2026년도 제1차 연합학습 기반
신약개발 가속화 프로젝트 사업
신규지원 대상과제 통합공고 과제제안요구서(RFP)

목 차

RFP 3-1 연합학습 플랫폼 활용 활성화

제안요구서 (세부 사업명)	Federated ADMET 예측 모델(FAM) 개발 (연합학습 플랫폼 활용 활성화)	공모 유형	품목 지정형	기술료 납부 대상	O
사업유형 해당여부	해당 사항 없음				
기획 시 참조 사항	- ADMET와 in-silico ADMET의 이해 (참고문헌 1-11번) - 연합학습의 이해 (참고문헌 12-15번) - 연합학습 적용 사례 (참고문헌 16-23번) - 연합학습 프레임워크 (참고문헌 24번) - 연합학습 기술 연구 동향(참고문헌 25-33번)				

▶ **지원 목적**

○ **AI 신약개발 가속화를 위한 연합학습 기반 신약개발 플랫폼(FDD: Federated Drug Discovery)에서 운영될 ADMET 및 PK 파라미터 예측 AI 솔루션인 FAM (Federated ADMET Model) 개발을 지원**

▶ **지원 대상**

○ **주관연구개발기관은 산·학·연·병 가능**

- 일반적인 사항은 공모안내서의 '신청 요건' 부분 참고

▶ **지원 규모**

지원분야	지원기간	연간 연구개발비 (1차년도)	선정 예정 과제수
연합학습 플랫폼 활용 활성화	3년(3) 이내	300백만원 이내 (150백만원)	5

- **1차년도 연구 기간** 및 연구비는 **6개월 이내**, 예산확보 상황에 따라 **연간 지원 예산 변동 가능**
- 매년 클라우드 비용은 정부 지원 연구개발비 중 일부(1차년도 10%)를 공용비로 배정해야 하며, 1년간 수행 후 사용량에 따라 연차별로 공용비율을 조정함
- 모델 개발에 필요한 장비는 위 공용 클라우드 비용이 아닌 자체 연구비에서 사용하거나 기확보한 장비를 사용해야 함

▶ **필수 연구 내용**

○ **타 세부과제와의 긴밀한 연계 협력(플랫폼 구축 및 개발, 신약개발 데이터 활용 및 품질 관리)**

- 태스크 정의 및 데이터 전처리 표준화 방안 수립을 위한 협의에 참여
- 사업단이 주관하는 운영협의체, 실무위원회 및 연합학습 워크숍에 적극 참여

○ **태스크 정의 및 데이터 준비**

- FAM의 적용 범위를 설정하기 위한 ADMET/PK 예측 태스크 정의 (연차별 확장 및 고도화)
- 모델 개발에 활용할 공개 및 비공개 데이터 현황을 파악 (in-vitro, in-vivo, in-human 데이터)

○ **데이터 전처리 도구 개발**

- 연합학습 환경에서 각 참여기관의 로컬 환경에서 독립적으로 수행하는 전처리 도구 개발 및 배포
- K-MELLODDY 표준데이터포맷을 준수하는 전처리 기능을 제공
  - Feature engineering, 표현 학습(Foundation Model), 데이터 품질 검증 및 이상치 처리 과정도 전처리 도구에 포함됨

○ **FAM 모델 개발 (TRL 5~8)**

- **ADMET 및 PK 파라미터 예측 모델 개발**
  - in-vitro, in-vivo, in-human 데이터를 활용한 ADMET 및 PK 파라미터 예측 모델 개발
    ※ 예: 단일 또는 복수 태스크 예측, PK 파라미터 예측, 중개 모델(예: PBPK, Population PK)
  - 제안기관은 **"K-MELLODDY 주요 태스크"**를 참고하여 모델의 태스크를 구체적으로 제시(4p)
- SOTA(State-of-the-Art) 대비 목표 성능, 벤치마크, 평가지표 및 외부 검증 전략 함께 제시

○ **FAM 운영 및 성능 개선**

- FAM 운영 현황 모니터링, 연합학습 적용 전후의 성능 비교를 포함한 성능 개선 방안 도출
  - 개발한 모델은 2차년도까지 FDD 플랫폼에 탑재하고, 연합학습 기반 성능 검증을 수행
- 기술 검증을 위해 국내외 학술지 논문 게재 및 학술대회 발표를 추진
- 성능 검증을 위해 SW 시험 인증, 사용자 만족도 조사, 실증기관 연계 등을 추진

▶ **선택 연구 내용(계획서에 제안하는 항목을 명시)**

○ **1. 다중 모델 융합 및 모델 연계 활용 기술 개발**

- 태스크별로 개발된 다양한 예측 모델을 조합하여 최적의 성능을 도출하기 위한 기술을 개발
- 다중 모델 융합(Ensemble), 모델 연합(Model Federation), 모델 선택 및 조합, 오케스트레이션(Orchestration) 등의 기술을 포함할 수 있음
  - 예시: Voting, Stacking, Weighted Ensemble, 전문가 혼합 모델(Mixture-of-Experts, MoE 게이팅 기반 동적 선택 네트워크 등

○ **2. 연합학습 집계(Aggregation) 알고리즘 기술개발**

- 신약개발 데이터의 희소성(sparsity), 레이블 불균형, 기관 간 분포 이질성, 태스크 편중, 노이즈 및 배치 효과를 고려한 취합 알고리즘 기술개발
  - 예시: 가중치 기반, 기여도 기반, 개인화된, 클러스터링 기반, 동적 기반 집계 방법 등

▶ **성과목표·지표(안)**

지원분야	성과목표	지원내용(예시)
연합학습 플랫폼 활용 활성화	FAM 개발 *지표별 1건씩	- FDD 플랫폼과 연동되는 FAM 개발 (1건 이상), SW 등록 (1건 이상), SW 공인시험인증 (1건 이상), 특허 출원 또는 등록(1건 이상)
	ADMET 태스크 예측 성능 개선 및 평가 지표 제시	- 기존 예측 도구 대비 성능 개선 목표를 평가 지표와 함께 제시 - 개발할 모델의 차별화 요소를 제시(예: 예측값 외에 오차 범위 신뢰도, 순위 등 추가 정보 제공) - SCI/SCIE급 논문 발표 (2건 이상)
	모델 사용자 매뉴얼 개발	- 모델 사용자 매뉴얼 개발 (1건 이상)
	FAM 활용 방안	- 개발한 모델의 외부 검증 또는 실증 결과 제시 (1건 이상) • 실증기관과의 공동연구, 위탁 연구 등 연계·활용 방안 제시

▶ **특기사항**

○ **연구개발계획서 작성 시 주요 사항**

- 사업공고문에 첨부된 사업개요서를 반드시 숙지하고 계획서 작성
- 최종 산출물의 기술성숙도(TRL)에 근거, 성과 목표 달성을 위한 전략 제시
- 총 연구 기간의 연차별 마일스톤(정량 지표)을 제시
  • 마일스톤은 연구개발 단계별로 달성해야만 하는 주요한 기술적인 실적으로 평가를 통해 실적 달성 여부를 판단 시 주요 기준으로 활용
- 주관연구개발기관이 공동연구개발기관 등을 자율로 구성하는 연구개발과제 형식으로 제안하여야 함
- 유사과제 수행 또는 참여하고 있는 경우는 중복지원을 지양함

○ **연구개발계획서 작성 시 참고 사항**

- 연구 종료 예정 기관*의 연구 성과를 활용·고도화하는 방안을 제안할 수 있음
  ※ 연구 종료 예정 기관(10개)은 세부3 과제 연구 내용(p.8 이후) 참고

○ **과제 수행 특기 사항**

- 수행 기관이 개발한 모델이 타 기관 모델과 앙상블되거나, 태스크별 성능 비교 대상이 될 수 있음
- 솔루션 개발은 기본적으로 각 기관의 자체 장비를 이용하여 개발하며, 연합학습 테스트는 안전한 데이터 보호를 위해 2세부의 기밀 클라우드 또는 온프레미스(자체 보유 장비)에서 진행
- 동 사업은 기술 변화가 빠른 분야로서 선정된 기관은 선정 이후 사업단과 연구개발 목표 및 범위를 지속적으로 협의·보완하여 고도화하여야 함
- 플랫폼의 지속적인 운영 및 활용성 제고를 위하여 선정기관은 본 과제를 통해 개발한 모델을 FDD 플랫폼에 탑재할 수 있도록 패키지화하여 제출하고, 소스 코드도 제출해야 함
  ※ 연구개발 성과의 소유는 국가연구개발혁신법 등 관련 법령을 준수함
- 수행 기간 중 개발된 모델은 사업단 및 참여기관에게 무상으로 활용하여 서비스를 제공할 수 있음
- 연구 종료 이후 지식재산권 귀속 및 수익 모델에 관한 사항은 사업 종료 전 사업화 계획 수립 과정에서 관련 기관 간 협의를 거쳐 정하여야 함

○ **일반적인 사항은「보건의료기술 연구개발사업 가이드라인」참고**

▶ 선정평가 기준

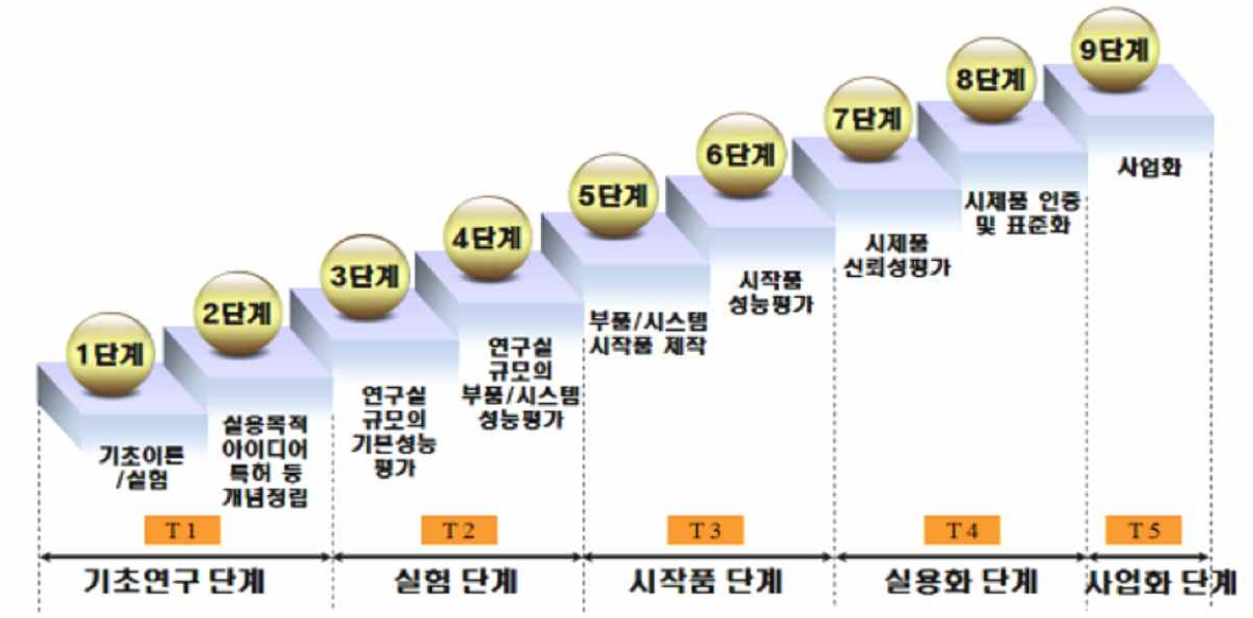
적용가점	해당사항 없음

구분	평가항목(배점)	
	대항목	소항목
발표 평가	1. 연구계획의 적절성 (50)	○ **사업목적에 대한 이해도(10)** - 제안요청서(RFP)의 목표와 지원 내용과의 부합성 - 사업목적 및 연구 방향성에 대한 이해도 ○ **연구 목표의 구체성 및 실현 가능성(10)** - 제시된 연구 목표의 타당성, 연구 결과의 현장 적용 가능성 ○ **연구 내용 우수성(30)** - 연구개발 수행 계획의 구체성(10) - 추가 제안된 선택 연구 내용의 창의성과 독창성(10) - 기존 수행 과제와의 차별성 또는 타연구와의 차별성(10)
	2. 연구개발 역량 (30)	○ **연구책임자 역량(15)** - 수행에 필요한 전문성·경력 보유 여부 ○ **연구개발기관 역량(15)** - 수행 기관의 인프라 및 자원 등 연구 수행 능력
	3. 연구개발 성과 (20)	○ **성공 가능성(10)** - 연구개발을 통한 기술·경제적·사회적 가치 창출 가능성 - 신약개발 현장 기관의 검증 및 의견 반영 ○ **파급효과(10)** - 신약개발 가속화 기여도 - 연구개발 성과 활용·확산 계획

※ 선정평가 계획 수립 시 일부 평가항목(배점) 및 내용이 달라질 수 있음

▶ 참고자료

○ 기술성숙도(Technology Readiness Level, TRL)



○ K-MELLODDY 주요 태스크(62개)

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

2. 참고문헌

□ ADMET/PK 분야 이해

❍ ADME 정의

1. Hodgson, J. (2001). ADMET–turning chemicals into drugs. Nature biotechnology, 19(8), 722-726.
2. [Chem Help ASAP]. (2020.08.24.). ADME & pharmacokinetics - playlist[Video]. Youtube. https://www.youtube.com/watch?v=K0jdqcui9G4&list=PLIzSRqjN72jdKApPyqqgWG7Q5gqvM2QI8

❍ PK 파라미터의 이해

3. 임동석,한승훈,한성필. (2021.09.23). 약동학 기초 이론(Ver1.0.3). PIPET(Pharmacometrics Institute for Practical Education and Training). https://pipetcpt.github.io/pharmapk/pharmapk.pdf
4. Ruiz-Garcia, A., Bermejo, M., Moss, A., & Casabo, V. G. (2008). Pharmacokinetics in drug discovery. Journal of pharmaceutical sciences, 97(2), 654-690.
5. Currie, G. M. (2018). Pharmacology, part 2: introduction to pharmacokinetics. Journal of nuclear medicine technology, 46(3), 221-230. https://tech.snmjournals.org/content/46/3/221#sec-5

❍ In-vitro에서 제공하는 ADMET assay

6. [NCATS NIH]. (2021.03.30.). In Vitro Assessment of ADME Properties of Lead Compounds[Video]. Youtube. https://www.youtube.com/watch?v=qUEkVJqeg0I

❍ In-vivo에서 제공하는 ADMET assay

7. Blass, B. E. (2015). Basic principles of drug discovery and development. Elsevier. 245-306. https://doi.org/10.1016/B978-0-12-411508-8.00006-2
8. Chung, T. D., Terry, D. B., & Smith, L. H. (2015). In vitro and in vivo assessment of ADME and PK properties during lead selection and lead optimization-guidelines, benchmarks and rules of thumb. https://www.ncbi.nlm.nih.gov/books/NBK326710/

