2026년도 「연합학습 기반 신약개발 가속화 프로젝트(K-MELLODDY)」 사업 개요서

2026.4 K-MELLODDY사업단

□ 사업명: 연합학습 기반 신약개발 가속화 프로젝트 (K-MELLODDY)

○ 과학기술정보통신부 보건복지부 공동 추진 (24.4 ~ 28.12, 5년간, 348억원)

□ 사업 목표

○ 연합학습 기반 신약개발(FDD, Federated Drug Discovery) 가속화 플랫폼 구축

○ 연합학습 기반 ADMET 예측 모델(FAM, Federated ADMET Model) 개발

□ 사업 필요성

○ 신약개발에서 ADMET 예측의 중요성

- ADMET(Absorption, Distribution, Metabolism, Excretion, Toxicity; 흡수, 분포, 대사, 배설, 독성)은 임상시험 성공의 가장 중요한 요소이며, 신약개발 R&D 비용의 22%를 차지함 (NIH)
- 현재 in-vitro 실험 결과 예측 AI 모델은 다수 구축되어 있으나 성능에 한계가 있고, 이 결과만으로 in-vivo(전임상) 및 임상시험 통과를 예측하기 어려움
- ADMET 임상시험 통과 예측은 임상시험(in-human) 데이터를 학습에 사용해야 하나, 데이터 부족 및 공유가 거의 불가하므로 모델 구현이 매우 어려운 상황임
- 기존 예측 모델의 한계를 극복하고 연속적으로 성능 개선이 가능한 in-vitro, in-vivo 및 임상시험(in-human) 전주기적(Longitudinal) 데이터 기반의 FAM 개발 필요

○ 연합학습의 중요성

- 바이오·헬스 분야에서 우수한 AI 모델 개발의 가장 큰 걸림돌이었던 개인정보보호, 데이터 보안, 다양한 데이터 확보 등 문제를 해결하는 연합학습(Federated Learning) 기술을 구글이 제안 (2017)
- 연합학습은 여러 기관이 보유한 데이터의 직접 공유 없이, 기계학습 모델 파라미터만 공유하여, 데이터 프라이버시를 보호하면서 AI 모델 성능을 협력하여 개선하는 방법임 (민감정보의 ‘보호’와 ‘활용’이 동시에 가능)

