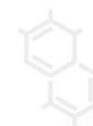


2025년도 연합학습 기반 신약개발 가속화 프로젝트
착수 보고회



연합학습기반 지속적인 멀티모달 앙상블 ADMET 예측 모델 개발

Development of a Continuous Multi-modal Ensemble
ADMET Prediction Model based on Federated Learning



주관연구책임자 |

고려대학교 의료정보학교실 **신용희**

공 동 연 구 자 |

아론티어 **황창하**

공 동 연 구 자 |

경희대학교 전자공학과 **최민석**

2025.7.25

연구 배경 및 필요성

ADMET 예측의 중요성

- 신약 개발 **성공률 향상**
: ADMET 예측 결과를 활용한 약물 구조 개선
- 개발 **비용 및 시간 단축**
: ADMET 예측을 통해 독성 이슈 조기 발견
- **효율적인 약물 개발 전략 수립**
: 약물 구조 개선을 통한 대사 안정성 향상

ADMET - AI 모델의 현황

- ADMET는 임상시험의 가장 중요한 요소
: 신약개발 실패의 주요 원인(독성)
- in vitro 실험 기반의 **AI 모델 성능 부족**
: in vivo 및 임상시험을 보장하지 못함
- ADMET 임상통과 예측 AI 모델 구축 위한
임상시험 **데이터 부족 및 공유 불가**

연합학습의 중요성

- **데이터 부족 문제 해결**
: 타기관 보유 데이터를 안전하게 공유 및 활용
- **AI 모델 성능 향상**
: 다양한 데이터 학습을 통한 AI 성능 향상
- **협업 강화 및 지식 공유 촉진**
: 제약사, 연구기관, 병원 등 협업 및 지식 공유

ADMET 예측 모델 FAM 개발 연구의 필요성

- 1 기밀성(privacy)과 지속성이 확보된 연합학습 모델 필요
- 2 멀티모달 활용 ADMET 예측 모델 연구 부족
- 3 다양한 형태의 데이터를 활용한 다중 작업용 예측 모델 필요
- 4 다양한 형태의 데이터가 **여러 기관에 분산 저장**
- 5 ADMET 예측모델의 **정확성**은 신약개발 비용 절감, 임상시험 성공률 향상, 환자 안전 보장 등 다양한 측면에서 중요한 역할



다양한 형태의 신약개발 데이터(**SMILES, 분자 지문, 그래프, 3차원 구조, 3DZD**)를 활용한 **ADMET 모델에 대한 연구 부족**



기밀성(privacy) 및 지속성이 담보된 연합학습 기반의 **ADMET 예측 모델에 대한 연구 부족**



Continuous ADMET Multi-Modal Ensemble Federated Learning (CAMEL)



연구 목표

연합학습기반 지속적인 멀티모달 양상불 ADMET 예측 모델 개발

Aim 1. 학습 데이터베이스
구축 및 전처리기 개발

Aim 2. 멀티모달 딥러닝
ADMET 예측 모델 개발

Aim 3. 지속가능 연합학습
알고리즘 개발



LLM을 이용한 데이터베이스 구축
3DZD와 결합에너지 프로파일
기반 전처리기 개발



ADMET-DNN, ADMET-DOPCNN
모델 구축



경희대학교
KYUNG HEE UNIVERSITY

Non-IID, 공격 취약성에 대응하는
연합학습 기술 개발

정량목표

- ✓ 모델 개발 (2건 이상)
- ✓ SW 등록 및 인증 (2건 이상)
- ✓ 특허 등록(2건 이상)
- ✓ 논문 (SCIE급) (4건 이상)

3DZD 생성, 결합 에너지 프로파일 생성
알고리즘 구현 및 전처리기 개발

- ◆ 학습용 입력 데이터 표준화 포맷 구성
 - 1차원 SMILES, 2차원 그래프 구조, 2차원 구조 기반 분자 지문, 분자 기술자(molecular descriptor), 3D Zernike Descriptor (3DZD), 결합 에너지 프로파일
- ◆ 3DZD 생성 모듈 구현
 - 3차원 분자의 표면으로부터 3DZD로 변환하는 파이썬 모듈 구현
- ◆ 결합에너지 프로파일 생성 파이프라인 구현
 - 역가상검색을 활용하여 결합에너지 벡터를 얻는 파이프라인 구축

LLM을 이용한 학습용 데이터베이스 구축

- ◆ LLM 모델을 이용한 전임상/임상 데이터 확보
 - PubMed, ChEBML 등 공개 데이터로부터 ADMET 관련 문헌 확보
 - LLaMa를 이용한 문헌 필터링
 - 다중 에이전트 시스템
- ◆ 기 구축된 50만여건의 독성 데이터 이외에 10만건 추가 확보를 목표로 함
- ◆ 전처리기를 활용하여 모델 학습 및 검증에 사용

Aim 1

학습용 데이터베이스 구축 및 전처리기 개발

▶ 전처리기: 3D Zernike Descriptor 및 결합 에너지 프로파일 도입

3차원 구조 기술의 필요성

분자의 3차원 구조를 고속으로 분석할 수 있는 기술이 필요

- 분자 특성 예측 성능을 위해서 3차원 구조가 중요함.
- 3D 좌표를 사용했을 때 성능이 12% 향상됨 (Fang, 2022).

nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-021-00638-4

2022년 발표된
해외 그룹 연구 결과

OPEN Geometry-enhanced molecular representation learning for property prediction

Classification (higher is better)

