연구개발기관-아이젠사이언스 소개

AIGEN SCIENCES
Harnessing AI for Real-World Drug Discovery

㈜아이젠사이언스의 강점

세계적 수준의 의생명 AI 역량 입증	의생명 특화 자체 LLM 보유	AI 및 신약개발 전문 인력 구성	혁신적 내부 신약개발 파이프라인 구축
의생명 AI 국제 경연대회 7회 우승	의사면허시험 통과 최초 sLLM MeerKat	글로벌 제약사 기술이전 경험 보유 전문가	16개 파이프라인 개발
SOTA(최고 성능) 약물 표현 모델	최초 최다 인용 의생명 언어모델 BioBERT	도메인 지식 반영 실용적 AI 모델 설계	선도물질 최적화 이상 10개 파이프라인 보유

인공지능 전문가
강재우, CEO, 대표이사

- 고려대학교 컴퓨터학과 교수 (2006 ~ 현재)
- North Carolina State Univ. 컴퓨터학과 조교수 (2003 ~ 2006)
- AT&T Labs Research, Technical Staff (1996 ~ 1997)

KOREA UNIVERSITY · NC STATE UNIVERSITY · AT&T

신약개발 전문가
이광옥, CSO, 부사장

- 영진약품 연구본부장 (2017-2021)
- 부광약품 연구소장 (2015-2017)
- 한미약품 합성신약/항암임상연구 팀장 (1995-2015)

영진약품 · 부광약품 · Hanmi 한미약품

임상 이행연구 전문가
이호정, CBO, 상무이사

- 플랫바이오 CSO (2018-2023)
- 한미약품 연구센터 약리/이행연구 총괄이사 (2017-2018)
- Univ. of Texas, MD Anderson Cancer Center, Cancer Metastasis Center, Post-doctoral fellow (2012-2016)

PLATBIO · Hanmi 한미약품 · GC 녹십자

인공지능 전문가
김선규, AI 연구실장

- 고려대학교 연구교수
- 고려대학교 컴퓨터학 박사

KOREA UNIVERSITY

Biology 전문가
김지혜, 약리 연구실장

- SK 바이오팜 약리/약효 평가 (2017-2020)
- 서울대학교 생물학 박사

SK 바이오팜

인력 구성 - Highlights

인공지능 전문가:
글로벌 의생명 특화 인공지능 경연대회 7회 우승팀

국내 Top Tier 제약사 출신 연구진:
글로벌 제약사 기술 이전 경험의 의약화학, 약효 평가, 임상이행팀 구성

혁신적 내부 신약개발 파이프라인 및 세계적 수준의 의생명 AI 역량

AIGEN SCIENCES
Harnessing AI for Real-World Drug Discovery

- 최고 수준의 의생명 AI 기술력을 실제 신약개발에 적용하고 있으며 혁신적인 **16개 내부 파이프라인 보유 중임**
- 구글, 얀센, 스탠포드 등의 세계 유수의 연구팀들을 제치고 7번의 우승 → AI 기술력 입증

DUB	KRAS	합성 치사	저분자 백신	ADC 페이로드
신개념 TPD* 약물성 확보 용이	Pan-KRAS 약물 부재 Best in class	내성암 극복 필요성 Best in class	신개념 저분자 백신 치료용 백신	신규 기전 페이로드 이중 페이로드
• USP1 • USP28 • 신규 DUB	• SOS1 , SOS1 + Target A • Pan-KRAS • G12C on	• PARP + Target B • 2세대 PRMT5i • 신규 표적, RNA 표적	• Novel Target • RNA 표적	• 암세포 대사 표적 • 내성암 극복 표적

AstraZeneca-Sanger Dream Challenge (2)
약물-치료 병용 효과 예측
2016.03
- Stanford 6th
- MIT 11th

NCI-CPTAC Dream Challenge (1)
암 유전체 관련 단백질 발현 예측
2017.11
- UCLA 2nd
- MD Anderson 16th

Multi-targeting Drug Dream Challenge (1)
다중 표적 억제 약물 발굴 및 예측
2018. 12
- Janssen 4th