□ 사업 구성

○ 플랫폼 구축(1개 과제): 연합학습 기반 신약개발(FDD) 플랫폼을 구축하고 FAM 솔루션을 운영

○ 데이터 공급·활용(20개 과제): 제약사, 병원, 연구소 등이 데이터를 공급하고 FAM 활용

○ AI 모델 개발(15개 과제): FAM 솔루션 및 응용 모델 개발

- 1차년도부터 3차년도까지 매년 5개 선정

○ 운영협의체 : 사업 운영에 필요한 세부사항 논의

○ 실무위원회 : 전체 과제 참여기관이 참여하여, 과제 수행에 대한 아이디어 수렴, 해결 방안 모색, 신기술 학습, 태스크 발굴 등을 논의함

○ 연합학습 워크숍(FEDTalk) : 사업 참여자 또는 외부 전문가의 강연을 통한 화학, 생물학, 약학, 의학, AI 등 폭넓은 분야의 이해

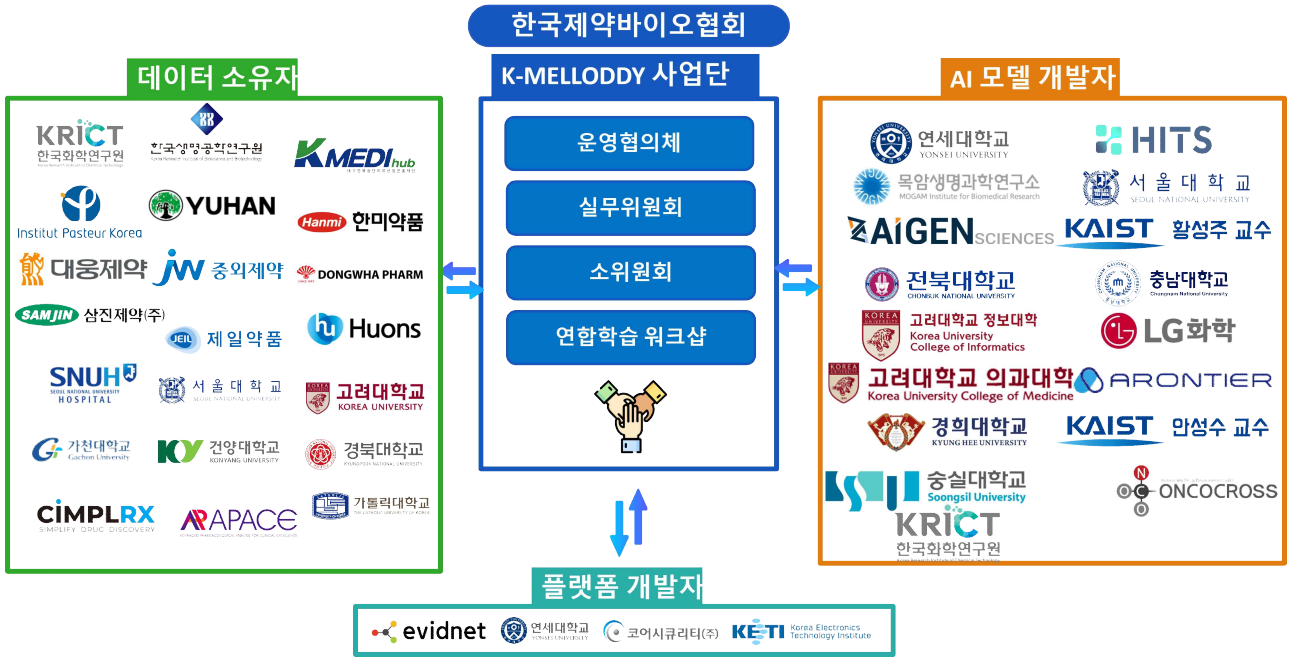

그림 1. 사업 구성 및 참여기관(26년 4월 기준)

□ 세부 사업별 역할

○ (1세부) 플랫폼 구축 및 개발 (플랫폼 개발자)
- 연합학습 기반 신약개발 플랫폼(FDD)을 구축하고 운영함
- FAM 솔루션 운영에 필요한 기능과 환경을 제공함

○ (2세부) 신약개발 데이터 활용 및 품질관리 (데이터 소유자)
- FAM 개발에 필요한 실험 데이터를 공급함
- 데이터 품질관리와 FAM 활용에 참여함

○ (3세부) 연합학습 플랫폼 활용 활성화 (AI 모델 개발자)
- FAM을 개발하고 성능을 고도화함
- 플랫폼과 연계하여 연합학습 기반 검증 및 활용을 수행함

□ 추진 전략

○ 제약사, 병원, 연구소가 보유한 실험 데이터를 직접 공유하지 않고 연합학습 방식으로 공동 활용함

○ 연합학습 기반 ADMET 세부 태스크를 정의하고, 목적에 맞는 데이터를 준비하여 FDD 플랫폼과 FAM 솔루션을 개발함

○ 과제 참여 기관의 기술 교류를 위한 연합학습 워크샵에 참여하고, 데이터, 솔루션, 플랫폼에 대한 상호 이해와 협력 방안을 논의하며 이를 과제 수행에 반영함

○ (플랫폼 개발자, 세부1)
- 데이터 소유 기관의 안전한 데이터 관리, 편리한 인터페이스 개발, FDD 구축 및 운영을 담당
- 데이터 기여에 대한 보상을 공정하게 산정하는 시스템을 구축하여, 데이터 기반 신약개발 생태계의 선순환 구조 구축과 개방적 협력이 가능하게 함
- 플랫폼은 안전한 클라우드에서 운영되며, 데이터 소유 기관, AI 모델 공급기관에 제공되는 로컬 작업공간 (독립 컨테이너)는 보안 클라우드로 제공
- 클라우드 이용 비용은 플랫폼 개발사, 데이터 소유 기관, 모델 개발기관에서 일정 비율로 클라우드 제공 기업에 지출하고 매년 사용량에 따라 조정