Dataset	BACE	BBBP	ClinTox	SIDER	Tox21	ToxCast	HIV	MUV	PCBA	Avg
No. molecules	1,513	2,039	1,478	1,427	7,831	8,575	41,127	93,087	437,929	
No. prediction tasks	1	1	2	27	12	617	1	17	128	
D-MPNN ¹	0.809 _(0.004)	*0.710 _(0.001)	*0.906_(0.006)	0.570 _(0.007)	0.759 _(0.007)	0.655 _(0.003)	0.771 _(0.001)	0.786 _(0.014)	*0.862 _(0.001)	*0.759
AttentiveFP ¹⁴	0.784 _(0.021)	0.643 _(0.008)	0.847 _(0.003)	0.606 _(0.002)	0.761 _(0.002)	0.637 _(0.002)	0.757 _(0.014)	0.766 _(0.001)	0.801 _(0.004)	0.734
N-Gram ¹⁵	0.779 _(0.003)	0.697 _(0.004)	0.775 _(0.004)	*0.668 _(0.001)	0.743 _(0.004)	—	0.772 _(0.004)	0.769 _(0.007)	—	—
N-Gram ¹⁶	0.791 _(0.003)	0.691 _(0.004)	0.875 _(0.001)	0.655 _(0.007)	0.758 _(0.008)	—	0.787 _(0.004)	0.748 _(0.003)	—	—
PretrainGNN ¹	*0.845 _(0.003)	0.687 _(0.003)	0.726 _(0.001)	0.627 _(0.004)	*0.781 _(0.006)	*0.657 _(0.004)	*0.799 _(0.005)	*0.813 _(0.003)	0.860 _(0.001)	0.755
GROVER ¹⁸	0.826 _(0.005)	0.700 _(0.001)	0.812 _(0.003)	0.648 _(0.004)	0.743 _(0.001)	0.654 _(0.004)	0.625 _(0.001)	0.673 _(0.008)	0.765 _(0.001)	0.716
BRUNNEN ¹⁹	0.890_(0.001)	0.739_(0.001)	0.912_(0.001)	*0.890_(0.001)	*0.795_(0.001)	0.695_(0.001)	0.802_(0.001)	0.819_(0.001)	0.830_(0.001)	0.729
GEM	0.856_(0.001)	0.724_(0.004)	0.901_(0.001)	0.672_(0.004)	0.781_(0.001)	0.692_(0.004)	0.806_(0.001)	0.817_(0.001)	0.866_(0.001)	0.791

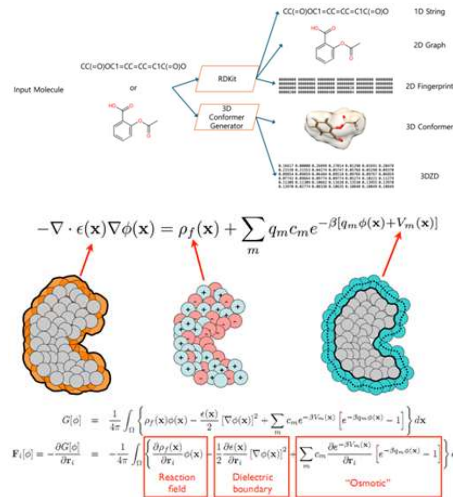
Table 2 | Performance of different GNN architectures for regression tasks

Method	RMSE			MAE		
	ESOL	FreeSolv	Lipo	QM7	QM8	QM9
GIN ¹	1.067 _(0.001)	2.346 _(0.022)	0.757 _(0.022)	110.3 _(1.2)	0.0199 _(0.0002)	0.00886 _(0.00001)
GAT ¹	1.556 _(0.001)	3.559 _(0.001)	1.021 _(0.021)	103.0 _(1.4)	0.0224 _(0.0001)	0.01117 _(0.00001)
GCN ¹	1.211 _(0.002)	3.174 _(0.001)	0.773 _(0.007)	100.0 _(0.3)	0.0203 _(0.0001)	0.00923 _(0.00001)
D-MPNN ¹	1.050 _(0.001)	2.082 _(0.002)	0.683 _(0.006)	103.5 _(0.8)	*0.0190 _(0.0001)	0.00814 _(0.00001)
AttentiveFP ¹⁴	*0.877 _(0.021)	*2.073 _(0.001)	*0.721 _(0.001)	*72.0 _(2.7)	0.0179 _(0.0001)	*0.00812 _(0.00001)
GTransformer ¹	2.298 _(0.001)	4.480 _(0.001)	1.112 _(0.001)	161.3 _(1.1)	0.0361 _(0.0001)	0.00923 _(0.00001)
SGCN ¹	1.629 _(0.001)	2.363 _(0.001)	1.021 _(0.001)	131.3 _(1.4)	0.0285 _(0.0001)	0.01459 _(0.00001)
DimeNet ¹	0.878 _(0.001)	2.094 _(0.001)	0.727 _(0.001)	95.6 _(1.1)	0.0215 _(0.0001)	0.01031 _(0.00001)
BRUNNEN ¹⁹	0.859_(0.001)	2.049_(0.001)	0.691_(0.001)	94.9_(1.1)	0.0199_(0.0001)	0.00859_(0.00001)
GeoGNN	0.832_(0.001)	1.857_(0.001)	0.666_(0.001)	59.0_(1.4)	0.0173_(0.0001)	0.00746_(0.00001)

