검증 결과 분석서 (4건)
신규 태스크 발굴								■	■	■			신규 태스크 발굴서 (2건 이상)

– 산출물

◦ 연합학습으로 만들어진 모델: FAM 모델을 이용하여 시간 및 비용 절감

→ **FAM모델을 cloud 또는 localization하여 연구자들이 지속적으로 이용 가능하게 함**



◦ 향후 연구를 위한 비임상 데이터 생산