○ (데이터 소유자, 세부2)
= 기존의 실험 데이터 및 추가 실험 데이터 제공
- 참여 기관 유형에 따라 각각 어떤 데이터를 공급할 수 있는지와 FAM에 어떻게 기여할 수 있는지를 제안

○ (AI 모델 개발자, 세부3)

- 모델 개발은 각 기관의 자체 장비를 활용하여 수행함
- 연합학습 테스트는 데이터 보호를 위해 클라우드 또는 온프레미스 환경에서 수행함
- 공개 데이터, 샘플 데이터, 자체 확보 데이터를 활용하여 초기 모델 개발 전략과 향후 활용 방안을 제안함

○ (공통)

- FAM의 향후 기능 확대 활용 방안을 제시

□ FDD 플랫폼

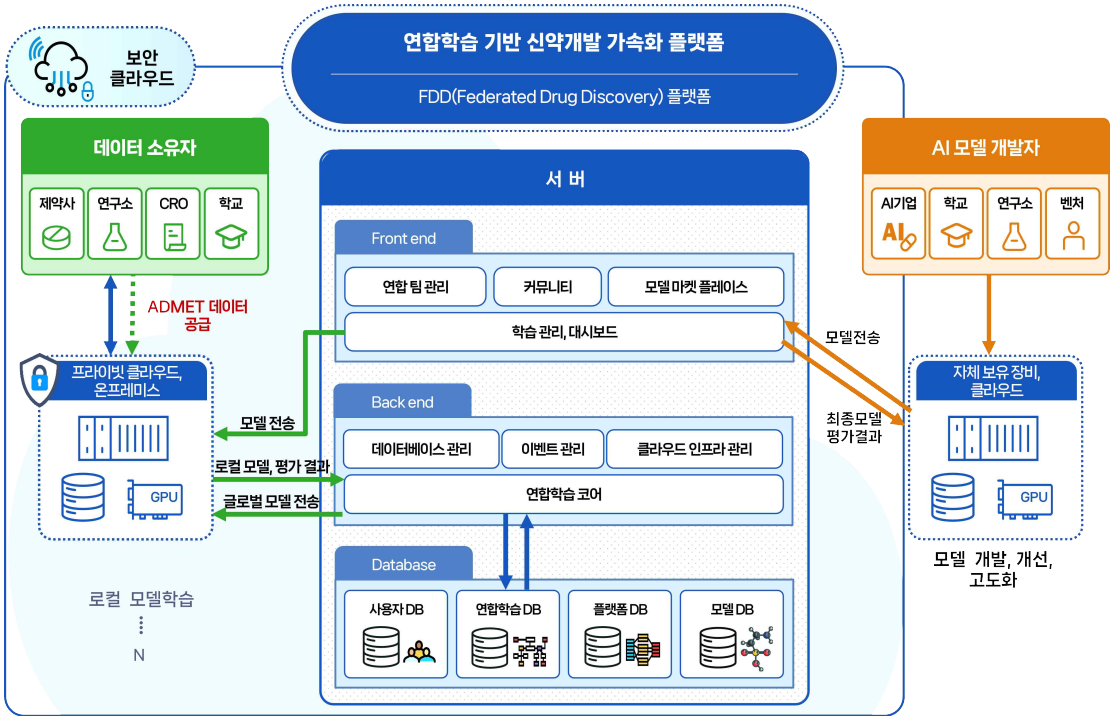

그림 2. FDD 플랫폼 개념도

○ (플랫폼 개발자, 세부1)