3DZD 개발 계획

사전연구를 통한 성능 및 속도 향상 확인

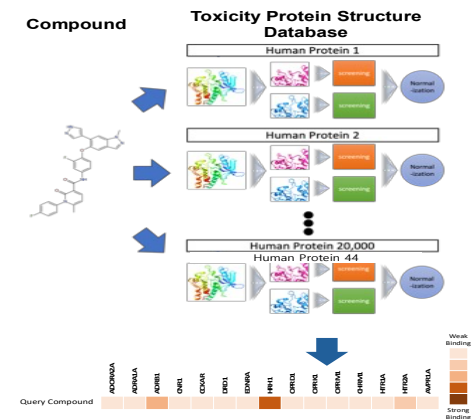
- 3차원 물체를 기술하고 빠르게 검색하는 목적으로 고안됨. (300배 이상 빠른 연산가능)
- 정전기 퍼텐셜, 소수성, 표면의 곡률 등을 이용하여 분자의 물리화학적 성질을 3DZD로 표현 가능.
- 분자의 3차원 구조 생성 및 전처리 기법 확립



결합에너지 프로파일

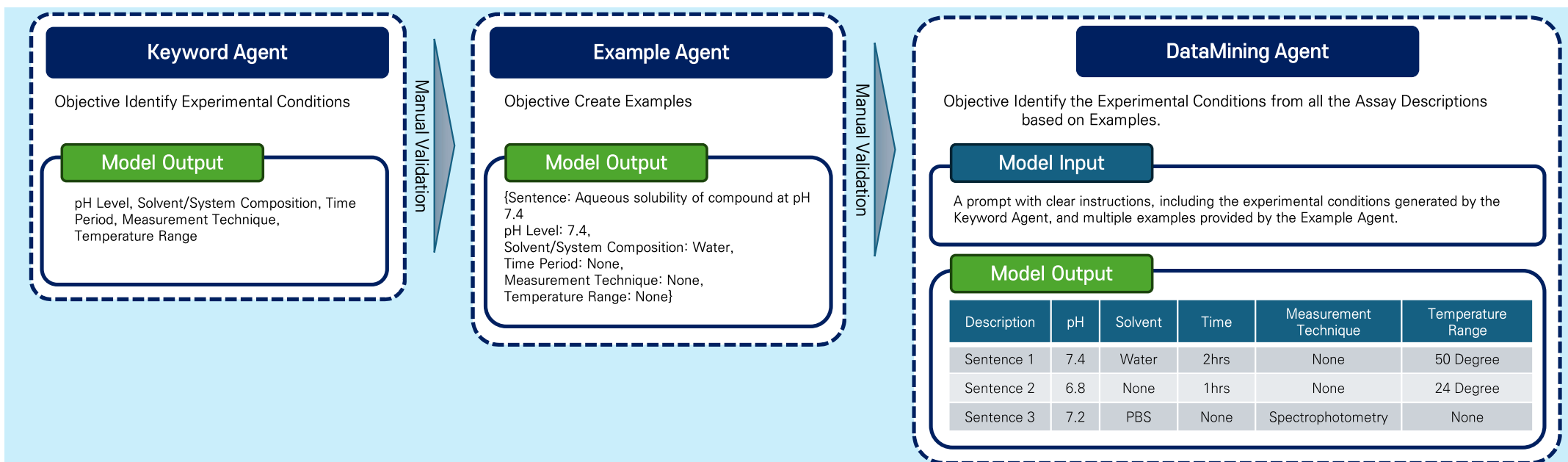
역가상검색 및 AlphaFold 구조를 이용

- 독성 등 약물의 부작용은 표적이 아닌 단백질에 결합하는 off-target binding에 의하여 발생하며 한 약물은 평균 6개의 단백질에 결합함 (Schieferdecker, 2022)
- Bowes et al. (2012) 등 독성과 관련된 단백질의 구조 데이터를 확보하고 역가상검색을 통해 얻어진 결합 에너지 프로파일을 입력 값들 중 하나로 삼음



▶ 학습데이터 확보: LLM기반 다중 에이전트를 이용한 10만 건 이상의 의료화학 문헌 정보 수집

- LLM 모델을 사용한 전임상/임상 및 화학 데이터 수집
 - PubMed, ChEMBL, BindingDB 등과 같은 생물의학 및 화학 데이터베이스에서 ADME-Tox 관련 논문 수집.
 - Llama3 모델을 사용하여 키워드 기반 검색으로 관련 논문을 자동으로 필터링함.
- LLM기반 다중 에이전트 시스템을 사용
 - KEA(Keyword Extraction Agent)를 기반으로 예제를 생성
 - EFA(Example Forming Agent)를 사용하여 KEA가 추출한 조건을 기반으로 예제를 생성한 뒤, Python 딕셔너리 형식으로 매핑
 - Data Mining Agent(DMA)를 통해 EFA가 생성한 예제를 학습하여 전체 데이터를 분석하고 실험 조건을 표준화된 형식으로 추출





멀티모달 기반 ADMET-DNN 모델 고도화

- ◆ 저분자 3D 구조 기반 ADMET 예측 **파운데이션 모델** 개발
 - E(n)-equivariant 기반 신경망과 Multi-head attention 메커니즘을 이용한 **EGNN** ADMET 예측 모델 개발
- ◆ **ADMET-DNN** 고도화
 - GCN, ChemBERTa 모델 고도화
 - EGNN 기반 예측 모델과 결합하여 양상블 예측 모델 개발