IDG Dream Challenge (1)
약물-표적 단백질 상호작용 예측
2019. 04
- UNC Chapel Hill Joint 1st
- UIUC Joint 1st

BioASQ Challenge Won 2yrs in a row (1, 1)
LLM기반 의생명 문헌 질의응답 시스템
2019, 2020
- Google 2nd

BioCreative VII NLM-Chem / DrugProt Track (1, 3)
LLM기반 의생명 문헌 지식 추출 시스템
2021. 11
- NVIDIA 3rd

RadSum Challenge (1)
LLM 기반 의료영상 진단추론 시스템
2023. 07
- Outperformed Stanford, UC London, Siemens

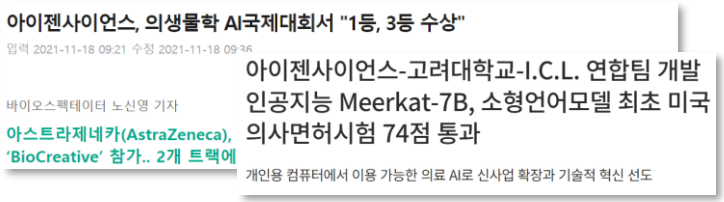
대규모 약물 표현 학습과 LLM 기반 문헌 마이닝을 활용한 연합학습 기반 ADMET 예측 모델 개발

AIGEN SCIENCES
Harnessing AI for Real-World Drug Discovery

Leveraging Point

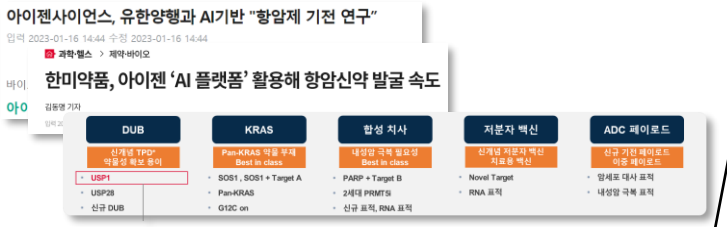
의생명 AI/ LLM 기술력

의생명 국제대회 7회 우승 및 최첨단 LLM 기술력 등 의생명/신약개발 AI 개발능력 보유



신약개발 전문회사

국내 Top-tier 신약개발 전문가 및 16개 내부 파이프라인 운영 등 신약개발 도메인 높은 이해도



㈜아이젠사이언스 R&D 포트폴리오

Our Approach

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

데이터 수급 및 효율 제고

- 문헌 기반 ADMET 데이터 추출
  - 의생명 특화 자체개발 LLM MeerKat
- 약물 표현 Foundation 모델
  - ADMET 예측 최고 성능 자체개발 모델 MolPLA

도메인 지식 반영

- 멀티태스크 학습
  - 신약개발 지식 기반 태스크 클러스터링
- 고도화된 데이터 전처리
  - 신약개발 지식 기반 데이터 보정

추진 전략

AIGEN SCIENCES
Harnessing AI for Real-World Drug Discovery

- 사업 목표를 달성하기 위해 아이젠사이언스의 강점을 바탕으로 사업 추진 전략을 설정함.

아이젠사이언스의 강점

- **글로벌 Top-tier 의생명 AI 역량 보유**
- **의생명 특화 자체 LLM 보유**
- **AI 및 신약개발 전문 인력 구성**
- **내부 신약개발 파이프라인 확보**