- AI 모델 개발자를 위한 인터페이스 개발
  * 태스크별 그룹 생성, 연합학습 요청, 모델 업로드, 학습 결과 분석 등
- 데이터 소유자를 위한 인터페이스 개발
  * 데이터 업로드, 전처리, 학습 모니터링, 학습 결과 조회, 모델 기여도 확인 등
  * 연합학습 과정에 참여하여 로컬 장비 또는 프라이빗 클라우드 상에서 데이터를 관리할 수 있도록 구성
  * 최종적으로 미세조정, 추론 모드 등 모델 활용을 위한 기능 제공

○ (데이터 소유자, 세부2)

- 로컬 장비(프라이빗 클라우드 또는 온프레미스) 사용
  * 자사의 보안 환경(클라우드/온프레미스 장비)에 데이터를 업로드
  * 데이터 전처리, 모델 학습 및 평가, 학습 중간 결과 산출, 미세조정 수행
- 로컬 모델 생성
  * 로컬 장비에서 학습된 로컬 모델의 파라미터(가중치)를 플랫폼 서버로 전송
  * 서버 측에서 글로벌 모델을 구축(연합학습), 다시 글로벌 모델을 각 로컬에서 내려받아, 반복 학습 수행 → 최종 수렴된 글로벌 모델 완성
- 글로벌 모델 미세조정
  * 최종 글로벌 모델 수령 후, 자체 데이터로 모델을 미세조정(Fine-tuning)
  * 미세조정 관련 기능은 세부3(모델개발)에서 제공, 플랫폼 인터페이스는 세부1(플랫폼)에서 제공

○ (AI 모델 개발자, 세부3)

- 초기 모델 개발
  * 데이터 소유자 보유 데이터 중 일부(샘플 데이터) + 공개 데이터(공개 DB 등)를 활용하여 초기 모델을 자체 장비 및 자체 클라우드에서 개발
  * 모델 개발에 전처리 모듈 개발
  * 일정 비율로 공동 부담하는 클라우드 비용은 초기 모델 개발에 사용하지 못하고 연합학습에 공동 사용됨
- 플랫폼 활용
  * 개발된 모델을 플랫폼 인터페이스로 업로드
  * 연합학습 참여: 학습 중간 산출물을 통해, 모델 구조와 성능을 최적화
  * 글로벌 모델뿐 아니라, 각 기관별 데이터로 모델을 미세 조정하는 기능 개발