❍ In-human에서 제공하는 ADMET assay

9. Lucas, A. J., Sproston, J. L., Barton, P., & Riley, R. J. (2019). Estimating human ADME properties, pharmacokinetic parameters and likely clinical dose in drug discovery. Expert opinion on drug discovery, 14(12), 1313-1327. https://doi.org/10.1080/17460441.2019.1660642
10. Kuerzel, U., Krone, V., Zimmer, M., Shackleton, G. (2011). The Human ADME Study. In: Vogel, H.G., Maas, J., Gebauer, A. (eds) Drug Discovery and Evaluation: Methods in Clinical Pharmacology. Springer, Berlin, Heidelberg. Chapter B.10. https://doi.org/10.1007/978-3-540-89891-7_11

❍ ADMET 데이터 포맷

11. 국가바이오스테이션, 국가생명연구자원정보센터, [자료] 화합물데이터 가이드북 안내, KBDS_화합물데이터가이드북_v2.0_2024.pdf, 2024.10.17., https://kbds.re.kr/portal/board/ed607150170611ee8157141877507e8b/view/635

□ 연합학습 기술 이해

❍ 연합학습의 이해

12. Ludwig, H., & Baracaldo, N. (Eds.). (2022). Federated learning: A comprehensive overview of methods and applications (pp. 13-19). Cham: Springer
13. Martin Keen. [IBM Technology]. (2023.07.07.). Training AI Models with Federated Learning[Video]. Youtube. https://www.youtube.com/watch?v=zqv1eELa7fs
14. Luzón, M. V., Rodríguez-Barroso, N., Argente-Garrido, A., Jiménez-López, D., Moyano, J. M., Del

Ser, J., ... & Herrera, F. (2024). A tutorial on federated learning from theory to practice: Foundations, software frameworks, exemplary use cases, and selected trends. IEEE/CAA Journal of Automatica Sinica, 11(4), 824-850, https://doi.org/10.1109/JAS.2024.124215

15. Qi, P., Chiaro, D., Guzzo, A., Ianni, M., Fortino, G., & Piccialli, F. (2024). Model aggregation techniques in federated learning: A comprehensive survey. Future Generation Computer Systems, 150, 272-293, https://doi.org/10.1016/j.future.2023.09.008, https://doi.org/10.1016/j.future.2023.09.008

❍ 연합학습 적용 사례

16. (유전체, EMR 데이터) Oasys Now. CoMPai Platform. https://www.oasysnow.com/products/compai
17. (구조 예측) Apheris. AI Structural Biology (AISB) Network - OpenFold3. https://www.apheris.com/join-a-network/aisb
18. (항체 연구) Apheris. Antibody Developability Consortium. https://www.apheris.com/join-a-network/antibody-developability-consortium
19. (비임상 시험) Eli Lilly and Company. Lilly TuneLab. https://tunelab.lilly.com/
20. (비임상 시험) Apheris. ADMET Network. https://www.apheris.com/join-a-network/admet
21. (생물학적 제제 특성) The Federated AI for Therapeutic Engineering (FAITE) Consortium. FAITE Consortium. https://faiteconsortium.org/
22. (외부 임상 검증) Ministry of Health and Welfare, Taiwan AI Center. (n.d.). National Strategy for External Clinical AI Validation. https://aicenter.mohw.gov.tw/AC/cp-7294-82667-208.html
23. (AI 암 연구) Cancer AI Alliance (CAIA). Cancer AI Alliance. https://www.canceralliance.ai/

❍ 연합학습 프레임워크

24. [NVIDIA Flare] Roth, H. R., Cheng, Y., Wen, Y., Yang, I., Xu, Z., Hsieh, Y. T., ... & Feng, A. (2022). Nvidia flare: Federated learning from simulation to real-world. arXiv preprint arXiv:2210.13291.

❍ 연합학습 기술 연구 동향

1. Frederik Wenkel, Hessam Mehrali, Patrick Kink, Eva Wolfangel, Dassyn Barr, Günter Klambauer. (2025). Insights into the Unknown: Federated Data Diversity Analysis on Molecular Data. arXiv preprint, arXiv:2510.19535. https://arxiv.org/abs/2510.19535
2. Yuhang Yao, Jianyi Zhang, Zhaoyi Liang, Ruofan Liu, et al. (2024). Federated Large Language Models: Current Progress and Future Directions. arXiv preprint, arXiv:2409.15723. https://arxiv.org/abs/2409.15723
25. Yang, Z., Zhang, Y., Zheng, Y., Tian, X., Peng, H., Liu, T., & Han, B. (2024). FedFed: Feature distillation against data heterogeneity in federated learning. Advances in Neural Information Processing Systems, 36.
26. Lee, R., Kim, M., Li, D., Qiu, X., Hospedales, T., Huszár, F., & Lane, N. (2024). FedI2p: Federated learning to personalize. Advances in Neural Information Processing Systems, 36.
27. Qiu, P., Zhang, X., Ji, S., Fu, C., Yang, X., & Wang, T. (2024). Hashvfl: Defending against data reconstruction attacks in vertical federated learning. IEEE Transactions on Information Forensics and Security, https://doi.org/10.1109/TIFS.2024.3356164
28. Sharma, A., & Marchang, N. (2024). A review on client-server attacks and defenses in federated learning. Computers & Security, 103801, https://doi.org/10.1016/j.cose.2024.103801
29. Peng, L., Luo, G., Zhou, S., Chen, J., Xu, Z., Sun, J., & Zhang, R. (2024). An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. NPJ Digital Medicine, 7(1), 127, https://doi.org/10.1038/s41746-024-01126-4

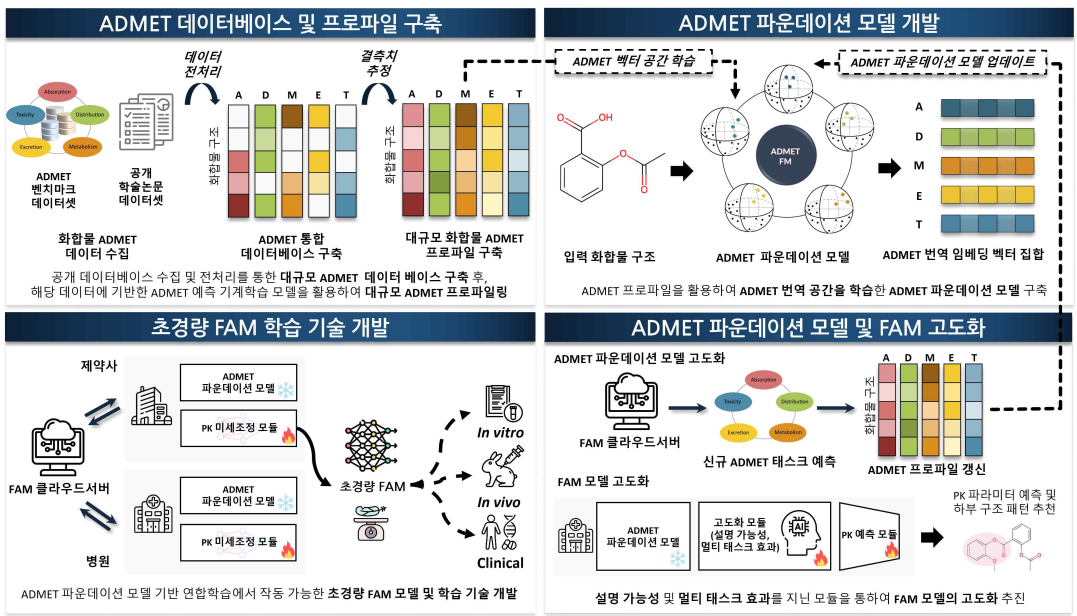
□ 세부3 과제 연구 내용

RS-2024-00460010 (세부3) FAM(Federated AMDET Model) 개발

메가스케일 ADMET-파운데이션 모델 기반 초경량 FAM 통합 솔루션 개발

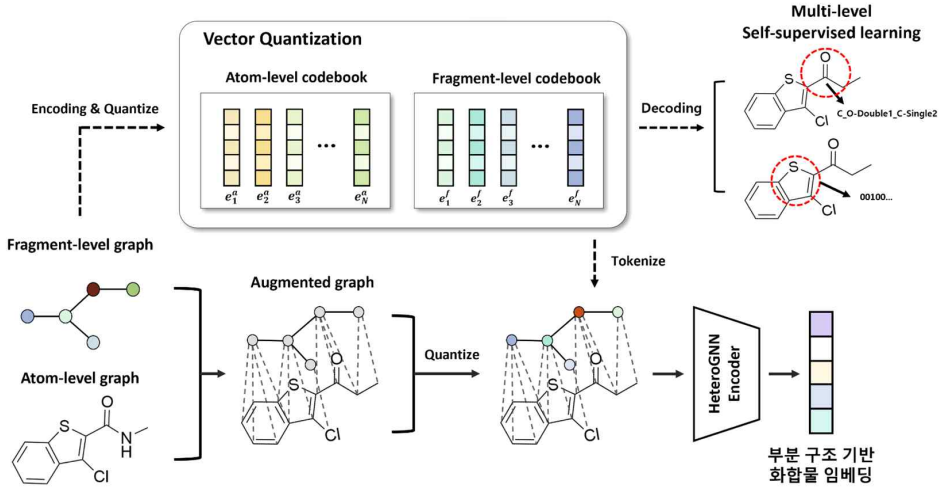
주관: 연세대학교
공동: 주식회사 히츠

요약문	□ 본 연구는 광범위한 ADMET 태스크에 적용가능한 파운데이션 모델 기반 초경량 FAM 모델 개발을 목표로 함 □ 이를 위해 화합물의 구조 정보 및 ADMET 프로파일 정보를 학습한 대규모 파운데이션 모델과 최소한의 파라미터로 구성된 초경량 FAM 학습 기술을 개발함 □ 또한, ADMET 임베딩 조합 모듈과 설명 가능성 모듈을 도입해 설명 가능성과 멀티태스크 효과를 강화함으로써 FAM 모델을 고도화함
특징 및 차별성	□ 데이터 전처리기 배포를 통한 데이터 보안 구현 • 데이터 구조 정보 노출 위험 없음 □ 메가스케일 ADMET-파운데이션 모델 기반 예측의 정확도 향상 • >10M 화합물 구조 정보, >50 태스크, ~60K 화합물, ~200K 활성 정보 • 임의의 ADMET 태스크에 사용 가능한 범용적 파운데이션 모델 • ADMET embedding을 기반한 multi-task learning 효과 구현 □ 초경량 FAM 기반 학습 효율성 제고 • ~10K개의 학습 파라미터를 요구하는 초경량 모델 • 대규모 태스크에 적용 가능 (태스크 당 A100 기준 1분 이내, RTX 4090 기준 2분 이내)
핵심 연구내용	□ 연구 개요도



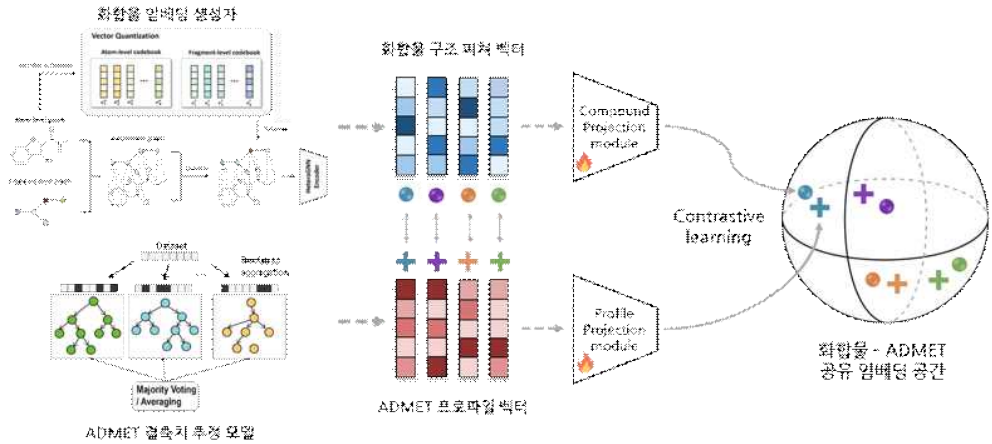
[그림 1] 제안하는 연구개발 내용의 개요도

	□ **1차년도** • ADMET 데이터베이스 구축 - 본 연구에서는 다양한 ADMET 태스크에 적용가능한 파운데이션 모델 개발을 위해 신약 개발의 각 단계(in-vitro, in-vivo, in-human) 별 ADMET 태스크를 포함한 데이터베이스와 프로파일을 구축함 - 이를 위해 ChEMBL 등의 대규모 공개 DB와 분산된 실험 논문 데이터를 수집 및 정제하여 다양한 태스크를 포함한 통합 ADMET 데이터베이스를 구축함 • ADMET 프로파일 구축 - 선행 ADMET 특성 예측 플랫폼 및 구축한 ADMET 데이터베이스를 기반으로 개발한 기계학습 예측 모델을 바탕으로 최적의 결측치 예측 모델을 선정하여 밀집된 형태의 데이터 프로파일을 구축함 - 최종적으로 50개 이상의 ADMET 태스크를 포함한 프로파일을 구축하고 이를 ADMET 파운데이션 모델 개발의 핵심 데이터로 활용함 • ADMET 파운데이션 모델 설계 - 본 연구에서는 ADMET 프로파일 정보를 활용하여 구조 기반 특징 벡터의 한계를 극복하고, 다양한 ADMET 태스크에 일반화할 수 있는 ADMET 파운데이션 모델을 제안함 - 해당 모델은 화합물 부분 구조 기반 인코더와 ADMET 번역 임베딩 생성자로 구성됨 - 1차년도에는 대량의 화합물 데이터베이스에서 부분 구조를 추출하고, 각 모듈의 최적 모델 구조 및 학습 방식을 선정함 • 데이터 전처리기 기획 - ADMET 데이터의 일관된 처리를 위한 전처리기 설계를 목표로, 데이터 기밀성을 유지하면서도 표준화 및 통합 가능 방안을 협의함 - 시험별 차이를 반영한 유연한 전처리 옵션을 제공하고, 정족수 기준 설정 및 중심구조 기반 데이터 분리를 통해 신뢰도 높은 학습 데이터 세트를 구축함 - 또한, HTS 기반 보조 데이터를 활용하여 FAM 모델의 성능을 향상시키며, 최종적으로 전처리기 기획 문서 및 프로토타입을 개발함
핵심 연구내용	□ **2차년도** • ADMET 파운데이션 모델 개발 - 2차년도에는 1차년도의 설계를 바탕으로 ADMET 파운데이션 모델 개발을 진행함 - 화합물 부분 구조 기반 인코더는 임의의 화합물에 대해 다양한 수준의 구조 정보를 함축한 다차원 특징 벡터로 인코딩하는 역할임 - 대량의 화합물 구조 데이터를 기반으로 원자와 부분 구조 정보를 담은 벡터 집합을 학습하고 이를 통합하여 분자 구조 기반 임베딩을 생성함 - 학습 과정에서는 이산 그래프 신경망을 통한 자기 지도 학습 방식을 활용함