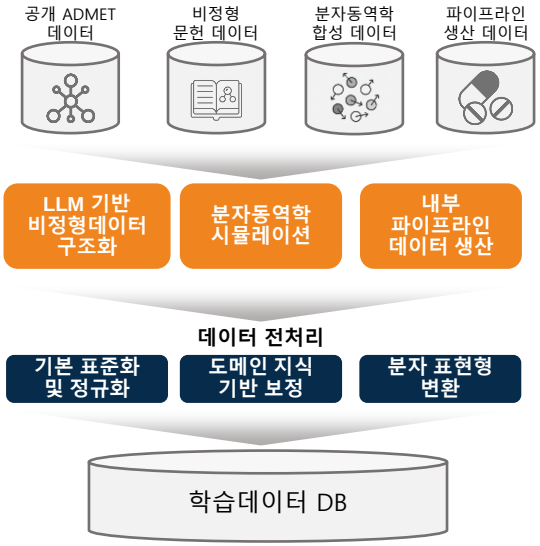
1. 학습 데이터 구축	2. 데이터 전처리 도구 개발	3. 약물 표현 AI 모델 구축	4. 연합학습 기반 ADMET AI 모델 개발
생명의료 특화 LLM 활용 문헌 정보 추출	신약개발 도메인 지식기반 신뢰도 높은 전처리 방식	사전 학습 모델 MolPLA	멀티-태스크 학습
공개 데이터베이스	실험 조건/측정 단위 표준화	오픈소스 사전 학습 (파운데이션) 모델	차등 프라이버시
내부 실험 데이터 확보	도메인 지식 기반 데이터 보정	Sequence/Graph/3D	연합 배치 정규화
생성AI/분자동역학 기반 데이터 합성	분자 표현 변환	메타 러닝	태스크 간 관계 모델링

학습데이터 구축

AIGEN SCIENCES
Harnessing AI for Real-World Drug Discovery

구축 방법론

공개된 데이터 뿐만 아니라 문헌 추출, 분자동역학, 내부 파이프라인을 통해 데이터 생산



대형언어모델(LLM) 활용 비정형 문헌 데이터 구조화

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

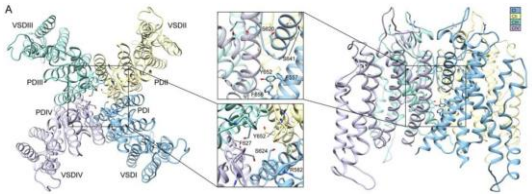
"In our study of novel kinase inhibitors, we evaluated the ADMET properties of three lead compounds: **XJ-5, XJ-7, and XJ-9**.
XJ-5 (4-(4-methylpiperazin-1-yl)-N-(4-(trifluoromethyl)phenyl)pyridine-2-amine) demonstrated good oral bioavailability in rats (F = 68%) and moderate plasma protein binding (87%). 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-yl)amino)pyrimidin-4-yl)amino)phenyl)acrylamide	2-((1H-indazol-5-yl)amino)-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

[1] Liang, Yanshu, et al. "Predicting Blood–Brain Barrier Permeation of Erlotinib and JCN037 by Molecular Simulation." *The Journal of Membrane Biology* 256.2 (2023): 147-157.
[2] Maly, Jan, et al. "Structural modeling of the hERG potassium channel and associated drug interactions." Frontiers in Pharmacology 13 (2022): 966463.

분자동역학 데이터 합성

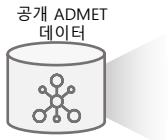
분자동역학 및 자유에너지 계산 등으로
추출 가능한 ADMET 데이터 합성



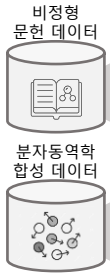
Structure modeling 기반 hERG Channel inhibition 예측 [2]

데이터 수집 대상

국내외 공개 DB, 다양한 논문 및 특허문헌 등 다양한 출처의 데이터



데이터 소스	대략적인 데이터 규모
PubChem	1억 1천만 개 이상의 화합물
ZINC	약 2억 3천만 개의 화합물
DrugBank	약 14,000개의 약물 정보
BindingDB	약 200만 개의 결합 친화도 데이터
MoleculeNet	약 70만 개 화합물, 17개 데이터셋
TDC	66개 데이터셋, ADMET 관련 22개
ChEMBL	200만 개 이상의 화합물
ADME@ NCATS	약 10만 개의 화합물, 다양한 ADMET 속성
Tox21	약 10,000개 화합물, 12개 독성 엔드포인트
ToxCast	약 9,000개 화합물, 1,000개 이상의 독성