□ 연합학습 모델 구축 절차와 용어 정의

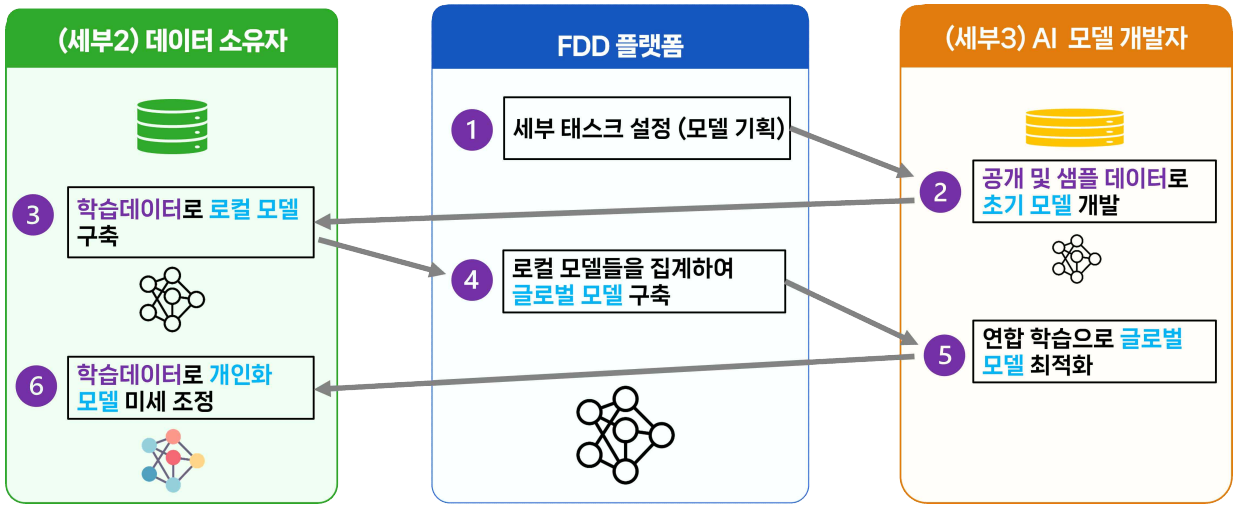

그림 3. 연합학습 모델 구축 절차

○ 데이터 용어 정의

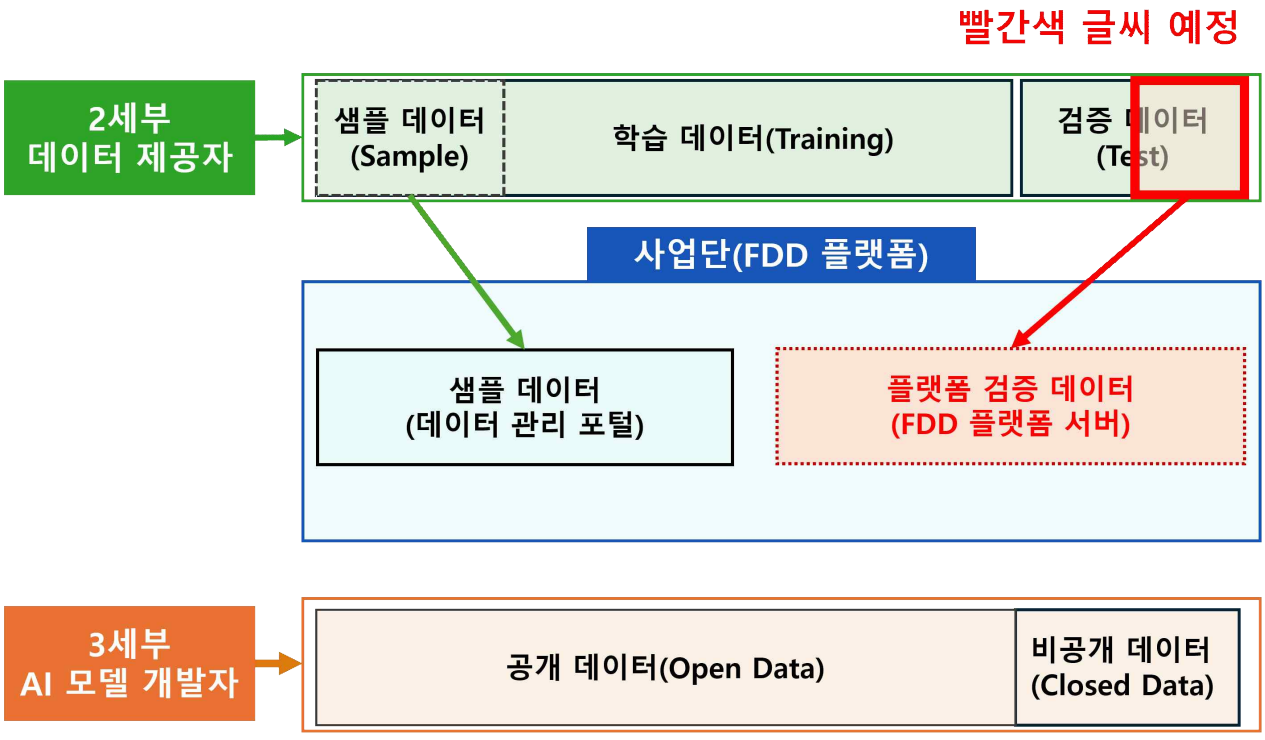

그림 4. 데이터 용어 정의

- **2세부 과제 수행기관이 제공하는 데이터**(데이터 소유자):
- 2세부가 제공하는 데이터는 in-vitro, in-vivo, in-human, PK 데이터에 해당하며, 각 실험 환경과 관련 파라미터를 설명할 수 있는 실험값이어야 함