[그림 2] 화합물 부분구조 인코더 모델 구조

- ADMET 번역 임베딩 생성자는 부분 구조 기반 임베딩에 ADMET 프로파일 정보를 추가 반영하여, 구조가 유사해도 서로 다른 ADMET 특성을 갖는 화합물들을 구분할 수 있도록 함
- 대조 학습 방식을 활용하여 ADMET 공간에 화합물을 투영하여 구조 정보와 ADMET 정보를 모두 포함한 최종 임베딩을 생성하여, 다양한 다운스트림 태스크에 대한 호환성과 일반화 성능을 제공함
- 이 두 모듈은 연합학습 시 직접 학습되지 않고 특징 벡터 생성에만 활용되어 모델의 경량성을 높일 수 있음



[그림 3] ADMET 번역 임베딩 생성자 구조

• FAM 모델 개발

- FAM 모델은 연합학습 환경에 최적화된 경량화 모델로, ADMET 파운데이션 모델과 PK 미세조정 모듈로 구성됨. 연합학습 과정에서 중앙 서버가 사전 학습된 ADMET 파운데이션 모델과 초기화된 PK 미세조정 모듈을 클라이언트에 배포하고, 각 클라이언트는 로컬 환경에서 PK 미세조정 모듈만 학습한 후 중앙 서버로 업데이트함
- PK 미세조정 모듈은 최소한의 파라미터를 가진 DNN으로 구성되어 적은 컴퓨팅 자원만으로도 안정적인 성능을 확보하며, 개발 단계에서는 이러한 경량성과 성능을 극대화할 수 있는 효율적인 PK 미세조정 모듈을 개발하는 것이

	목적임 • 전처리기 모델 개발 - 전처리기 모델 개발은 공개 데이터와 기관별 기밀 데이터를 일관되게 처리하고, FAM 모델에 적합한 데이터 표현을 제공하는 것을 목표로 함 - 이를 위해, 전처리기 내부 자동화 모듈과 사용자 개입을 바탕으로 태스크 유형, 이상치 및 중복 데이터 검사 및 제거, 데이터 표준화 등이 가능한 전처리기를 개발하고 이에 대한 테스트를 진행함
핵심 연구내용	□ 3차년도 • ADMET 파운데이션 모델 및 FAM 모델의 고도화 - FAM 고도화 단계에서는 앞서 개발한 PK 미세조정 모듈을 두 가지 방안을 기반으로 고도화함 - 첫째, ADMET 임베딩 조합 모듈을 통해 ADMET 파운데이션 모델에서 생성된 A, D, M, E, T 임베딩을 조합한 특징 벡터를 생성함 - 이는 다양한 특성을 포함하는 멀티 태스크 효과를 통한 성능 향상을 기대할 수 있음 - 둘째, 예측에 중요한 하부 구조 패턴을 추출하는 분자 구조 기반 설명가능성 모듈을 도입하여 모델의 해석성과 화합물 구조 정보의 표현력을 강화하여 예측 성능을 향상함. 두 고도화 방안 모두 최소한의 파라미터로 구성하여 FAM 모델의 초 경량성을 유지할 계획임 • 전처리기 최종화 - 전처리기 최종화 단계에서는 최종적인 구축과 성능 평가를 중점으로 진행함 - 최종 구축 단계에서는 사업 과정에서 받은 데이터를 바탕으로 전처리기-FAM 워크플로우의 작동을 검증하고 최적화하여 플랫폼 탑재를 완료함 - 전처리기 성능 평가는 ADMET 데이터의 레이블 균형, 표본추출, 결측치 처리, 데이터 증강의 측면에서 연합학습의 효과를 측정하는 방식으로 평가함 - 또한, 전처리기의 신뢰성과 재현성을 보장하기 위한 자동화된 테스트 및 검증 프레임워크를 구축하여, 단위 변환·pH 보정·엔드포인트 매핑 등 핵심 모듈에 대한 단위 테스트와 통합 테스트를 수행하고, CI/CD 파이프라인을 통한 지속적 품질 관리 체계를 마련함

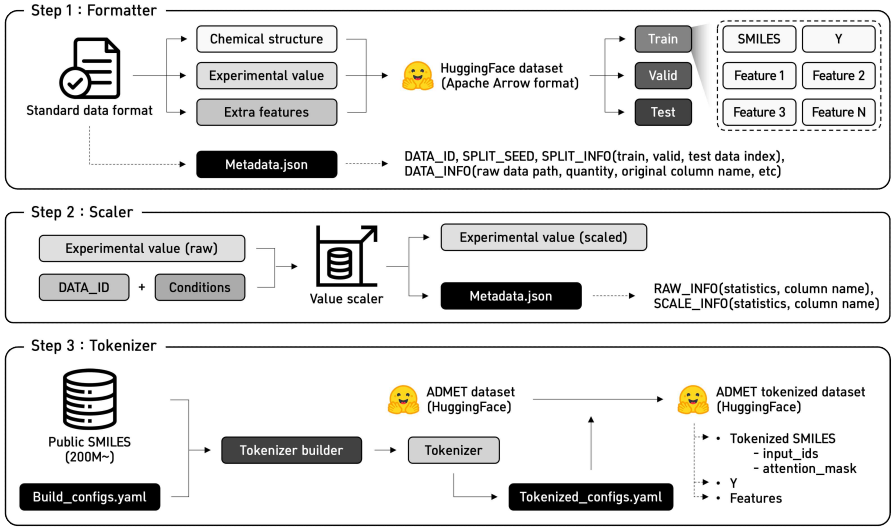
연합학습 기반 ADMET 예측을 위한 언어-그래프 앙상블 모델 개발

주관: 목암생명과학연구소
공동: 서울대학교

요약문	□ 다각화되는 ADMET 예측 태스크에서 유연하게 활용할 수 있는 파운데이션 앙상블 모델 개발을 위해 공개 화합물 데이터베이스를 수집하고 전처리하여 거대 언어 모델 및 그래프 모델의 학습을 위한 기반을 마련하고 ADMET 예측을 위한 미세조정을 거쳐 연합학습 플랫폼에 탑재할 수 있는 확장된 FAM(Federated ADMET Model)모델 개발을 목표로 함
특징 및 차별성	□ 사전 학습 모델은 특정 태스크나 도메인에 맞추어 학습하는 과정인 미세조정을 통해 모델을 최적화하는 방법을 시도할 수 있음. 미세조정은 사전 학습된 모델을 활용하기 때문에 특정 태스크에 대한 소량의 데이터만으로도 모델을 효과적으로 학습하는 특징을 지니고 있음 □ 거대 언어 모델에 SMILES 데이터를 입력하면 구조 정보를 기반으로 화합물의 특성을 학습할 수 있으며, 이를 활용하여 신약개발 관련 다양한 태스크에 맞게 모델을 미세 조정할 수 있음 □ 분자의 물성은 원자 간의 상호작용에 의존하며, 원자 간 상호작용은 그래프의 구조에 직접적으로 연관되어 있음. 따라서, 원자 간 상호작용을 효과적으로 이해하기 위해서는 분자 그래프에 담긴 구조 정보를 활용하는 것이 매우 중요함 □ 대규모 공공 화합물 데이터를 기반으로 사전 학습(pre-trained)된 파운데이션 모델을 활용한다면, 다양한 ADMET 예측 태스크에 대해 미세조정을 수행할 수 있으므로, 각 태스크에 대해 비교적 적은 데이터로 높은 성능을 도출할 수 있음 □ 연합학습을 적용하여 여러 기관에서 생성한 ADMET 데이터를 보안 유지 상태로 활용할 수 있으므로 모델의 일반화 능력을 향상할 수 있으며, 새로운 데이터가 생성될 때마다 모델을 실시간으로 업데이트할 수 있음. 이를 통해 ADMET 예측 모델의 성능 향상을 극대화할 수 있음
핵심 연구내용	□ 데이터 수집 • 알려진 대규모 공개 화합물 데이터베이스를 통해 구조적 다양성을 가진 화합물 데이터 수집 • 학습 최적화를 위해 SMILES 길이와 약물 디자인에 사용되는 원소를 기준으로 데이터 필터링 • 선별된 데이터는 토크나이저 학습에 사용되었으며, 추후 거대 언어 모델의 사전 학습(Pre-training)용으로도 활용 예정 • 대규모 공개 화합물 데이터베이스 중 충분히 많은 양을 확보한 데이터베이스에서 무작위로 추출하여 설명 데이터 생성 • 생성된 설명 데이터는 추후 그래프 파운데이션 모델 학습에 활용할 예정

□ 데이터 전처리

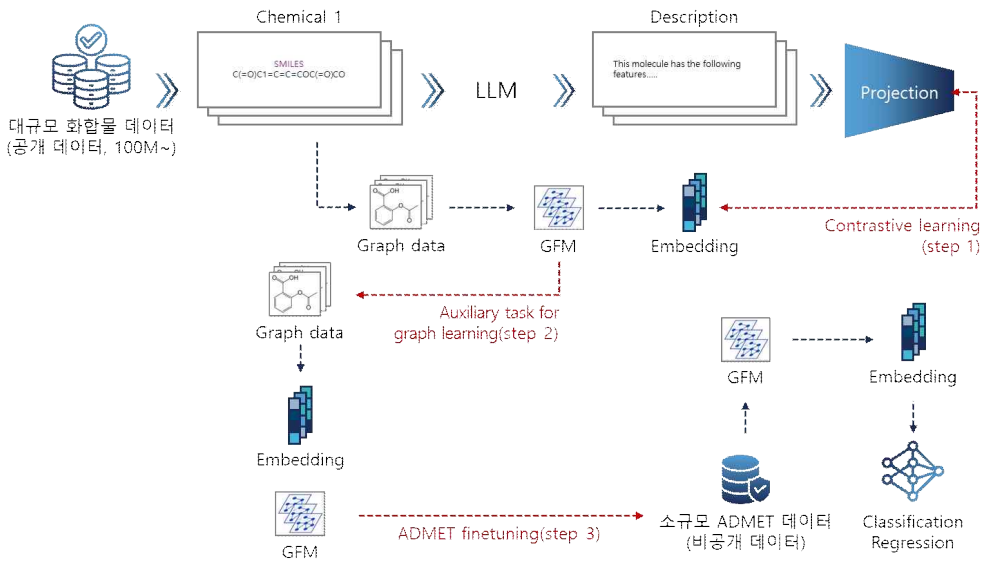
- 운영협의체, 실무위원회 및 연합학습 워크숍 참여를 통해 활용가능한 ADMET 데이터 파악 및 데이터 제공 형식(표준 데이터 포맷)을 기준으로 원본 데이터에서 필요한 데이터를 추출하고 가공하는 3단계 전처리 도구 설계



[그림 1] 거대 언어 모델 구조도

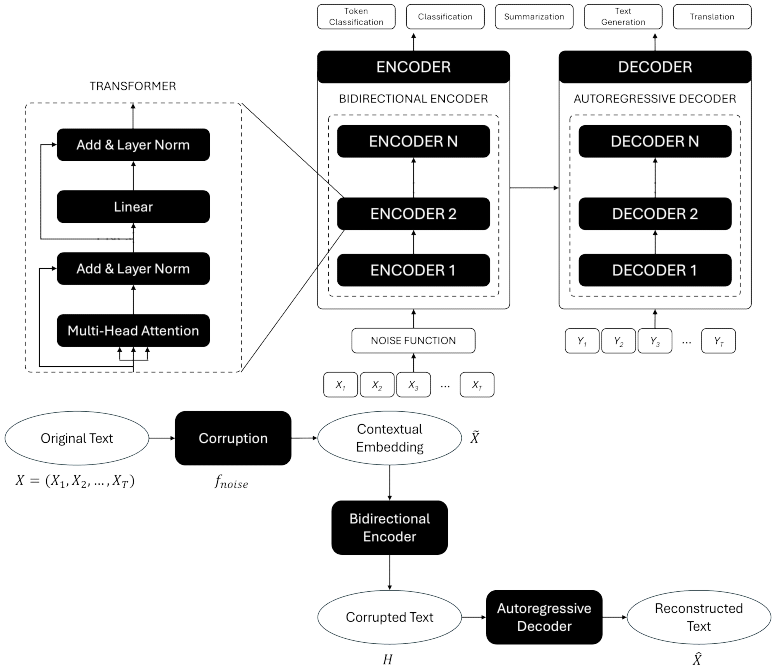
□ 모델 개발

- SMILES 기반의 GNN Encoder와 Description 기반의 Text Encoder를 활용한 대조 학습(Contrastive Learning) 기반 멀티모달 학습 방법을 연구. 그래프와 텍스트 임베딩 간의 의미적 정합성을 높이는 방향으로 모델 설계 및 실험 진행
- 멀티모달 학습을 통해 높은 표현력을 가지는 그래프 파운데이션 모델을 개발. 그래프와 텍스트 간의 상호 보완적 정보를 활용하여 다양한 화합물 표현 학습 가능성을 탐색. 이후 ADMET 등 다양한 하위 태스크에 대한 파인 튜닝 방법을 연구. 멀티태스킹 학습을 통해 그래프 모델의 일반화 성능 향상을 목표로 최적의 학습 전략을 설계함



[그림 2] 그래프 기반 모델 구조도

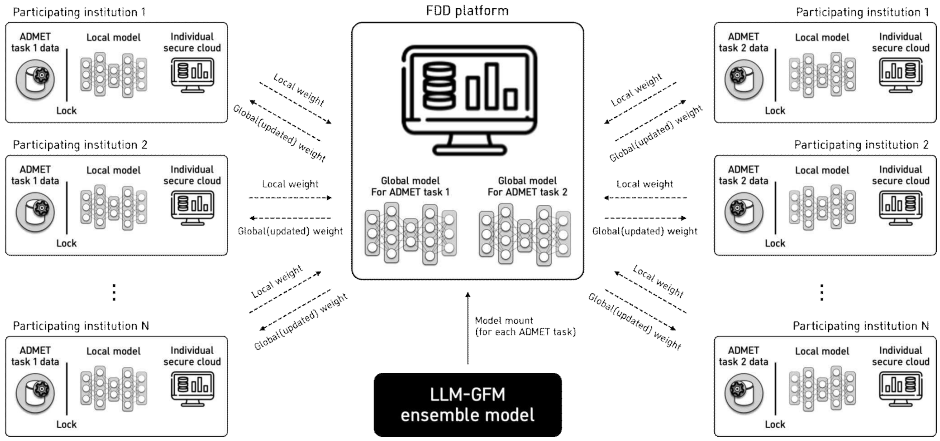
- BART를 바탕으로 언어 모델을 설계함. BART는 시퀀스-시퀀스 모델로 구축된 노이즈 제거 자동 인코더(de-noising auto-encoder)로, 임의의 노이즈 함수를 사용하여 텍스트를 손상시키고, 모델이 원래의 텍스트를 재구성하는 방식으로 학습을 수행함