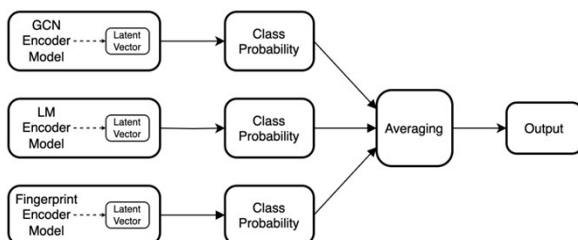
멀티모달 기반 ADMET-DOPCNN 모델 개발

- ◆ **증류기법** 기반 ADMET-DOPCNN 개발
 - GCN을 교사 모델, ADMET-OPCNN을 학생 모델로 증류 기법을 사용
 - ADMET-OPCNN 모델의 **과적합 문제 해결** 및 일반화 성능 향상
- ◆ ADMET-DOPCNN **고도화**
 - 분자의 3차원 구조를 이용한 EGCN, SMILES 문자열을 반영하는 언어모델등을 교사모델로 추가하여 DOPCNN을 고도화

▶ 멀티모달 기반 ADMET-DNN 모델 고도화

ADMET-DNN

선행연구



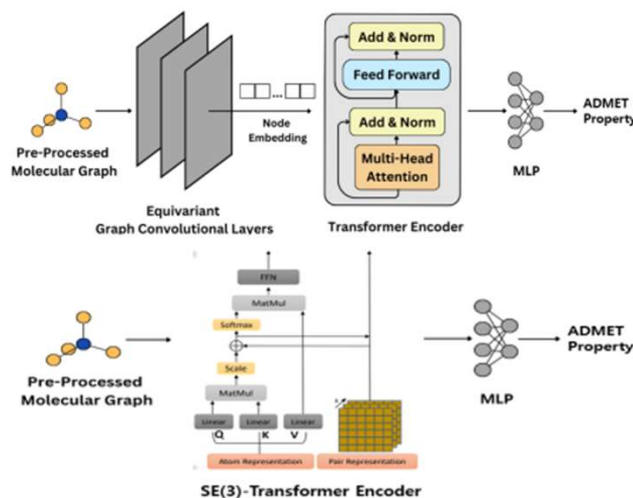
- 분자의 2차원 구조를 이용한 GCN, 1차원 SMILES string을 이용한 언어모델, ECFP 분자 지문 encoding을 이용한 앙상블 ADMET 예측 모델.
- 개별 모델 대비 앙상블 모델이 높은 성능을 보임

Model	AUC
GNN	0.813
ChemBERTa	0.806
FP-MLP	0.803
Ensemble	0.834

3D 구조 기반 ADMET 예측 모델 개발

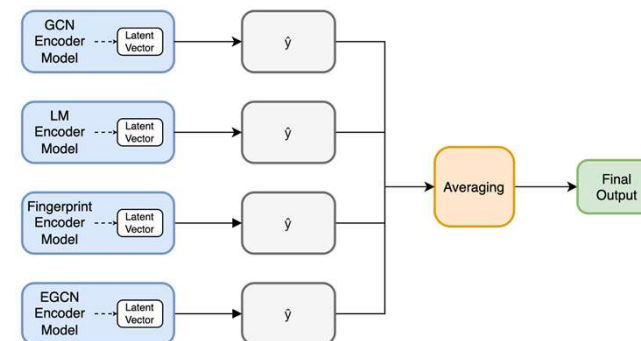
EGCN, SE(3)-equivariant Transformer를 이용

- E(n)-equivariant 기반 그래프 신경망 또는 SE(3)-equivariant Transformer를 사용하여 embedding vector를 만들고 MLP를 사용하여 ADMET를 예측
- 1차년도의 독성 예측을 시작으로 ADME와 PK등으로 예측 범위를 확대



ADMET-DNN 고도화

파운데이션 모델을 이용한 예측 모델 고도화



- 선행연구에서 개발되었던 GCN, ChemBERTa 기반 모델 고도화
- SE(3)-equivariant Transformer 또는 EGCN을 사용한 3차원 구조 기반 모델과 결합
- MLP 대신 파운데이션 모델을 이용 ADMET 예측

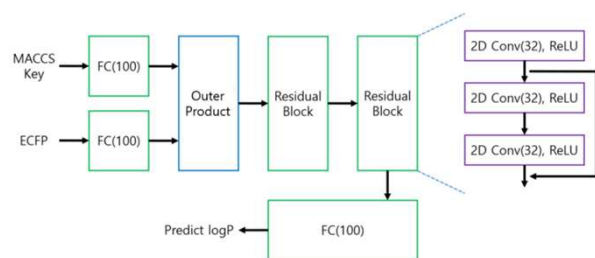
Classification		Regression	
AUPRC	MCC	ρ	R^2
> 0.92	> 0.92	> 0.85	> 0.85

▶ 증류기법을 이용한 ADMET-DOPCNN 모델 개발

ADMET-OPCNN

선행연구

- 두 종류의 분자지문을 사용하는 독성 예측 모델



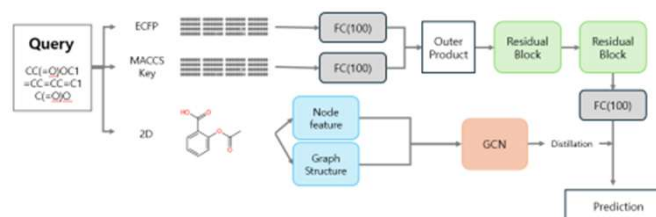
- DNN과의 logP 예측 성능을 비교한 결과 높은 성능을 나타냄

Metric Model		All Data		Approved Data	
		R2	RMSE	R ²	RMSE
vs DNN	ToxOPCNN	0.911	0.685	0.821	0.968
	DNN	0.899	0.729	0.816	0.982
vs LSTM	ToxOPCNN	0.911	0.685	0.821	0.968
	LSTM	0.885	0.776	0.743	1.161