- 학술 문헌 데이터
- 특허 문헌 데이터 (WIPO, USPTO 등)
- 임상 시험 보고서 데이터
- 단백질 구조–약물 결합 데이터
- 분자동역학 시뮬레이션 데이터
- 자유에너지 perturbation 데이터

데이터 전처리 도구 개발

AIGEN SCIENCES
Harnessing AI for Real-World Drug Discovery

- 다양한 소스에서 추출한 데이터를 통합하기 위한 전처리 도구 개발
- 단순 통계를 넘어 신약개발 도메인 지식을 토대로 보정함으로써 데이터 품질 향상

표준화 및 정규화

기본 보정을 통한 데이터 일관성 확보와 품질 개선

보정 항목	보정 방법
Unit conversion	• 농도 단위 변환 (예: μg/mL → mol/L) • 온도 단위 변환 (예: °F → °C)
Data normalization	• Min-Max / • Z-score
Outlier Detection and Removal	• IQR (Interquartile Range) / • Z-score 방법
Batch Effect Correction	• 배치 간 평균 조정 / • ComBat 방법
Threshold Value Handling	• LOQ (정량한계) 이하 값의 1/2 LOQ 대체 • LOD (검출한계) 이하 값의 0 대체

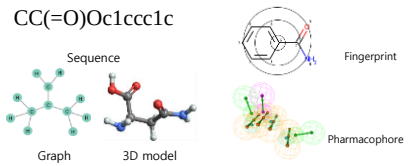
도메인 지식 기반 보정을 통한 예측력 향상

보정 항목	보정 방법 (1st principle & Data-driven)
pH / 온도에 따른 용해도	• Henderson-Hasselbalch / van't Hoff 방정식 • pH-용해도 관계를 학습한 신경망 모델
대사 안정성 종간 차이	• 알로메트릭 스케일링 • 종간 대사 안정성 차이를 학습한 트랜스포머 기반 모델
막 투과성 (세포주 간 차이)	• Fick의 확산 법칙 기반 계산 • 다양한 세포주 간 투과성 관계를 학습한 신경망 모델
LogP/LogD (실험 조건 차이)	• Hansch-Leo 방법 또는 원자 기여도 방법 • 다양한 실험 조건에서의 LogP/LogD 값을 학습한 그래디언트 부스팅 모델
pKa (이온화 상태 영향)	• Hammett 방정식 기반 계산 • 분자 구조와 pKa 관계를 학습한 딥러닝 모델