[그림 3] BART 기반 모델 구조도

□ **연합학습 테스트**

- 운영협의체, 실무협의회, 연합학습 워크숍을 통해 협의된 ADMET 태스크별로 연합학습 클라이언트(client)를 정의하고, ADMET 태스크별로 클라이언트를 그룹화함
- FDD 플랫폼을 서버(server)로 정의하여 각 클라이언트의 모델 업데이트를 수집하고 글로벌 모델을 업데이트하는 역할을 부여함
- FDD 플랫폼에 탑재된 모델은 FAM(Federated ADMET Model)으로 작동하며, 연합학습의 글로벌 모델 역할을 수행함. 연합학습 최초 동작 과정에서 ADMET 태스크별로 각 참여기관의 보안 클라우드에 전송됨
- LLM-GFM 앙상블 기반 FAM은 규모가 큰 모델임. 따라서, PEFT(Parameter Efficient Fine-Tuning)를 통해 모델의 전체 파라미터 중 일부만을 선택적으로 조정함으로써, 모델의 핵심 구조를 유지함과 동시에 특정 ADMET 태스크에 맞게 모델을 조정할 수 있음



[그림 4] 연합학습 테스트 시나리오

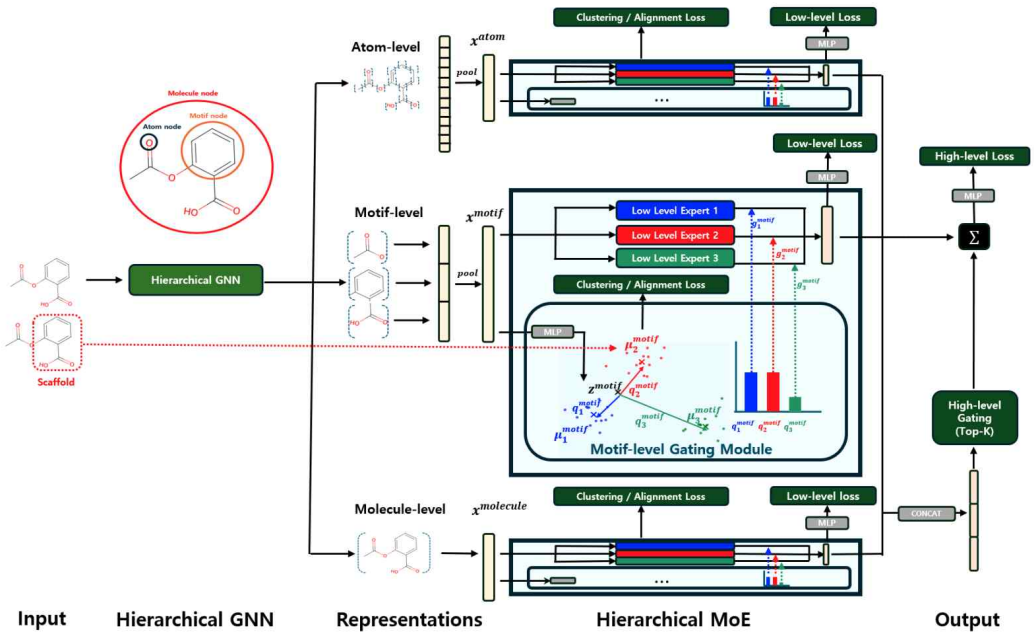
대규모 약물 표현 학습과 LLM 기반 문헌 마이닝을 활용한 연합학습 기반 ADMET 예측 모델 개발

주관: (주)아이젠사이언스

요약문	□ **본 연구는 연합학습(Federated Learning) 기반의 ADMET 예측 모델을 개발하여 신약개발의 실패 원인 중 하나인 부적합한 ADMET 특성을 사전에 정밀하게 예측함으로써 임상시험 성공률을 높이고, 신약 개발 비용과 기간을 단축하는 것을 목표로 함** □ **이를 위해 대규모 화합물 데이터를 활용하고, 최신 Foundation 모델(오픈소스 모델 및 자체 개발 모델)과 신약개발 특화 언어모델을 융합하여 약물 표현을 극대화하며, 문헌 마이닝을 통한 ADMET 정보 추출 및 해석 가능성을 확보함**
특징 및 차별성	□ **멀티모달 약물 표현 모델** • 분자 구조를 SMILES, 2D/3D 그래프 등 다양한 모달리티로 표현하고, 코어 스캐폴드와 R-group 정보를 동시에 반영하는 계층적 임베딩 기법을 개발하여, 기존 단편적 표현 한계를 극복함 □ **신약개발 특화 LLM 기반 문헌 마이닝** • 자체 개발한 신약개발 특화 LLM(Meerkat)을 활용해 특허, 학술 논문, 임상 보고서 등 다양한 비정형 문헌에서 ADMET 정보를 자동으로 추출하고, 정확한 데이터 구조화를 통해 모델 학습 데이터 세트의 신뢰도를 크게 향상시킴 □ **연합학습 환경 최적화 모델** • 다기관의 이질적 데이터를 효과적으로 반영할 수 있도록, 각 기관의 로컬 모델 업데이트 정보를 정교하게 집계하고 노이즈 보정 및 기여도 기반 가중치 조정 기법을 적용하여, 연합학습 환경에서 최고 수준의 예측 성능과 해석력을 가진 ADMET 예측 모델을 구현함
핵심 연구내용	□ **약물 표현 Foundation 모델 구축 및 데이터 통합** • 문제 제기 - (분자 표현의 한계) 기존 모델은 SMILES, 2D 그래프 등 단편적 정보에 의존하여, 분자의 코어 구조(Scaffold)와 R-group 등 세부 특성을 충분히 반영하지 못함. 분자 동역학, 3D 구조 및 복합적 상호작용 정보를 효과적으로 융합하지 못해 신약 후보 물질의 특성 분석에 한계 존재 - (데이터 사일로 및 불일치 문제) PubChem, ChEMBL, ZINC15 등 공개 데이터베이스 간 중복 및 포맷 불일치로 통합 분석에 어려움. 제약·바이오 기업 및 학계에서 축적한 내부 ADMET 데이터와 화합물 데이터가 별도로 관리되어 데이터 활용도가 낮음 □ **연구 내용**

• 대규모 데이터셋 구축 및 통합 전처리 플랫폼 개발

- PubChem, ZINC15, ChEMBL 등 공개 데이터베이스에서 1억 개 이상의 화합물 데이터 수집
- 신약 개발 시 약물 상호작용 및 대사 안정성 예측에 필수적인 CYP(Cytochrome P450) 저해능(Inhibition) 데이터를 자체 실험을 통해 구축함. 2025년 12월 기준, 총 74종의 약물 각각에 대하여 주요 대사 효소 5종(1A2, 2C9, 2C19, 2D6, 3A4) 전체에 대한 실험을 개별적으로 모두 수행함으로써, 빈틈없는(Full-panel) 대사 저해 프로파일 데이터를 확보하였음. 지속적인 추가 자체 생산 데이터 확보 예정
- RDKit 등의 도구를 활용하여 SMILES 정규화, Tanimoto 유사도 기준 중복 제거 및 기본 분자 특성(분자량, LogP, TPSA 등) 계산
- 내부 파이프라인에서 자체 생산한 ADMET 및 효능 데이터와 외부 데이터를 통합하는 자동화 전처리 시스템 구축
- 데이터 스키마 표준화, 온톨로지 기반 보정 및 배치 효과(Batch Effect) 교정 기법을 적용하여 데이터 품질 및 일관성 강화



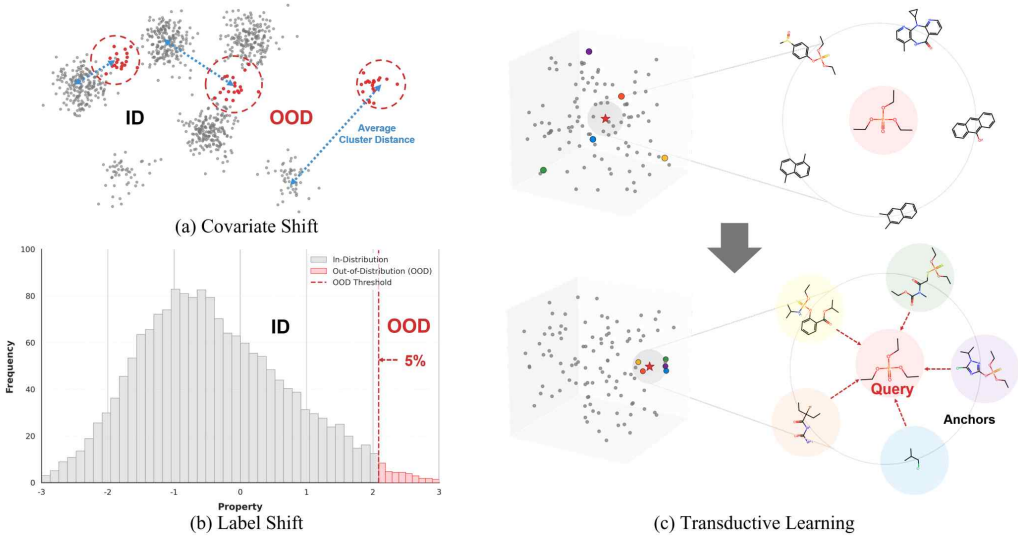
[그림1] Hierarchical MoE 모델 모식도

• 계층적 MoE 기반의 멀티스케일 분자 표현 기술 구축

- 기존의 단순한 벡터 결합(Concatenation) 방식을 넘어, 분자 구조의 계층적 정보를 동적으로 통합하는 Hierarchical MoE(Mixture of Experts) 아키텍처로 모델을 고도화함
- 계층적 특징 추출 및 동적 벡터 조합: 분자를 구성하는 Atom(원자), Motif(부분 구조), Molecule(전체 분자)의 3단계 레벨에서 각각 독립적인 전문가(Expert) 네트워크를 구축하여 다면적인 구조 정보를 추출함
- Gating Network 기반의 최적화: 각 입력 분자의 특성에 맞춰 어떤 레벨의 특징을 얼마나 반영할지 결정하는 Gating Module을 도입함. 이를 통해

Cross-attention과 유사하게 각 계층 간의 중요도를 동적으로 산출하고 최적의 비율로 벡터를 조합하여 표현력(Representation)을 극대화함

- Local-Global 정보 균형 확보: GNN의 한계인 과도한 평활화(Over-smoothing)를 방지하고, 용해도(Global feature)와 대사 안정성(Local feature) 등 서로 다른 스케일의 특성을 모두 포괄할 수 있는 최적의 임베딩 조합 기술을 구현함



[그림2] Transductive Learning 기반 OOD 모델 모식도

- Transductive learning 기반 고강건성 예측 모델 구축
  - 다양한 ADMET 속성을 동시에 예측하는 Multi-task 환경에서, 데이터 희소성 및 분포 차이로 인한 성능 저하를 극복하기 위해 Transduction 기법과 Task-Adaptive Routing을 적용함
  - Transduction 기반 OOD(Out-of-Distribution) 예측 강건성 확보: Multi-task 학습 시 학습 데이터의 화학적 공간(Chemical Space)을 벗어나는 새로운 구조(OOD)가 입력될 경우 예측 성능이 급락하는 문제를 해결하기 위해 트랜스덕티브 러닝(Transductive Learning)을 적용함. 추론 단계에서 테스트 샘플의 분포 정보를 능동적으로 참조하여 Covariate Shift를 보정함으로써, 미지의 약물 구조에 대해서도 안정적인 예측 성능을 유지함
  - Task 별 전문가(Expert) 라우팅 최적화: 모든 태스크가 동일한 파라미터를 공유하는 기존 방식 대신, 예측하려는 ADMET 속성(예: 독성 vs 흡수)에 따라 계층적 전문가(Hierarchical Experts) 중 가장 적합한 경로를 선택적으로 활성화하는 라우팅 메커니즘을 적용하여 Multi-task 효율성을 10% 이상 개선함

핵심 연구내용

□ **LLM 기반 문헌 마이닝 및 정보 추출 시스템 개발**

- 문제 제기
  - (비정형 문헌 데이터의 산재 및 활용 한계) ADMET 관련 정보가 학술 논문, 특허, 임상 보고서 등 다양한 비정형 문헌에 분산되어 있어 수작업 기반의 정보 추출은 비효율적. 기존 LLM은 의생명 분야 특화 데이터가 부족해, ADMET 관련 전문 용어와 문맥을 정확히 파악하지 못함

- (추출 데이터의 품질 및 구조화 문제) 문헌으로부터 추출한 데이터의 신뢰도와 구조화가 미흡하여, 후속 학습 데이터 세트에 바로 활용하기 어려움. 추출된 ADMET 정보의 정량적 신뢰도가 낮아 모델 학습에 영향을 미칠 수 있음

□ **연구 내용**

• [신약개발 특화 LLM 개발 및 최적화]

- 자체 개발한 의생명 특화 특화 언어모델을 최적화하여, 의학·생명과학 분야의 전문 용어와 문헌 구조에 최적화된 정보 추출 시스템 구축
- USMLE와 같은 평가에서 기존 대형 모델(GPT-4)보다 우수한 성능을 보인 기술력을 바탕으로, ADMET 관련 문헌 데이터에 적용

• 자동 문헌 마이닝 및 데이터 구조화 시스템 구축

- 특허, 학술 논문, 임상 보고서 등에서 ADMET 관련 데이터를 자동으로 추출하고 구조화하는 알고리즘 개발
- Few-shot learning, 프롬프트 최적화 및 자동 검증(크로스체크, 신뢰도 점수 산출) 시스템 도입
- 추출된 데이터를 내부 ADMET 데이터 세트와 통합하여, 데이터 부족 현상을 극복하고 모델 학습의 신뢰도 향상