  * **학습 데이터**(Training Data) : 연합학습에 활용할 데이터로 자체 실험, 위탁 실험 등을 통해 생산된 기관 고유 실험 데이터

    ※ 학습 데이터 자체는 외부로 공유되지 않으며 연합학습에만 사용됨

※ 신약개발 파이프라인 진행 중에 실패한 데이터도 필요하며, 신약 후보에서 탈락한 시점 정보도 필요함

* **샘플 데이터(Sample Data)** : 학습 데이터 중 일부 데이터를 AI 초기모델 개발을 위해서 사업에 참여하는 1, 2, 3세부 과제 수행기관에게 공개하는 데이터

※ 샘플 데이터는 사업 참여자에게만 공개되며 외부로 유출되지 않음

※ 표준 데이터 포맷(마스터 테이블)에 맞게 변환하여 제공해야 함

* **검증 데이터(Test Data)** : 학습 데이터 및 샘플 데이터에 포함되지 않은 데이터로 AI 모델의 성능을 자체 테스트하는 데 사용할 데이터 (데이터 부족 시 추가 데이터 확보 필요)

- **3세부 과제 수행기관이 제공하는 데이터**(AI 모델 개발자) :

* **공개 데이터(Open Data)** : 공개 DB, 논문 데이터 등 누구나 접근가능한 데이터로서 AI 모델 개발에 사용할 데이터

* **비공개 데이터(Closed Data)** : 3세부에서 자체 실험, 문헌 분석, 외부기관 협력 등을 통해 별도로 확보한 비공개 데이터

※ 3세부는 공개 데이터와 비공개 데이터 그리고 2세부에서 제공하는 샘플 데이터를 이용하여 AI 초기모델을 구현하게 됨

※ 3세부는 모델 개발에 사용할 공개 및 비공개 데이터 내역을 제안서에 명시하고 변경 시 사업단에 보고해야 하며 외부 기관 데이터를 활용하거나 구매한 경우 향후 데이터 및 모델 사용에 소유권 문제가 없도록 해야 함

- 플랫폼 검증(벤치마크) 데이터 : 세부 과제 2, 3 수행기관이 보유한 검증 데이터 중 일부를 FDD 플랫폼에 기탁하게 하여, 서버 차원에서 글로벌 연합학습 모델 평가에 활용하는 데이터

○ 모델 용어 정의

- **초기 모델(Baseline Model)** : 연합학습 시작 이전에, 샘플 데이터, 공개 데이터, 비공개 데이터를 활용해 설계한 최초의 AI 모델
- **로컬 모델(Local Model)** : 각 클라이언트(세부 과제2 수행기관)에서 자신의 학습 데이터를 활용하여 초기 모델을 개별적으로 학습한 모델
- **글로벌 모델(Global Model)** : 여러 클라이언트에서 학습된 로컬 모델의 파라미터(가중치)를 집계하여 생성된 통합 모델
- **개인화 모델(Personalized Model)** : 글로벌 모델을 클라이언트의 데이터로 추가 학습(Fine-tuning)하여 최적화한 모델

□ FAM 솔루션 정의

○ FAM(Federated ADMET Model)

- FAM은 in-vitro, in-vivo 및 임상시험(in-human) 데이터를 연계 사용하여 최종적인 endpoint로 ADMET 및 임상시험 PK 파라미터까지 예측하는 모델임 (그림 2, 3 참조)
- 단발성 모델이 아니라, 데이터가 축적될수록 지속적으로 성능을 개선할 수 있는 구조를 지향함