ADMET-DOPCNN

증류 모델을 이용한 ADMET 예측 모델

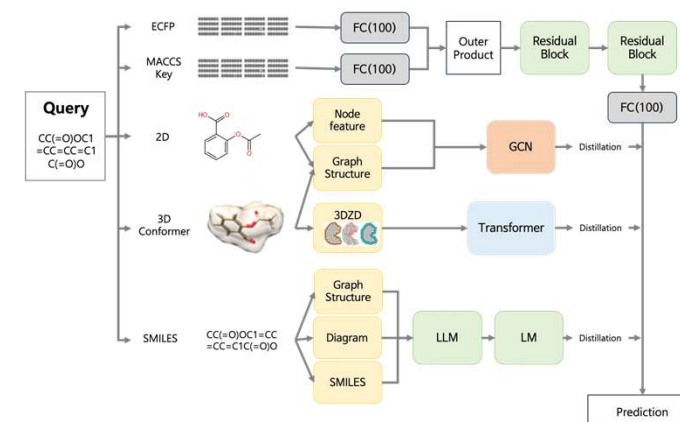
- 각 데이터에 해당하는 교사 모델의 지식을 학생 모델로 전달하여 성능을 높이는 증류 기법을 적용 (distillation)
- 훈련 데이터가 적을 때 과적합을 줄이는 효과
- 교사 모델: GCN, Transformer, LM
- 학생 모델: ADMET-OPCNN
- GCN은 분자의 구조적 특성(graph-based)을 학습하고, ADMET-OPCNN은 분자 지문 기반 특성을 학습



모델 고도화

교사 모델 확장

- EGCN, SE(3)-equivariant Transformer 등 3차원 구조 기반 ADMET 예측 모델과 3DZD 기반 ADMET 예측 모델을 교사 모델에 추가, ADMET-DOPCNN 고도화 수행



연구목표



불균형 데이터와 자원에 대응하며 지속적인 모델 업데이트가 가능한 연합학습 ADMET 예측 모델 개발

학습 모델 개인화 및 지속적인 모델 업데이트를 위한 연합학습 기술 개발

- ◆ 기관별 모델 개인화 및 서버의 **지속적 모델 업데이트 자동화**를 위한 연합학습 설계
 - 기관별 보유 데이터 및 선호 태스크 맞춤형 ADMET 예측 모델 개인화
 - 서버의 모델 업데이트와 기관별 개인화를 동시에 수행할 수 있는 모듈화 연합학습
- ◆ 지속적 모델 업데이트 연합학습 시나리오에 알맞은 **초기모델 사전 훈련 방법** 고안
 - 서버 업데이트 모듈과 클라이언트의 글로벌 모듈을 함께 사용 일반화 성능 개선
- ◆ Aim 2에서 개발한 ADMET 예측 모델에 대해 알고리즘 검증

기관별 불균형한 데이터, 자원, 기여도를 고려한 연합학습 기술 개발

- ◆ 기관별 **불균형 데이터와 자원에 대응**하는 LoRA 기반 개인화 연합학습
 - 개인화와 일반화 성능 간의 균형을 조절할 수 있는 파라미터 효율적인 연합학습
 - 서로 다른 자원에 호환되는 LoRA 개발
- ◆ 학습 기여도에 기반한 성능 **인센티브** 부여
 - 연합 학습 참여자 기여도 측정 방식 개발
 - 높은 기여도 참여자에 높은 개인화 성능을 제공하는 모델 종합 방식 개발
- ◆ Aim 1에서 개발한 ADMET 데이터셋의 불균형한 데이터 분포 환경 세팅

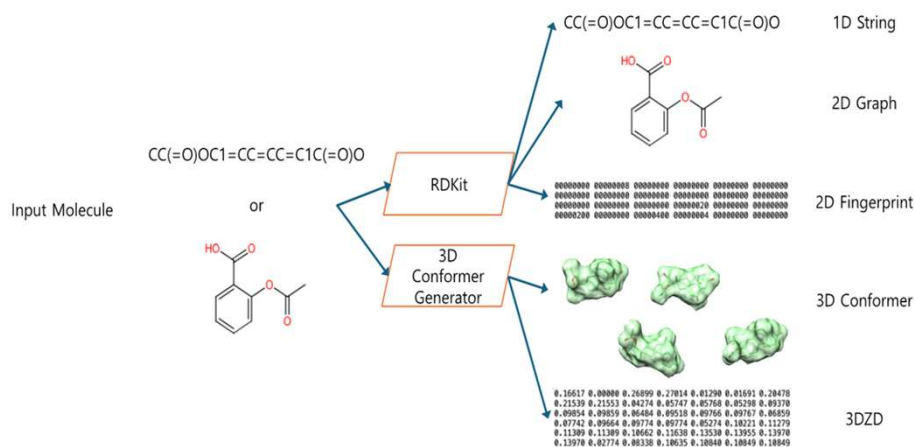
기관별 이질적 모달리티, 공격 취약성에 대응하는 연합학습 기술 개발

- ◆ 기관별 **이질적인 모달리티에 대응**하는 멀티모달 연합학습 기술 개발
 - 멀티모달 파운데이션을 위한 LoRA 고도화
 - 서로 다른 모달리티 데이터를 보유한 기관들에 호환될 수 있는 LoRA 기술개발
- ◆ **보안 취약성에 대응**하는 연합학습 개발
 - 기관별 이질적인 데이터 모달리티로 인해 서로 다른 파라미터를 업로드하는 경우의 secure aggregation 적용 방법 고안
- ◆ Aim 1과 2에서 개발한 ADMET 데이터셋과 예측 모델에 대해 개발 알고리즘 검증