• 데이터 증강 및 합성 전략

- 분자동역학 시뮬레이션 및 자유에너지 계산을 통해, 문헌에서 추출하기 어려운 ADMET 데이터를 합성하여 보완
- LLM 기반 문헌 추출과 시뮬레이션 데이터의 결합으로, 다양한 ADMET 속성을 다각적으로 분석하는 데이터세트 구축

의생명 특화 LLM Meerkat 및 다른 LLM을 이용하여
논문, 특허, 임상 보고서 등 비정형 문헌 데이터로부터 ADMET 관련 데이터 추출

막대한 인력, 자원 투입이 필요한 문헌 기반 **ADMET** 데이터 추출 작업 수행이 가능한 **LLM** 기술역량 보유

ADMET 관련 Description에 대하여 실제 LLM을 이용한 데이터 구조화 결과

"In our study of novel kinase inhibitors, we evaluated the ADMET properties of three lead compounds: **XJ-5, XJ-7, and XJ-9**.
XJ-5 (4-(4-methylpiperazin-1-yl)-N-(4-(trifluoromethyl)phenyl)pyridine-2-amine) demonstrated good oral bioavailability in rats (F = 68%) and moderate plasma protein binding (87%). Its metabolic stability in human liver microsomes was acceptable, with a half-life of 52 minutes. The compound showed no significant inhibition of major CYP450 enzymes at concentrations up to 10 μM. In the Ames test, XJ-5 was found to be non-mutagenic.
XJ-7 (N-(3-((2-((1H-pyrazol-3-yl)amino)pyrimidin-4-yl)amino)phenyl)acrylamide) exhibited poor aqueous solubility (0.02 mg/mL at pH 7.4) but high permeability in the Caco-2 assay (Papp = 25 × 10^-6 cm/s). Its plasma protein binding was high (98%), which may limit its tissue distribution. The compound was rapidly metabolized in human liver microsomes, with a half-life of only 15 minutes. XJ-7 showed moderate inhibition of CYP3A4 (IC50 = 5 μM) but no significant effects on other CYP450 enzymes.
Lastly, **XJ-9** (2-((1H-indazol-5-yl)amino)-6-(2,6-difluorophenyl)pyrimidin-4(3H)-one) demonstrated the most promising ADMET profile. It had good aqueous solubility (0.5 mg/mL at pH 7.4) and moderate Caco-2 permeability (Papp = 8.5 × 10^-6 cm/s). Its oral bioavailability in rats was excellent (F = 85%), with moderate plasma protein binding (75%). XJ-9 showed good metabolic stability with a half-life of 120 minutes in human liver microsomes. No significant inhibition of CYP450 enzymes was observed. In vitro safety studies revealed no hERG inhibition at concentrations up to 30 μM, and the compound was negative in the Ames test."

속성	XJ-5	XJ-7	XJ-9
IUPAC	4-(4-methylpiperazin-1-yl)-N-[4-(trifluoromethyl)phenyl]pyridin-2-amine	N-[3-[[2-(1H-pyrazol-3-ylamino)pyrimidin-4-yl]amino]phenyl]prop-2-enamide	2-(1H-indazol-5-ylamino)-6-(2,6-difluorophenyl)pyrimidin-4(3H)-one
용해도 (mg/mL, pH 7.4)	-	0.02	0.5
Caco-2 투과성 (10^-6 cm/s)	-	25	8.5
경구 생체이용률 (%)	68	-	85
혈장 단백질 결합 (%)	87	98	75
대사 안정성 (반감기, 분)	52	15	120
CYP450 억제	10 μM까지 유의한 억제 없음	CYP3A4 IC50 = 5 μM	유의한 억제 없음
변이원성 (Ames test)	음성	-	음성
hERG 억제	-	-	30 μM까지 억제 없음

[그림3] 대형 언어 모델(LLM)활용 비정형 문헌 데이터 구조화

핵심 연구내용

□ **신약 개발 태스크 특화 추론 기법 및 추론 과정 설명 제시**

• 문제 제기

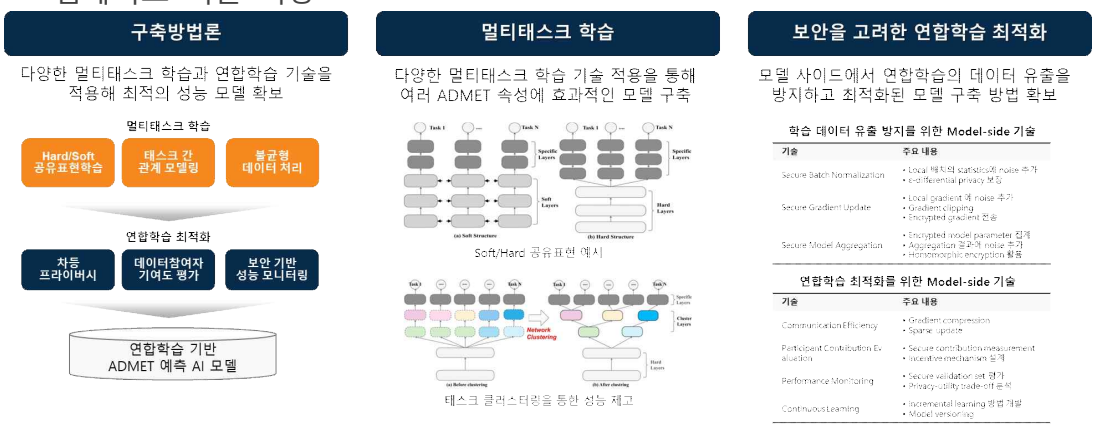
- (단일 기관 데이터 한계 극복 필요) 기존 ADMET 예측 모델은 제한된 기관 데이터에 기반해 학습되어 일반화 성능이 낮음. 다기관 데이터 활용 시, 데이터 불균형 및 이질성 문제로 모델 성능이 저하됨
- (연합학습 환경 특성에 최적화된 모델 부재) 연합학습에서는 각 기관의 로컬

모델 업데이트 후 집계하는 방식으로, 통합 과정에서 정보 손실이나 노이즈 발생 가능성이 있음. 블랙박스 모델 구조로 인해 예측 결과 해석이 어려워, 각 기관의 데이터 특성을 반영한 모델 최적화가 미흡함

- (모델 해석력 및 불확실성 정량화 부족) 연합학습 환경에서 각 기관의 특성을 통합한 모델은 불확실성 평가와 해석 가능성이 중요한데, 기존 접근 방식은 이에 대한 대응이 부족함

□ 연구 내용

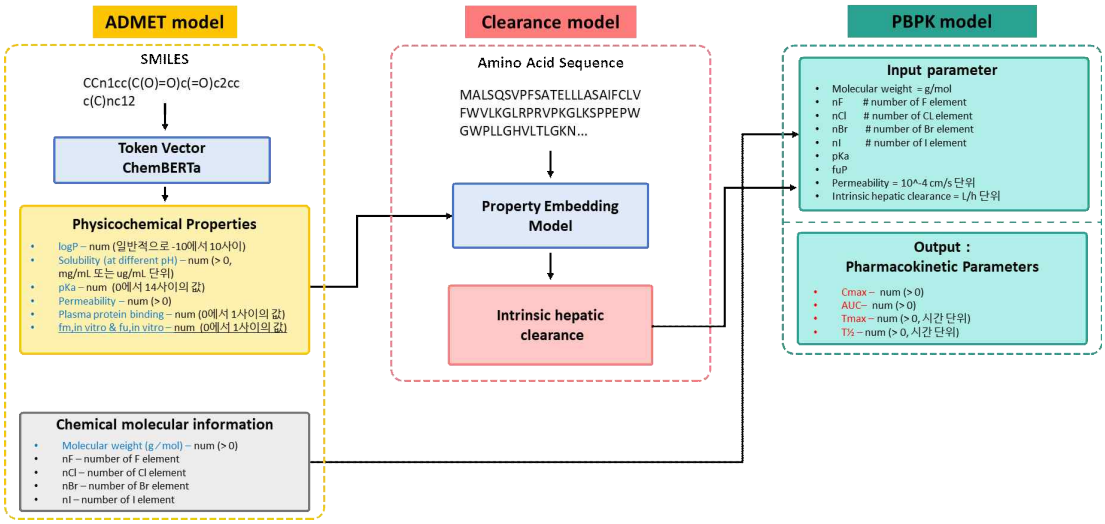
- [연합학습 환경 최적화 모델 아키텍처 개발]
  - 각 기관의 데이터 특성을 효과적으로 반영할 수 있는 모듈형 아키텍처 설계
  - 로컬 모델의 업데이트 정보를 정교하게 집계하고, 노이즈 보정을 위한 가중치 조정 및 분산 최적화 기법 적용
  - 태스크 간 관계를 모델링하여 ADMET 속성별 예측 성능을 동시에 최적화하는 멀티태스크 학습 기법 도입
- 불확실성 정량화 및 해석 가능한 AI 기법 적용
  - 해석가능한 AI(XAI) 기법을 도입하여, 각 기관 데이터의 기여도와 예측 결과 신뢰성을 명확히 평가
  - 모델 예측의 불확실성을 정량화하여, 위험 요소를 사전에 식별하고, 보정 알고리즘을 통한 안정성 향상 도모
- 연합학습 환경에 특화된 성능 개선 전략 수립
  - 각 기관의 데이터 이질성을 반영한 차등 프라이버시 및 기여도 기반 가중치 업데이트 전략 개발
  - 로컬 모델 업데이트 시 발생하는 통신 효율 문제를 고려한 압축 및 스파스 업데이트 기술 적용



[그림 4] 연합학습 환경 최적화 ADMET 예측 알고리즘 개발

연합학습 기반 ADMET-PK/PD 모델 개발

주관: 전북대학교
공동: 충남대학교

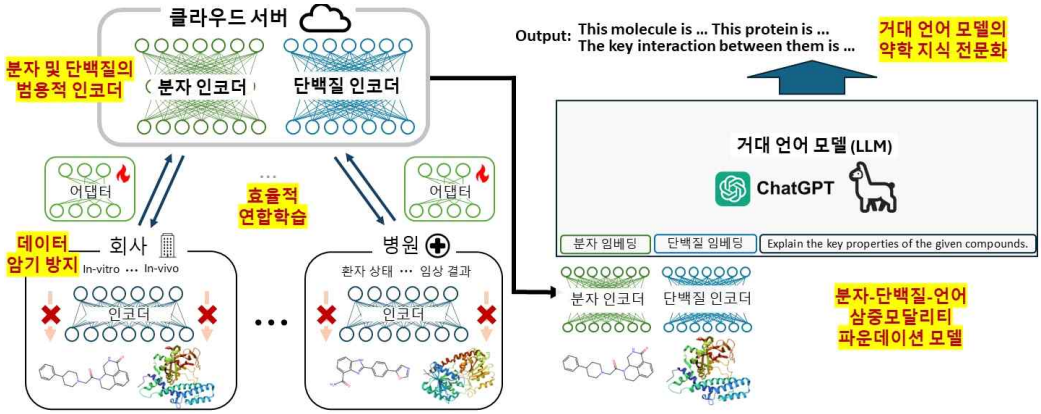
요약문	□ 사업 참여기관 제공 데이터, 공개된 in-vitro / in-vivo feature를 활용하여 연합학습 기반으로 약물 특이적 약동력학적 성질 예측
특징 및 차별성	□ Physiologically based pharmacokinetics (PBPK), systems pharmacology 등과 기전적인(mechanistic) 방법론을 AI 모델과 융합하여 연합학습 기반 예측 모델을 고도화함
핵심 연구내용	 **[그림 1] PBPK 예측 모델 구조도** • 화합물 구조로부터 약동력학적 성질을 예측하는 선행 연구 조사 및 기존 연구의 개선점 탐색 • Drugbank, Drug Interaction Database (DIDB®) 등의 DB에서 약물 연구실 측 자료의 수집 • SMILES로부터 in-vitro feature를 예측하는 모델 개발・탐색 • 화합물 구조, Compound parameter, Physiology parameter를 입력값으로 하여 PBPK 모델 시뮬레이션을 수행할 수 있는 흐름 개발 • 실측과 유사한 수준의 시뮬레이션을 해낼 수 있는 PBPK 모델을 개발하기 위해 AI 기반 입력값을 추정 및 조정함 • 확보한 시간-혈중농도 프로파일로 AUC, Cmax 등의 약동학파라미터 예측

연합학습 기반 설명가능한 임상 결과 예측 분자-단백질-언어 삼중모달리티 파운데이션 모델

주관: 한국과학기술원 KAIST 황성주 교수

요약문	□ **본 연구는 연합학습 기반 설명가능한 임상 결과 예측 분자-단백질-언어 삼중모달리티 파운데이션 모델을 연구 및 개발하여 1) 데이터 보안을 유지하며 방대한 약학 지식을 학습한 파운데이션 모델을 개발하고, 2) 데이터 증강기법을 통한 데이터 희소성 및 불균형 문제를 해결하며, 3) 이를 기반으로 설명가능한 임상 결과 예측 기법을 개발하여 신약 개발을 가속하는 것을 목표로 함**
특징 및 차별성	□ **특징** • **연합학습 적용**: 데이터 보안을 유지하면서 다양한 기관의 데이터를 통합하여 학습하는 방법 연구 • **삼중 모달리티 학습**: 분자, 단백질, 언어 정보를 함께 학습하는 파운데이션 모델 개발 • **설명가능한 AI**: AI의 추론 과정과 근거를 제공하여 연구자가 신뢰할 수 있는 신약 개발 모델을 구축 □ **차별성** • 기존 자연어 및 이미지 중심 거대모델과 달리 신약 개발을 위한 약학 특화 거대모델을 개발 • 단순한 약물-단백질 상호작용 예측이 아닌 임상 결과까지 예측가능한 모델을 구축 • 데이터 보안 이슈를 해결하기 위해 연합학습을 도입, 기관 간 협력을 촉진하며 활용성을 극대화 • 설명가능성(XAI)을 제공하여 AI 예측 결과의 신뢰성을 높이고 연구자의 의사결정 지원
핵심 연구내용	□ **[1단계] 연합학습 기반 분자-단백질-언어 삼중 모달리티 파운데이션 모델** • 문제 제기 - (데이터 보안) 신약 데이터는 개인정보 및 지식재산권을 보유하고 있기에, 데이터 유출은 중요한 문제임. 하지만 현재의 기술 수준이 이를 완벽하게 수행하기 어려우며, 데이터의 공유 및 암기 없이 범용적 지식을 학습한 모델의 개발이 필요함 - (데이터 다양성) 데이터 구조 및 형태가 상이하므로 통합된 모델의 개발 필요 - (약학 특화 거대모델) 범용 지식을 학습한 거대모델은 활용도가 높지만, 임상 결과 및 Drug-Target Interaction (DTI) 예측은 복잡한 약학적 지식이 있어야 하는 태스크임. 따라서, 거대모델의 약학 전문화를 위한 추가학습이 필요함 • 연구 내용 - 다양한 신약 데이터 정보를 분자 및 단백질 인코더에 함축적으로 인코딩한다. in-vitro, in-vivo, in-human 데이터는 그 형태와 종류가 다양하기에 이를 분자 및 단백질 인코더에 함축적으로 임베딩 할 수 있는 인코더를 개발함