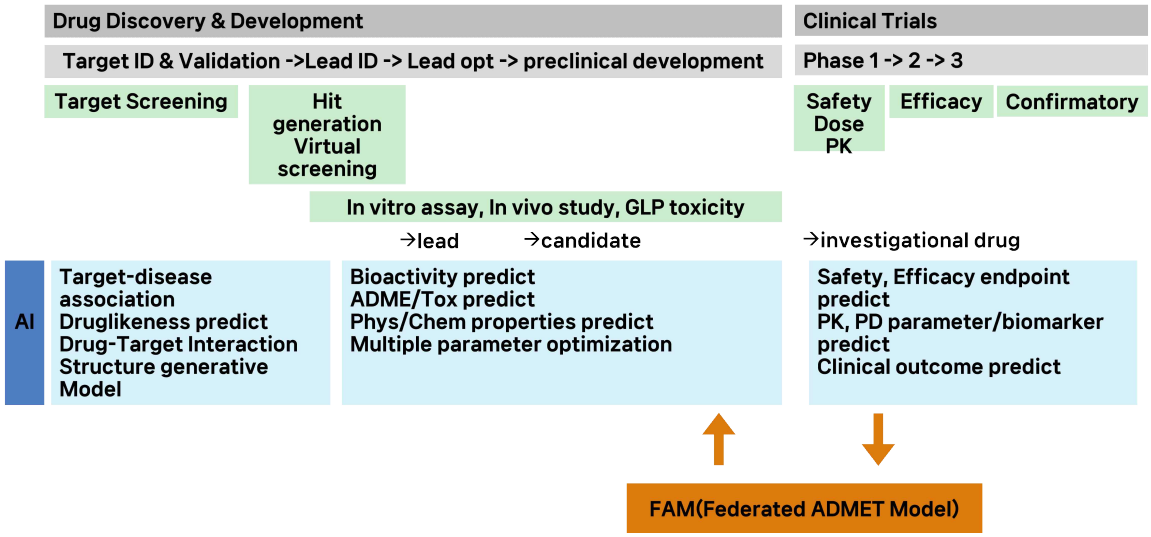

그림 5. FAM 정의: ADMET 및 임상 PK 파라미터까지 예측하는 모델

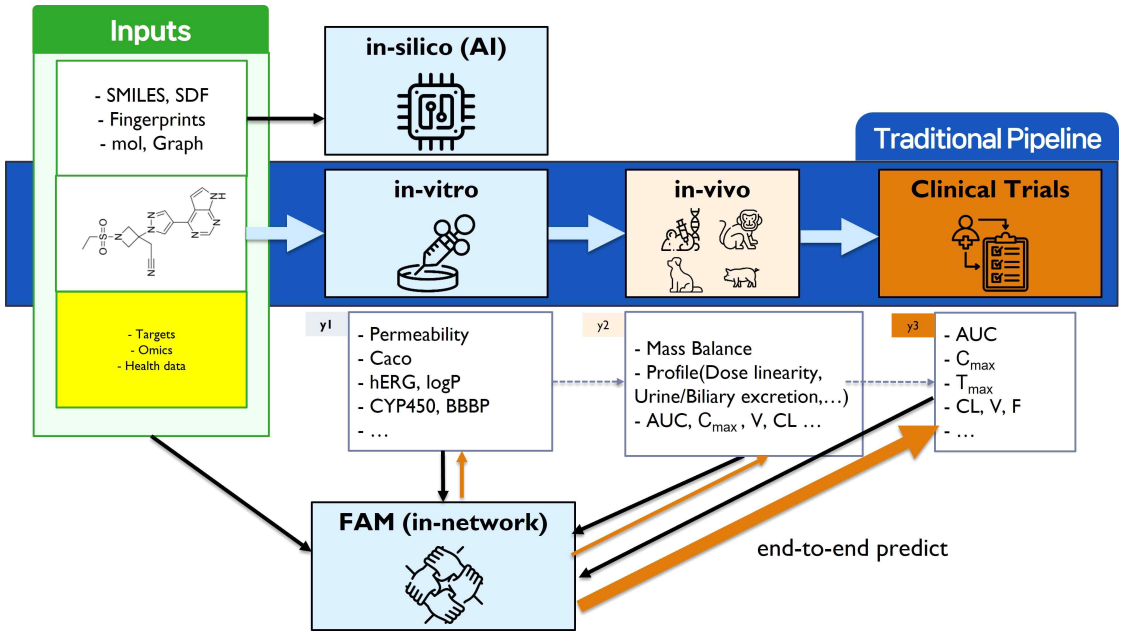

그림 6. 그림 6. FAM 모델 : in-vitro, in-vivo in-human 데이터를 연계 사용하여 endpoint로 ADMET 및 임상 PK 파라미터까지 예측하는 모델 (노란색 데이터는 옵션 입력)