선행 연구: 연합학습

ADMET 예측 연합학습

- 데이터: 약물 독성 데이터 (20,728)
- Input feature



- 모델 선정

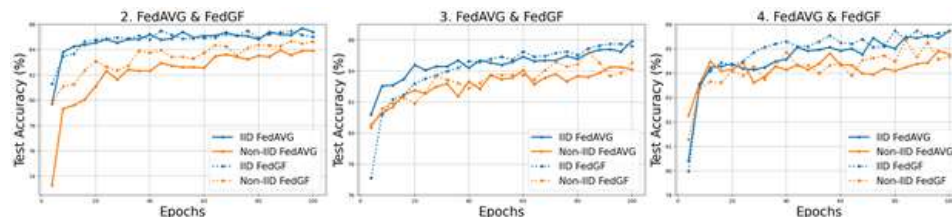
Model	Accuracy	Epoch
TabNet	0.840	34
FT-Transformer	0.833	17
CNN	0.844	13

- non-IID 환경 모사

1. Fingerprint를 이용한 분자간 distance 계산 후 balanced k-medoids clustering을 동일 개수의 샘플을 가지는 18개의 군집으로 나누고 군집별 데이터를 클라이언트에 분배
2. Standard k-medoids clustering을 활용하여 군집별 데이터를 클라이언트에 분배
3. Elbow method를 이용하여 군집 수를 결정 후 군집별 데이터를 클라이언트에 분배