- 데이터 암기 정도를 정량화하고 암기 방지 기법을 개발하여, 인코더 성능은 유지하되 인코더가 학습한 데이터를 추론하지 못하도록 함
- 분자 및 단백질 정보를 신약 데이터에 대한 정보 유출 없이 인코딩함. 연합학습에 기반하여 데이터 유출 없이 신약 데이터에 특화된 통합 분자 및 단백질 인코더를 학습함
- 핵심 파라미터 연합학습을 통해 연합학습의 효율성을 높임. 인코더 연합학습 시, 어댑터를 활용하여 핵심 정보를 학습하고 핵심 파라미터만 연합할 수 있게 함
- 분자-단백질-언어 삼중 모달리티 파운데이션 모델을 개발하여 약학 지식에 전문화함. LLM에 기반하여 분자 및 단백질 그래프 임베딩을 입력으로 받아 성질 묘사 및 상호관계 생성 등 다양한 태스크에 추가학습 시켜 약학 지식에 전문화함

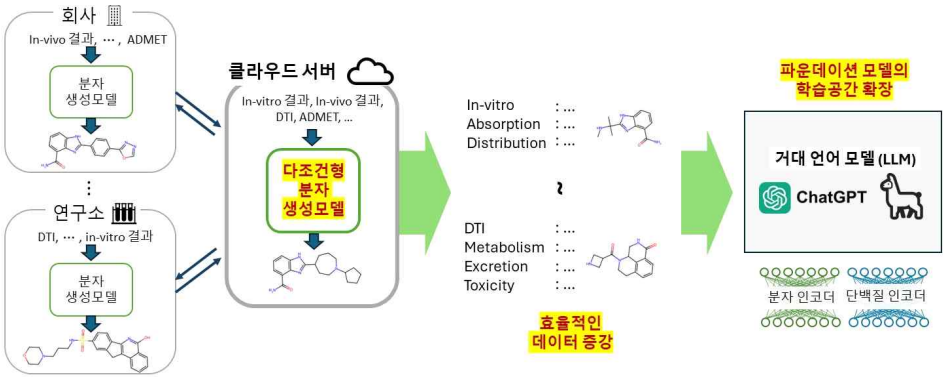


[그림 1] 연합학습 기반 분자-단백질-언어 삼중모달리티 파운데이션 모델

핵심 연구내용

□ **[2단계] 조건형 생성 모델 기반 데이터 증강을 통한 학습데이터 희소성 및 불균형 해결**

- 문제 제기
  - (데이터 희소성) 실험적으로 검증된 신약 데이터는 약 물질 탐색공간에 비해 소수이므로, 새로운 약 후보 물질에 대한 예측이 불확실함
  - (레이블 불균형) 임상을 통과한 물질은 일부이기에 임상 결과 예측 학습을 편향되게 만듦. 이러한 편향된 학습을 보완할 수 있는 학습 기법이 필요함



[그림 2] 조건형 생성 모델 기반 데이터 증강으로 데이터 희소성 및 불균형 해결

- 연구 내용
  - 다양한 조건을 만족시키는 조건형 생성 모델을 학습함 데이터 증강을 위해 주어진 조건을 만족시키는 조건형 분자 생성 모델을 학습함
  - 연합학습을 통해 조건형 생성 모델의 일반성을 보장함. 각 클라이언트에서 학습된 조건

형 생성 모델을 서버 모델에 통합하여, 데이터 유출 없이 생성의 질과 정확도를 높임

- 생성 모델을 이용하여 효율적으로 데이터를 증강함. 데이터 희소성 및 불균형 문제를 효율적으로 해결하기 위해 추가 임상 실험의 수행 없이 생성된 분자 및 조건을 데이터화함
- 증강 데이터의 확실성에 기반하여 파운데이션 모델을 추가 학습시킴. 생성된 분자의 조건 만족 확실성이 낮은 경우, 파운데이션 모델의 추가학습 손실 값에 대해 불이익을 부여하는 방법을 개발하여 추가학습을 안정화함

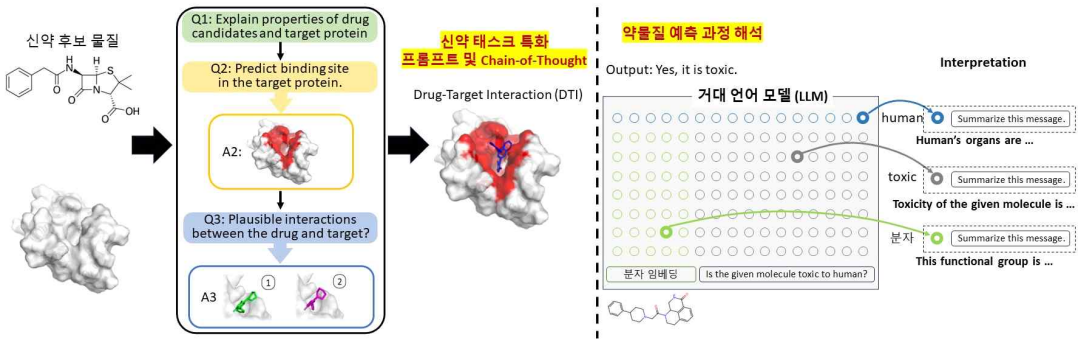
핵심 연구내용

□ **[3단계] 신약 개발 태스크 특화 추론 기법 및 추론 과정 설명 제시**

• 문제 제기

- (임상 결과 예측의 어려움) 신약 후보 물질은 임상을 통과하지 못하는 경우가 많으며, 이는 신약 개발 비용을 증가시키는 요인임. 초기 단계에서 약물 구조로부터 ADMET를 예측하는 모델의 개발이 필요함
- (설명 가능성) 신약 개발 분야에서는 모델 추론뿐 아니라 그 추론에 대한 설명이 뒷받침되어야 하므로 모델 신뢰성을 위한 설명가능성 역시 중요한 문제임. 예측 결과 및 과정에 대한 분자 구조에 대한 설명 기법이 필요함
- (레이블의 불균형) 다양한 신약 후보 물질 중 극히 일부만이 임상을 통과하기 때문의 불균형한 임상 결과 레이블 하의 정확한 예측 기법이 필요함

• 연구 내용



[그림 3] 신약 개발 태스크 특화 추론 기법 및 추론 과정 설명 제시

- 신약 실험 결과 예측에 특화된 프롬프트 기법을 개발함 신약 태스크에 최적화시키기 위해, LLM 기반 파운데이션 모델이 학습한 지식을 추출할 수 있는 최적화 된 프롬프트를 개발함
- 합리적인 추론을 위한 Chain-of-Thought 기법을 개발함. 단계별 추론을 통해 추론 과정에 대한 설명을 제공하는 동시에 합리적인 추론 과정을 통해 레이블의 불균형 문제를 완화함
- 추론 결과에 대한 추가적 해석을 제공함. 단계별 추론 결과에 대한 설명을 추가적으로 제공하여 추론 결과에 포함되지 않은 기저 정보를 파악하고 합리성을 검토할 수 있게 함

연합 메타학습 기반의 확장 가능한 모듈형 융합 프레임워크 및 질량 분석 모듈 개발

주관: 고려대학교(감태의)
고려대학교 정보대학 Korea University College of Informatics

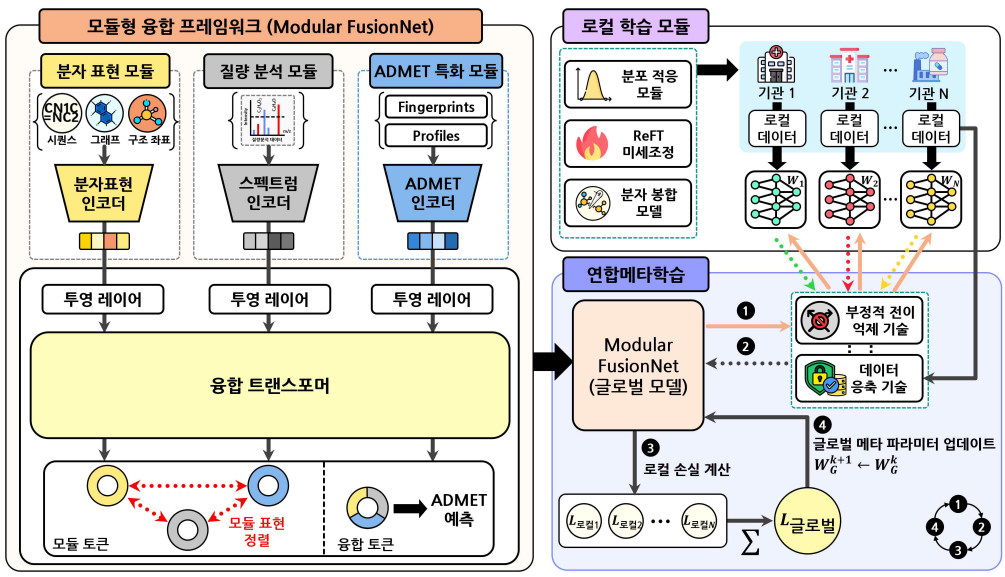
요약문

□ **본 연구는 신약개발 전 과정에서 핵심적으로 요구되는 ADMET 태스크를 고도화하기 위해, 연합 메타학습 기반의 모듈형 융합 프레임워크와 질량 분석 모듈을 통합적으로 개발하고자 함**

- 트랜스포머 기반의 확장 가능한 모듈형 융합 프레임워크를 개발하여 ADMET 예측모델을 개발함. 각 모듈은 독립적으로 개선 및 교체가 가능하도록 설계하여 기능 확장 역시 가능하게 함
- 질량 분석 모듈을 개발하여 대사체 분석 정보를 분자 표현 학습 과정에 통합함으로써, 단순 구조 기반 예측을 넘어 실험 기반 생체 반응 정보를 반영하는 융합 프레임워크로 고도화함
- 기관별 소량·이질성 데이터를 효과적으로 활용하기 위해 연합 메타 학습을 적용함. 각 기관은 로컬 데이터로 모델을 빠르게 적응시키고, 이를 통해 범용성과 개인화 성능을 동시에 확보하는 구조를 구현하고자 함

핵심 연구내용

□ **전체 연구 개요도**



[그림 1] 제안하는 연구개발 내용의 개요도

□ **1차년도: 확장 가능한 모듈형 융합 프레임워크 및 질량 분석 데이터 전처리기 개발**

- TDC, PubMed, ZINC, MoleculeNet 등 공개 분자 데이터셋을 예측 항목별로 정리하여 데이터 통합 파이프라인을 구축함. 기관별 데이터 형식 차이에 대응하는 범용 전처리 체계를 확장하고, 3D 구조 정보가 부족한 분자에 대해서는 conformer 생성을 적용하여 2D·3D 데이터를 통합함. 모든 데이터를 일관된 입력 형식으로 변환하여 ADMET 예측 모델 학습에 활용함

- 1D(SMILES), 2D(Graph), 3D(Conformer) 등 다중 분자 표현에 최적화된 자연어 및 그래프 기반 모델을 활용하여 분자 표현 모듈을 설계함. TDC 벤치마크를 기반으로 각 ADMET 태스크별 특화 모듈을 개발하고, 출력 임베딩을 표준화하여 융합 트랜스포머와의 효율적 결합이 가능하도록 설계함
- 사전 학습된 분자 표현 모듈을 기반으로 모듈형 융합 프레임워크의 사전학습을 수행함. 모듈 간 의미적 일관성을 확보하기 위해 단일 모듈 토큰뿐 아니라 융합 토큰 조합을 고려하는 조합 손실(Combinatorial Loss)을 도입하여 정렬을 유도함
- 오프라인 강화학습 기반 분자 봉합(Molecular Stitching) 기술을 개발하여 기존 분자를 조합함으로써 신규 후보 분자를 생성함. SMARTS 기반 화학 반응 템플릿을 적용하여 생성 전 과정에서 화학적으로 유효한 구조만 산출되도록 설계함
- 질량 스펙트럼 데이터를 통합하여 분자 구조와 스펙트럼 정보를 동시에 반영하는 멀티모달 학습 기반을 구축함. MassSpecGym, GNPS, HMDB 등 공개 스펙트럼 데이터셋을 통합하여 질량 분석 확장 학습 데이터를 구성함

□ 2차년도: 질량 분석 모듈 개발 및 모듈형 융합 프레임워크 고도화

- 입력 SMILES와 질량 스펙트럼을 정렬하기 위한 스펙트럼-그래프 매칭 신경망을 개발하여 주어진 SMILES와 유사한 질량 스펙트럼을 검색하는 검색기를 구축함
- 검색된 스펙트럼을 참조하여 더 정확한 스펙트럼을 생성하는 질량 스펙트럼 검색 증강 생성기(MS-RAG)를 개발함
- 질량 스펙트럼의 메타정보와 분자 파편 정보를 활용하여 피크를 화학식 단위로 해석하는 GNN 기반 스펙트럼 인코더를 설계하고, 스펙트럼에 나타나지 않는 중성 손실(neutral loss)을 예측·보정하는 네트워크를 개발함. 이를 Fragment Formula Transformer로 통합하여 파편 간 관계를 반영하는 고차원 질량 스펙트럼 임베딩 모델을 구축함
- 질량 스펙트럼 생성기(MS-RAG)와 인코더를 통합하여 해석·생성·임베딩이 가능한 통합형 질량 분석 모듈을 설계·구현하고, Docker 기반 배포가 가능한 구조로 개발함. SMILES 또는 질량 스펙트럼을 입력받아 양방향 분석 결과를 제공하는 인터페이스를 구현하고, 대사 및 배설 분석 등 실제 규제·실험 환경에서 활용 가능한 정량 분석 도구로 완성함
- 질량 분석 모듈을 멀티모달 입력 구성요소로 통합하여 모듈형 융합 프레임워크의 확장성을 강화하고, 경량 미세조정으로 신규 모듈이 융합 트랜스포머에 쉽게 결합되도록 설계함. 특히 대사(Metabolism) 및 배설(Excretion) 태스크에 특화하여 질량 스펙트럼 기반 분자 파편 정보를 반영함으로써 SOTA 수준의 멀티모달 예측 성능 달성을 목표로 고도화함