□ K-MELLODDY 주요 태스크

○ 주요 태스크는 데이터 보유량과 중요도가 높은 ADMET 및 PK 지표를 중심으로 결정하였으며, 태스크는 매년 확대되고 있음

표 2. K-MELLODDY 주요 태스크(26년 4월 기준)

Test	Test_Type
Physicochem	logP, ClogP, AlogP, pKa, HBA, HBD, MW
Solubility	Solubility
Permeability	Caco-2, GIT_PAMPA, MDCK
Brain_penetration	Total_BBB, Unbound_BBB, BBB_PAMPA
Plasma_protein_binding	PPB
Efflux_transporter	P_gp, BCRP
Plasma_stability	Plasma_stability
Metabolic_stability	Liver_microsomes, Liver_Microsomes_Phase_II, Hepatocytes, Hepatocytes _Phase_II
CYP_Inhibition	CYP1A1, CYP1A2, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP3A4_MDZ, CYP3A4_TST
Toxicity	hERG, Ames, Cytotoxicity, Genetoxicity
In_vivo_PK	PK_iv, PK_ip, PK_sc : AUClast, AUCinf, Clearance, T1/2, Tmax, Cmax, Co, Vd, Vz, Vss PK_po: AUClast, AUCinf, Clearance/F, T1/2, Tmax, Vd/F, BA, Cmax, Co, MRT
Human_PK	PK_Parameter: Cmax, AUClast, AUCinf, Tmax, T1/2, CL/F, V/F PK_Concentration: Substance, Administration, Concentration, etc

□ 연차별 세부 사업 추진 일정(안)

<table>
<tr><th rowspan="2">연도
세부 과제</th><th>1차년도</th><th>2차년도</th><th>3차년도</th><th>4차년도</th><th>5차년도</th></tr>
<tr><th>2024.7~</th><th>2025</th><th>2026</th><th>2027</th><th>2028</th></tr>
<tr><td>(1세부)
연합학습 플랫폼
구축 및 개발</td><td colspan="5">1개 과제 선정 완료 (4.5년)</td></tr>
<tr><td>(2세부)
신약개발 데이터
활용 및 품질관리</td><td colspan="5">20개 과제 선정 완료 (4.5년)</td></tr>
<tr><td rowspan="3">(3세부)
연합학습 플랫폼
활용 활성화</td><td colspan="3">5개 과제 1차 선정 완료 (2.5년)</td><td></td><td></td></tr>
<tr><td></td><td></td><td colspan="2">5개 과제 2차 선정 (2.5년)</td><td></td></tr>
<tr><td colspan="2"></td><td></td><td colspan="2">5개 과제 3차 선정 (2.5년)</td></tr>
</table>

□ 참고자료

○ 연합학습 프레임워크

- NVIDIA FLARE 프레임워크를 채택(24년)
  * NVIDIA FLARE는 분산된 환경에서 데이터 프라이버시를 보호하면서 협력적으로 AI 모델을 학습할 수 있도록 지원하는 연합학습(federated learning) 프레임워크
  * License: Apache-2.0(오픈소스)
  * 공식 홈페이지: https://developer.nvidia.com/flare
  * Github 페이지: https://github.com/NVIDIA/NVFlare

○ 연합학습 클라이언트 장비 스펙(현재 사용중)

- AWS EC2: g6e.4xlarge
- CPU: AMD EPYC 7R13 Processor ×16
- RAM: 128 GB
- GPU: Nvidia L40s 48GiB vRAM ×1