- FedGF: 손실함수의 flat minima를 만드는 optimizer를 이용하여 글로벌 모델의 일반화 성능을 높임

- 연합학습 시뮬레이션 실험 결과

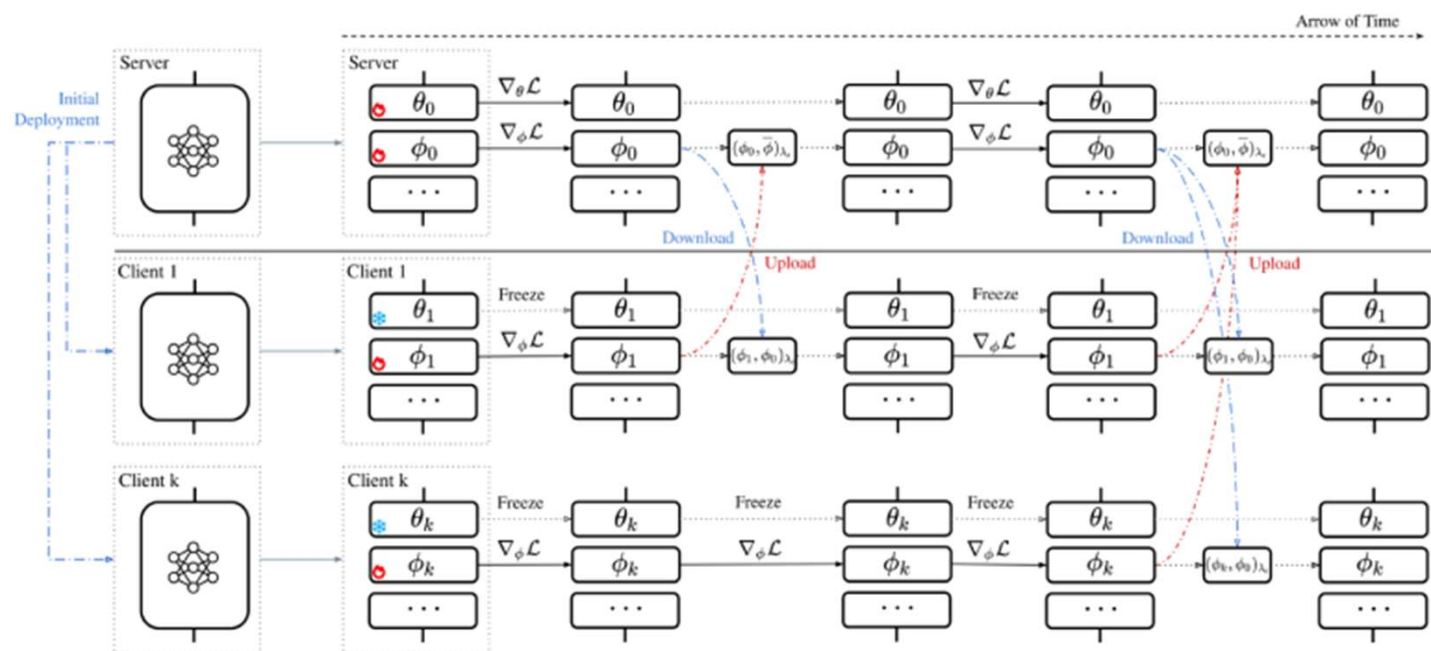


- 연합학습 시뮬레이션 실험 해석

FedGF 알고리즘이 non-IID 환경에서 FedAVG보다 **우수한 성능**을 보임

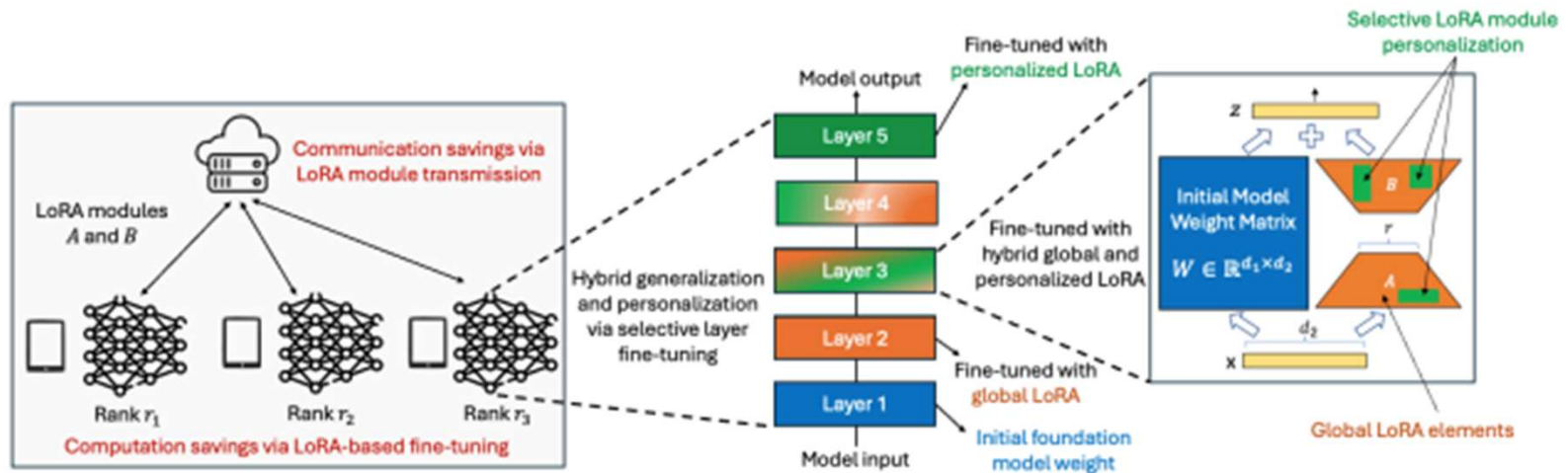
▶ 학습 모델 개인화 및 지속적인 모델 업데이트를 위한 연합학습 기술 개발

- 기관별 모델 개인화 및 서버의 지속적 모델 업데이트 자동화를 위한 연합학습 기술 개발
 - 기관별 보유 데이터 및 선호 태스크 맞춤형 모델 개인화를 위한 파라미터 효율적인 연합 학습 기술
 - 서버의 일반화 성능 개선을 위한 모델 업데이트와 기관별 개인화를 비동기적으로 동시에 수행할 수 있는 모듈화 연합 학습 기술
- 초기모델 사전 훈련 방법 고안
 - 기관별 모델 개인화와 서버의 지속적 모델 업데이트 자동화를 제공하는 연합 학습 시나리오에 알맞은 메타 학습 기반의 초기모델 사전 훈련 기법
 - 기관별 데이터 분포가 변화하는 환경에서 서버가 업데이트한 모델을 지속적으로 배포할 수 있는 환경 구축