□ 3차년도: 연합 메타학습 기반의 범용적 글로벌 모델 학습 기술 개발

- 기관별 데이터 분포 차이로 인한 초기 모델 성능 저하를 해결하기 위해, 혼합 샘플링 기반 공개 데이터 구성과 다양한 분포의 가상 데이터 생성을 통해 분포 적응 모듈(distribution adaptor module)을 사전 학습하고, 로컬 데이터를 초기 모델의 학습 분포에 맞게 변환함으로써 다양한 환경에서도 안정적으

	로 대응할 수 있도록 설계함 • 기관별 소량 데이터 한계를 극복하기 위해 스펙트럼-그래프 매칭 기반의 정밀 분자 봉합 기술과 스펙트럼 기반 파편 중요도 분석 기법을 개발하고, 분자 선호 최적화 및 다목적 점진적 최적화 전략을 적용하여 ADMET 조건을 반영하는 조건부 분자 생성 모델로 고도화함으로써 화학적 타당성과 목표 특성을 동시에 만족하는 증강 데이터셋을 구축함 • 분포 적응 모듈과 분자 봉합 기반 데이터 증강을 통합 적용하여 로컬 데이터의 이질성과 소량 문제를 동시에 완화하고, 정렬·증강된 데이터를 기반으로 경량 학습 기법을 적용함으로써 안정적이고 확장 가능한 로컬 모델 학습 기술을 확립함 • 다양한 기관을 메타-작업으로 간주하는 연합 메타학습(Meta-FAM) 기반 범용 글로벌 모델을 구축하고, 기관 간 부정적 전이(negative transfer)를 억제하는 conflict-free 학습 전략과 차등 개인정보 보호 기반 데이터 응축(data condensation) 기술을 적용하여, 데이터 공유 없이도 새로운 환경에 빠르게 적응 가능한 견고한 글로벌 모델을 완성함

연합학습기반 지속적인 멀티모달 앙상블 ADMET 예측 모델 개발

주관: 고려대학교(신웅희)
공동: 경희대학교(최민석)
공동: 아론티어(황창하)

요약문	□ **본 연구는 멀티모달 기반 지속가능한 연합학습 ADMET 예측 모델 개발을 통해 임상시험 성공률을 높이고, 신약 개발 비용과 기간을 단축하는 것을 목표로 함** • 분자의 1·2·3차원 구조, 3D Zernike 기술자(3DZD), 단백질-약물 결합에너지 프로파일을 통합한 **멀티모달 ADMET 예측 인공지능 모델**을 개발하고자 함 • 오프타깃(off-target) 단백질과의 결합 정보를 반영하여 기존 분자 기반 예측의 한계를 극복하고, 독성 및 ADME 예측 정확도를 고도화함 • 기관별 데이터 보안과 이질성을 고려한 **지속가능한 연합학습(PEFT 기반 개인화·모듈화)**체계를 구축함
특징 및 차별성	• 3D 구조 기반 고도 표현(3DZD) 도입 - 기존 SMILES, 2D 그래프 중심 ADMET 예측을 넘어, 분자 표면을 복셀화하여 3D Zernike 기술자로 변환 - 회전 불변성을 가지는 벡터 표현으로 구조 정렬 없이 고속으로 유사도 계산이 가능함 • 오프타겟 단백질 기반 결합 에너지 프로파일 활용 - 독성 관련 단백질 구조 앙상블을 활용한 역가상검색 수행 - 약물 1개당 다차원 결합 에너지 벡터 생성을 통한 원인 단백질까지 설명 가능한 구조적 정보 확보 • 멀티모달 딥러닝 아키텍쳐 고도화 - GCN, LLM, 분자지문, 3DZD등을 통합 - ADMET-DNN, DOPCNN등 목절별 특화 모델 설계 • PEFT 기반 개인화 연합학습 체계 - 서버는 공공데이터로 사전학습 및 버전을 업데이트하고, 클라이언트는 일부 파라미터만 개인화 - 기관별 데이터 분포, 자원, 모달리티 이질성에 댕으하는 모듈화 연합학습
핵심 연구내용	□ **멀티모달 ADMET 데이터 인프라 구축** • 공개문헌 기반 데이터 확보 - PubChem, ChEMBL, Tox21등에서 독성 및 ADME 데이터 수집 - LLM 기반 다중 에이전트 시스템을 활용한 정보 추출 • 전처리 파이프라인 구축 - RDKit 기반 1,2차원 특징 생성

- 3차원 구조 생성후 분자 표면 복셀화를 통하여 3DZD 변환
- 결합에너지 프로파일 생성
- 다중입력을 통합하는 멀티모달 데이터셋 구축

□ **3DZD 및 결합에너지 기반 구조 정보 생성 기술**

• 3D Zernike Descriptor 생성 알고리즘 개발
- 분자 표면 생성 후 복셀화
- 다앵한 구조 생성 알고리즘 비교 후 선정

• Safety44 기반 역가상검색
- 다중 구조 앙상블 구축 후 AutoDock-GPU로 도킹을 수행, AK-Score2로 결합에너지 재평가
- 약물별 44차원 결합에너지 프로파일 생성
- 향후 hERG, CYP등 ADME 관련 단백질로 확장

□ **멀티모달 ADMET 예측 모델 개발**

• ADMET-DNN
- GCN, ChemBERTa, 분자 지문 인코더 구성
- 잠재벡터의 평균, 병합, attention 기반 융합

• ADMET-(D)OPCNN
- 분자지문 기반 outer-product CNN
- 증류 기법을 사용하여 교사-학생 구조를 차용. 데이터가 적은 환경에서 과적합을 완화
- 기관별 소규모 데이터 환경에 적합

□ **지속 가능한 개인화 연합학습 체계 구축**

• PEFT 기반 모듈화 연합학습
- 서버: 전체 모듈 사전학습 및 주기적 업데이트
- 클라이언트: 일부 파라미터만 학습
- 개인화화 일반화 성능의 메타학습 기반 조정

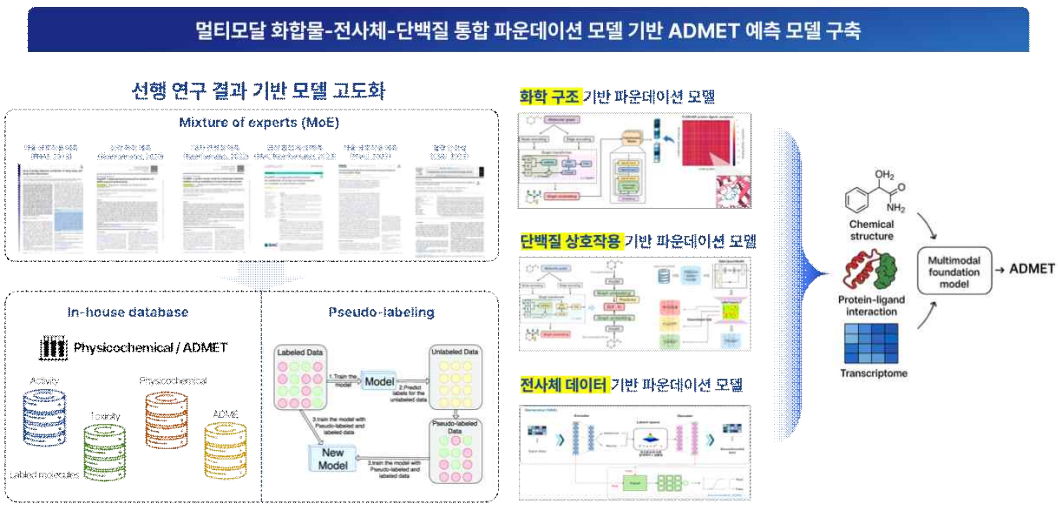
• Non-IID 및 자원 불균형 대응
- Balanced/standard k-medioids 기반 데이터 분포 모델링
- Dirichlet 분포 기반 non-IID 시뮬레이션
- FedAvg, FedGF등을 통해 일반화 성능을 향상

• 강건 연합학습
- Straggler 대응: staleness 기반 가중평균
- Adversary 대응: 엔트로피, 손실 기반 필터링
- 모델 poisoning, backdoor 공격에 대한 방어 설계

• 기여도 기반 인센티브 설계
- 기관별 학습 기여도 측정
- 기여도에 따른 성능 차등 배포

멀티모달 화합물-전사체-단백질 통합 파운데이션 모델 기반 고정밀 ADMET 예측 기술 개발

주관: 숭실대학교(류재용)
공동: 한국화학연구원(장우대)
공동: 온코크로스(김이랑)

요약문	□ 본 연구팀은 화합물 구조 정보, 단백질 상호작용 데이터, 세포 내 유전자 발현(전사체) 정보를 통합 학습하는 '멀티모달 화합물-전사체-단백질 통합 파운데이션 모델'을 구축하여, 생물학적 기전이 반영된 고정밀 ADMET 예측 기술을 개발함
특징 및 차별성	□ 생물학적 맥락이 반영된 차세대 모델링: 기존의 화학 구조(Chemical Structure) 중심 QSAR 방식의 한계를 넘어, 실제 세포 반응인 약물 유도 전사체(Transcriptome) 데이터와 단백질-리간드 결합 데이터를 결합하여 약물의 생체 내 작용 기전을 직접적으로 모델링함  [그림 1] 연구개발 내용 □ 기전 해석이 가능한 설명 가능한 AI (XAI): 블랙박스 형태의 단순 예측을 지양하고, 결과 도출에 기여한 생물학적 경로(Pathway) 활성도를 함께 제시하여 연구자가 예측 결과의 타당성을 직관적으로 검증할 수 있음 □ 데이터 희소성 극복 및 일반화 성능 확보: 레이블이 부족한 ADMET 데이터의 한계를 대규모 전사체 데이터 기반의 사전 학습(Pre-training)과 전이 학습(Transfer Learning)으로 극복하여 과적합을 방지하고 높은 예측 성능을 확보함
핵심 연구내용	□ 단백질-리간드 복합체 파운데이션 모델 구축 • 단백질-리간드 복합체 데이터 확보 - 단백질과 화합물 간 물리적 상호작용은 ADMET를 예측하는데 있어서 중요하

게 활용될 수 있음. 이러한 정보를 이용하여 모델을 구축하기 위해, 구조적 다양성을 갖는 15,000개 이상의 library 화합물과 주요 GPCR, Kinase, Druggable target protein, 구조 정보가 존재하는 단백질을 포함하여 약 4,900개에 해당하는 단백질의 구조 정보를 확보하였음

- 구조 정보가 존재하지 않은 단백질의 경우 AlphaFold2를 통해 구조 정보를 확보하고, binding site를 예측하여 분자 도킹 위치를 결정하였음
- 기 구축된 비공개 데이터로서 총 75,000,000개의 protein-ligand complex의 binding affinity 정보를 확보함

• 단백질-리간드 복합체 데이터 기반 모델 구축

- 파운데이션 모델은 chemical structure를 graph 등 embedding 하는 부분과, 이를 transformer 구조로 binding affinity를 self-supervised learning으로 학습하는 전략으로 구축하고자 함
- 이 과정에서, MiniMol, MolGPS 등 알려진 파운데이션 모델 아키텍처를 평가하고 성능을 고도화 하고자 함
- 단백질-리간드 복합체 파운데이션 모델을 구축함으로써, 화합물의 미세한 구조적 차이가 미치는 영향을 반영한 ADMET 예측이 가능할 수 있으며, 특히 단백질 상호작용이 중요한 종류의 ADMET 예측에 있어서 성능향상이 기대됨

□ 약물 구조-전사체 연계 멀티모달 파운데이션 모델 구축

• 도메인 특화 인코더(Encoder) 설계 및 임베딩

- 공개 데이터베이스에서 수집한 화합물 구조(SMILES)는 Transformer 기반 모델을 사용하여 분자 시퀀스 특성을 학습하고 고차원 벡터로 압축함
- 전사체 데이터는 AVAE(Variational Autoencoder + GAN) 구조를 활용하여, 유전자 발현 프로파일을 저차원 잠재 공간(Latent Space)으로 정교하게 압축·임베딩하며, 이 과정에서 적대적 학습(Adversarial Learning)을 통해 데이터 분포의 연속성과 일반화 성능을 극대화함

• 크로스모달(Cross-modal) 학습 및 정렬

- 이종 데이터인 화학 구조와 전사체 반응을 연결하기 위해 대조 학습(Contrastive Learning)과 매칭 목적함수(Matching Objectives)를 동시에 적용함
- 동일 약물의 구조-전사체 임베딩 간 거리는 좁히고 다른 약물은 멀어지게 학습하며, 임베딩 간 코사인 유사도 예측을 통해 두 도메인 간의 연관관계를 정밀하게 정렬(Alignment)함

• 멀티태스크 기반 기전 내재화 및 ADMET 예측

- 모델에 Pathway Regression Head와 Alignment Head를 도입한 멀티헤드 구조로 학습하여, 실제 생물학적 경로(Pathway) 활성도를 예측하게 함으로써 기전 정보를 모델 내부에 내재화함
- 이후 사전 학습된 인코더를 고정(Freeze)하고 ADMET 예측 전용 분류층(Classifier)만 미세조정(Fine-tuning)하는 전략을 사용하며, Pathway 예측 손실을 보조 손실로 활용하여 과적합을 방지하고 예측 결과에 대한 생물학적 해석을 제공함

□ 화학구조 기반 파운데이션 모델 구축

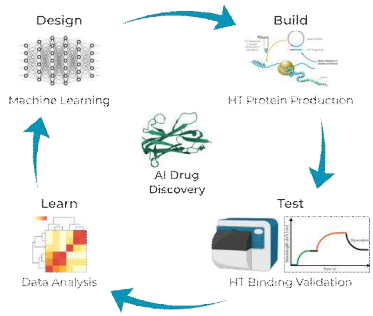
• 공동 임베딩 예측 아키텍처(Joint-Embedding Predictive Architecture, JEPA) 기반의 파운데이션 모델 구축

- 공개 화합물 데이터베이스 ZINC-22 내의 lead-like 분자 약 130억종과 한국화합물은행 화합물 78만종을 활용
- Graph transformer 모델을 통해 화합물에 대한 임베딩을 수행하고, 자기지도 학습 과정은 JEPA 기반으로 파운데이션 모델을 구축
- 원본 분자와 손상된 분자 사이의 핵심적인 특징(공통정보)에 집중하여 노이즈와 불필요한 정보는 필터링하는 방식으로 학습되며, ADMET와 같은 다운스트림 태스크 모델 개발 시에 일반화 성능이 우수한 모델을 개발할 수 있을 것으로 예상
- ChEMBL API를 통해 ADMET 항목과 물리화학 정보 수집(Solubility: 34,130개, 혈장단백질결합(PPB): 5,590개, 세투투과도(Permeability): 6,108개, LogP/LogD: 40,465개, hERG: 10,550개, CYP: 13,894개)
- In vivo PK 항목에 대한 총 65,000 건에 대한 사전학습 데이터 수집하였으며, 사전 수집된 ADMET와 In vivo PK 데이터로 멀티태스크 모델 구축

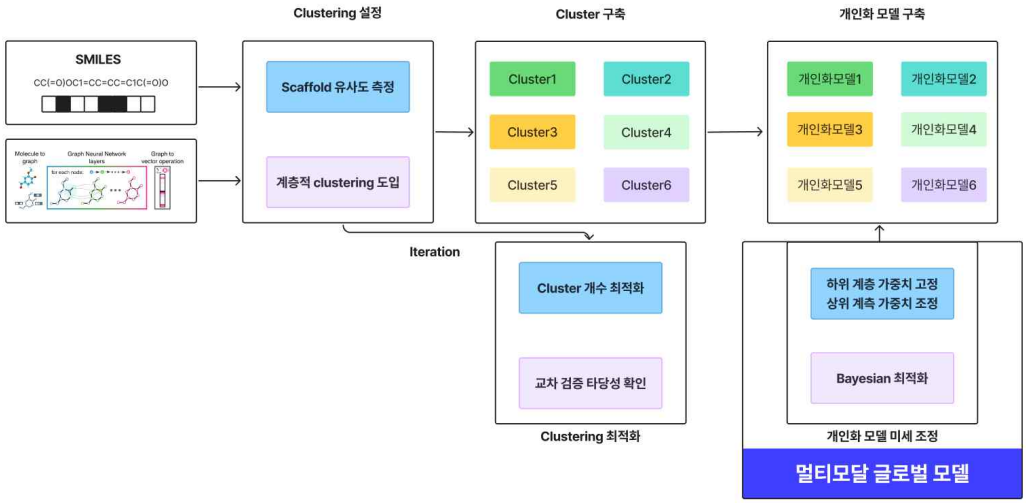
Transformer와 GNN 기반 분자 클러스터링 특화 FAM 모델 개발

주관: LG화학(김승하)

요약문	□ **연합학습 기반 분자 클러스터링 특화 ADMET 예측 모델을 개발하기 위해 PubChem/ ChEMBL/ TDC 등 공개 DB와 LG화학 내부 ADMET 데이터를 통합해 10만 건 이상 분자 데이터를 확보하고 자동화 전처리 파이프라인으로 품질을 관리함으로써, 신약개발 초기 단계에서 ADMET 리스크를 정밀 예측해 후보물질 실패를 조기에 줄이고 개발 비용·기간을 단축하는 것을 목표로 함** □ **SMILES - Transformer와 분자 그래프** • GNN을 결합한 멀티모달 글로벌 모델을 구축하고, 스캐폴드 기반 클러스터별 개인화 (미세조정) 및 Secure Aggregation/ Differential Privacy 기반 연합학습 PoC를 통해 보안/ 성능/ 운영성을 동시에 달성함
특징 및 차별성	□ **멀티모달 글로벌 분자 표현(Transformer+GNN)** • SMILES 시퀀스와 2D 분자 그래프를 Cross-attention / 공통 Transformer 인코더 / late-fusion으로 융합해 상호보완적 분자 표현을 생성하고, 노드/엣지/토큰 마스킹 기반 자기지도학습·대조학습·다중작업학습으로 라벨 희소성에도 강건한 표현학습 체계를 구축함 □ **내부 5만+α HTS 화합물 기반 실전형 ADMET 데이터 자산화** • 내부 저분자 화합물 약 5만여 종+α(HTS) 및 과거 데이터를 체계적으로 정리하고, 표준화·검증 프로세스로 신뢰성 높은 ADMET 학습데이터를 제공함 □ **스캐폴드 기반 클러스터링·개인화로 화학공간별 예측력 극대화** • 분자 스캐폴드 기반으로 구조적 유사성 그룹을 형성하고 클러스터별 특화 미세조정을 수행하여, 데이터 분포가 다른 화학공간에서도 성능/안정성을 향상시킴 □ **보안형 연합학습 운영(PoC) + 표준화** • 참여기관 간 스키마/메타데이터를 일치화하고, Secure Aggregation으로 업데이트를 암호화 공유하며 Differential Privacy로 단일 데이터 포인트 영향도를 제한해 재구성 공격을 방어함. 또한 FedAvg/FedProx 비교를 통해 이질적 분포에서도 안정적인 연합학습 전략을 도출하고 PoC로 구현 가능성을 검증함
핵심 연구내용	□ **[1단계] ADMET 데이터 통합·전처리 플랫폼 구축 및 분자 클러스터링 PoC** • 문제 제기 - (데이터 사일로) 공개 DB와 내부 실험데이터가 분리·비정형으로 관리되어 통합 분석/학습 효율이 낮음 - (품질/표준화) 기관별 스키마·메타데이터 불일치로 학습 시 잡음/편향이 발생할 수 있음

	• 연구 내용 - 공개 DB(PubChem PUG-REST, ChEMBL Web Services, TDC 등) 수집 자동화 및 증분 업데이트 가능한 ETL 파이프라인 구축(전처리 자동화로 데이터 품질 보장) - 내부 5만+α HTS 화합물 및 과거 데이터 정리, 실험데이터 표준화/검증 프로세스 확립, 추가 ADMET 실험 데이터 확장 - 분자 클러스터링 방법 테스트 및 MoE 기반 접근을 활용한 스캐폴드 구조 구분 테스트(개인화 단계의 기반 확보)  **[그림 1] 데이터 자산화·품질관리·연구 선순환 체계**
핵심 연구내용	□ **[2단계] 멀티모달 글로벌 ADMET 예측 모델(Transformer+GNN) 개발** • 문제 제기 - (표현 한계) 단일 모달(SMILES만/그래프만)로는 화학 결합/ 토폴로지/ 문맥 정보를 충분히 반영하기 어려움 - (라벨 비용) ADMET 라벨은 실험 비용이 높아 희소·불균형이 발생하기 쉬움 • 연구 내용 - SMILES + 분자 그래프를 Cross-attention/ Transformer encoder/ late-fusion으로 통합하여 상호보완적 표현 생성 - 화학 결합 타입·토폴로지 구조를 효과적으로 인코딩하는 화학 특화 Transformer 변형 (GTN/MAT 등) 검토/ 적용 및 대규모 학습 최적화 - 마스킹 기반 자기지도학습·대조학습·다중작업학습 결합으로 라벨이 제한적인 상황에서도 표현학습 성능 확보
핵심 연구내용	□ **[3단계] 스캐폴드 기반 개인화 + 성능 평가·검증** • 문제 제기 - (화학공간 이질성) 스캐폴드/시리즈별 분포 차이로 글로벌 단일 모델의 국소 성능 저하가 발생 - (신뢰성 요구) ADMET 예측은 End-point별 지표 기반의 정량 검증과 일관된 평가 체계가 필요 • 연구 내용 - 스캐폴드 기반 개인화 모델 구축: Tanimoto 유사도 기반 정밀 클러스터링을 수행하고, 클러스터별 특화 미세조정 및 하이퍼파라미터 최적화를 통해 화학 공간별 예측 성능을 극대화함

- 성능 평가·검증 체계 확립: BBB, PPB, 간 마이크로좀 대사 안정성, PAMPA, hERG, CYP450, Ames 등 in-vitro End-point를 대상으로 성능을 검증하고, TDC ADMET Benchmark 기반 AUROC/Spearman/MAE 지표로 연차별 목표를 설정하여 최종 성능 개선을 달성함
- 운영화 연계: End-point 패널 기반 리포팅(평가 항목 한 장 요약)과 모델 업데이트/검증 루프를 플랫폼 실증 흐름에 연결하여, 개인화 모델의 실사용 가능성을 확보함



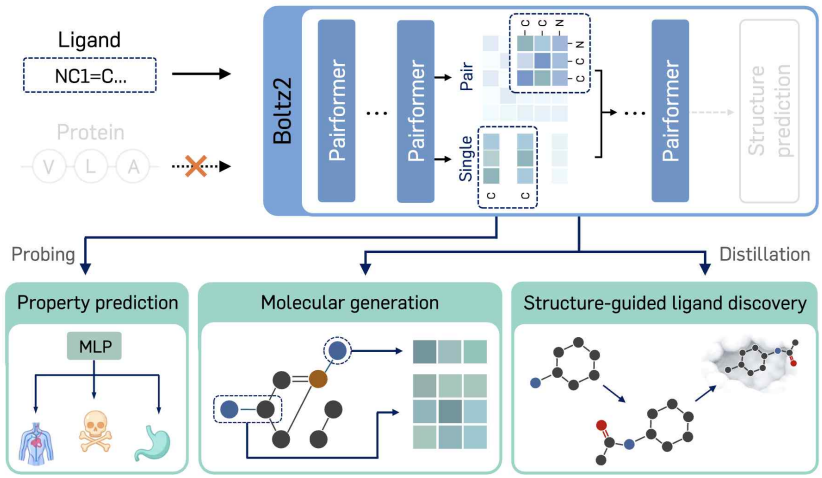
[그림 2] 스캐폴드 기반 개인화(클러스터링→미세조정→검증) 흐름도

ADMET 예측을 위한 초다중모달(약물, 단백질, 유전체, 세포 이미징, 문헌) 파운데이션 모델 개발

주관: 한국과학기술원(안성수)

KAIST 안성수 교수

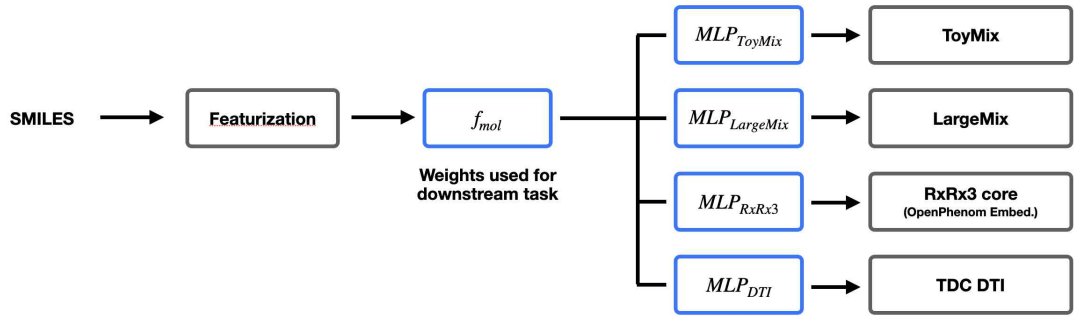
요약문	□ **기존 단일 모달리티 기반 예측은 복합 생체 상호작용 학습이 제한되며, 약물 데이터는 규모 한계로 파운데이션 모델의 표현력을 활용하기 어려움. 이에 본 연구는 약물과 초다중모달 신약 데이터를 이용하는 파운데이션 모델을 개발하고자 함** □ **MolsGPS 기반 계획에서 공개 제약으로 인해, Boltz-2 기반 분자 임베딩 전략으로 전환하여 ADMET downstream task에서 성능을 확보함**
특징 및 차별성	□ **특징** • 약물, 단백질, 유전체, 세포 이미지, 문헌을 통합한 초다중모달 학습 구조 • Boltz-2 기반 구조 임베딩을 활용한 고표현력 분자 인코더 • 세포 이미지 기반 멀티모달 pre-training □ **차별성** • 기존의 단일 모달리티가 아닌, 생체 상호작용에 대해 학습하는 통합 모델 • 모달리티 간 단순 결합이 아닌, 각 모달리티의 파운데이션 모델을 활용하여 표현 정렬 기반 학습 전략
핵심 연구내용	□ **약물 관련 멀티모달 멀티모딩 데이터베이스 구축** • 본 연구에서는 ADMET 파운데이션 모델 학습을 위해 대규모 공개 데이터베이스와 분산된 실험 논문 데이터들을 정제하여 통합 ADMET 데이터베이스를 구축하고자 함. 최종 평가 지표인 TDC benchmark을 바탕으로, 약물 분자와 관련된 여러 멀티 모달 데이터셋을 완성함



[그림 1] Boltz-2 기반 분자 인코더 아키텍쳐 및 다운스트림 태스크

□ **Boltz-2 기반 분자 인코더로의 전환**

- 초기 계획에서는 MolGPS 기반 분자 임베딩을 활용한 멀티모달 정렬을 목표로 하였으나, 해당 모델의 세부 스펙 및 가중치가 비공개됨에 따라 새로운 분자 인코더가 필요해짐. 이에 대해 Boltz-2 기반 분자 표현으로 전략을 전환하였으며, 기존 단백질-리간드 결합 생성 기반의 구조적 표현을 학습하는 모델을 원자 수준 표현 모델로 재사용함
- 변형된 Boltz-2를 통해 원자 수준에서 고차원 임베딩 제공이 가능해졌으며, 이를 바탕으로 다중 모달 임베딩과 정렬하는 pre-training 파이프라인을 설계중. 사전 학습 없이 앙상블 기법만 사용한 경우의 기존 분자 인코더들과 비교 시 TDC ADMET 22개의 태스크 중 11개에서 최고 성능을 달성함



[그림 2] LargeMix, RxRx3core, TDC DCTI 멀티모달 임베딩 pre-training 파이프라인

□ **세포 이미지 기반 멀티모달 pre-training 효과**

- 세포 이미징 데이터를 약물 처리 후 발생하는 표현형 변화를 직접적으로 반영하는 중요한 생물학적 신호를 포함함. 본 연구에서는 OpenPhenom을 활용하여 RxRx3-core 세포 이미지로부터 384차원의 임베딩을 추출하고, 이를 GPS++의 분자 임베딩과 연결하는 정렬 학습을 수행을 Boltz-2 기반 아키텍쳐 개발과 병렬적으로 진행함.
- 구체적으로, 멀티모달 pre-training에서는 기존에 사용하는 LargeMix 데이터셋에 세포 이미지 임베딩을 회귀 목표로 추가한 설정을 실험함. 해당 경우, TDC ADMET benchmark 22개 태스크 중 절반 이상에서 성능 향상을 확인함. 이는 세포 수준 표현형 정보가 구조 기반 분자 표현과 상보적으로 작용하여 ADMET 예측의 일반화 성능을 향상시킬 수 있음을 시사함.

□ **향후 계획**

- 향후에는 단백질, 유전체, 문헌 임베딩을 포함한 보다 확장 초다중모달 pre-training 파이프라인을 구축하고, 각 모달 간 정렬 전략을 고도화할 계획임. 또한 새로 개발한 Boltz-2 기반 분자 인코더의 구조적 표현을 pre-training 파이프라인에 맞추어, ADMET 파운데이션 모델의 범용성과 성능을 동시에 향상시키고자 함