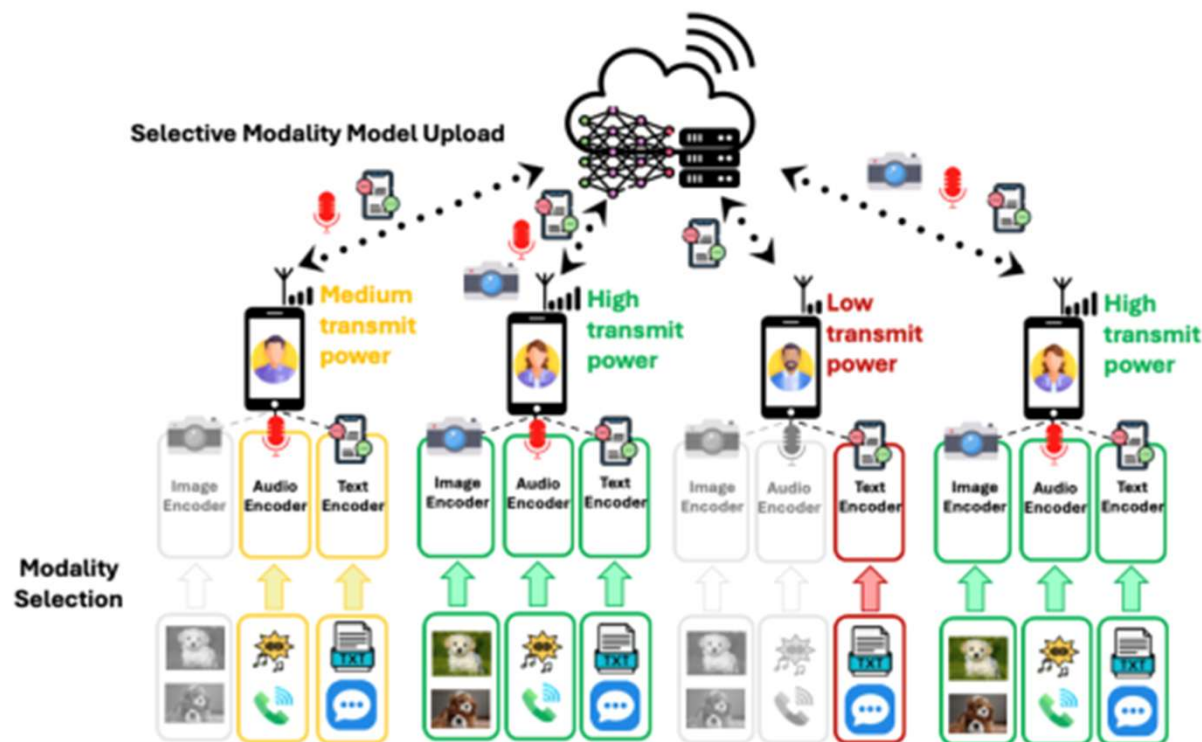
▶ 불균형 데이터, 자원, 기여도를 고려한 연합학습 기술 개발

- 기관별 불균형 데이터와 자원을 고려한 LoRA 기반 연합학습
 - 연산 및 통신자원에 따라 로컬 모델간 rank를 할당하여 전체 시스템의 자원 효율성과 학습 지연 시간의 최적화를 수행
 - 전체 모델이 아닌 특정 레이어에만 LoRA 기반의 개인화를 적용 일정 수준의 일반화 성능을 유지
- 학습 기여도에 따른 인센티브 부여
 - 로컬 모델과 글로벌 모델의 파라미터 비교등을 통하여 각 로컬 모델의 기여도를 측정
 - 기여도가 높은 기관의 개인화 성능에 가중치를 높게 부여하여 성능 측면에서 인센티브를 부여할 수 있는 모델 종합 방법 개발
- ADMET 예측 모델 탑재
 - ADMET-DNN과 ADMET-DOPCNN을 LoRA 기반 연합학습, 지속적인 모델 버전 업데이트 연합학습에 적용



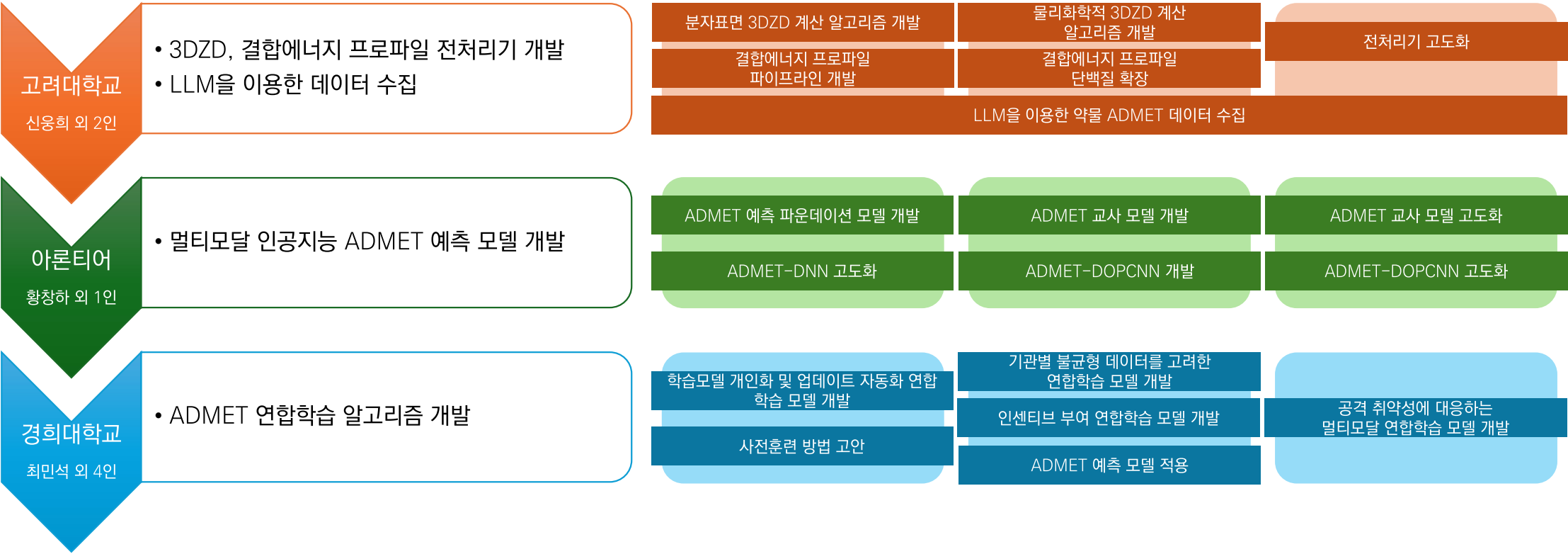
▶ 이질적 모달리티, 공격 취약성에 대응하는 멀티모달 딥러닝 연합학습 기술 개발

- 기관별 이질적 모달리티에 대응하는 멀티모달 연합학습 기술
 - 멀티모달 파운데이션 모델을 위한 LoRA 기술 고도화를 통한 사용자 맞춤형 모델 개발
 - 기관별 서로 다른 모달리티 데이터를 보유할 때 호환 될 수 있는 LoRA 파라미터 선택 방법 개발
- 공격 취약성에 대응하는 멀티모달 딥러닝 연합학습 기술 개발
 - 기관별 이질적인 데이터, 자원, 모달리티로 인해 서로 다른 파라미터를 업로드하는 경우의 secure aggregation 적용 방법
 - Gradient leakage problem에 대응하는 secure aggregation 방법 적용



연차별 추진 계획

PART4 • 추진 체계 및 연구진 구성



2025년도 연합학습 기반 신약개발 가속화 프로젝트
착수 보고회



감사합니다

연합학습기반 지속적인 멀티모달 앙상블 ADMET 예측 모델 개발

Development of a Continuous Multi-modal Ensemble
ADMET Prediction Model based on Federated Learning