분자 표현형 변환

1st principal 기반 물질 표현 활용

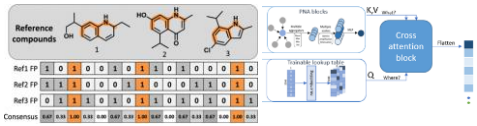
1st



구조 기반 표현 | 특징 기반 표현

데이터 기반 물질 표현

2nd



유사성 기반 표현 | 사전학습모델 기반 표현

연차별 전처리도구 개발계획

1년차 기본 전처리 도구를 통해 모델 개발 이후 성능 향상을 위한 전처리 집중

1차년도

기본 전처리 도구 개발

- ADMET 데이터 형식 표준화 프로토콜 수립
- 데이터 정규화 기법 구현
- 기본 분자 표현 변환 기능 구현

2차년도

고급기능 추가 및 성능 최적화

- 도메인 지식 기반 보정 기법 개선
- 대용량 데이터 처리를 위한 성능 최적화
- 기본 사용자 매뉴얼 초안 작성

3차년도

자동화 및 통합 시스템 구축

- 연합학습 플랫폼과의 통합
- 머신러닝 기반 자동 클리닝 및 보정

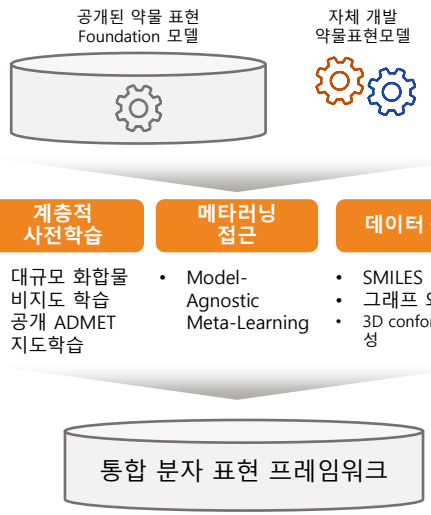
약물 표현 파운데이션 AI 모델 구축

AIGEN SCIENCES
Harnessing AI for Real-World Drug Discovery

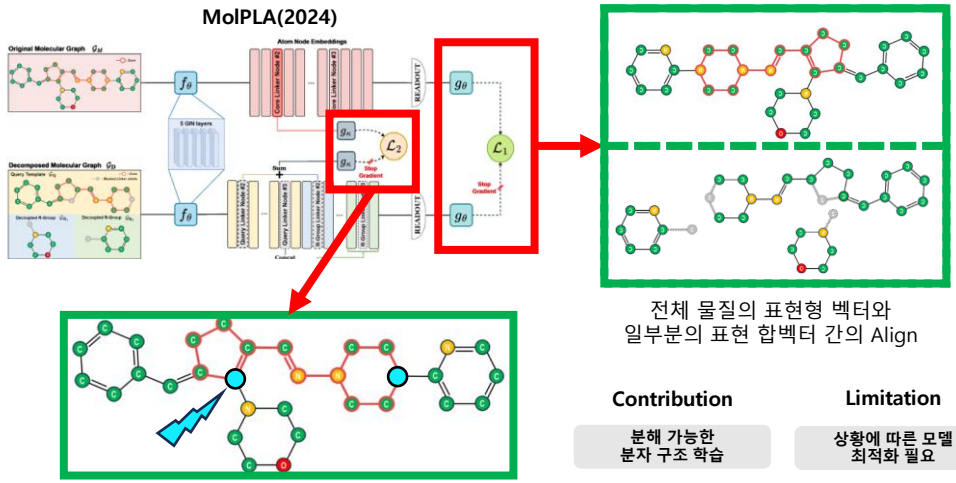
- 최고 성능 수준의 자체개발 약물 표현 모델과 오픈소스 모델 같이 운용

구축 방법론

공개 및 자체개발 약물표현 모델을 토대로
통합 분자 표현 프레임워크 구축



자체 개발 약물 표현 Foundation 모델 MolPLA



Contribution	Limitation
분해 가능한 분자 구조 학습	상황에 따른 모델 최적화 필요
자기지도학습 가능한 사전학습	사전학습 데이터 부족

Gim, Mogan, et al. "MolPLA: a molecular pretraining framework for learning cores, R-groups and their linker joints." *Bioinformatics* 40.Supplement_1 (2024): i369-i380.

오픈소스 표현모델 수집 대상

국내외 공개된 문헌을 기반으로다양한 데이터구조의 표현모델 수집

그래프 기반
약물 표현 모델

모델 이름	발표 연도	발표된 저널(학회)
MoMu	2023	Nat. Mach. Intell.
GeoGNN	2022	Nat. Mach. Intell.
MolCLR	2021	NeurIPS
Graphormer	2021	NeurIPS
GROVER	2020	NeurIPS

3D 기반
약물 표현 모델

모델 이름	발표 연도	발표된 저널(학회)
Uni-Mol V2	2023	ArXiv
GeoMol	2021	NeurIPS
3D Infomax	2021	ArXiv
DimeNet++	2020	ICML
SchNet	2018	J. Chem. Phys.

시퀀스 기반
약물 표현 모델

모델 이름	발표 연도	발표된 저널(학회)
MolFormer	2023	NeurIPS
SELFormer	2023	ArXiv
MolGPT	2022	JCIM
Regression Transformer	2022	Nat. Mach. Intell.
ChemBERTa	2020	ArXiv

연합학습 기반 ADMET AI 모델 개발

AIGEN SCIENCES
Harnessing AI for Real-World Drug Discovery

- 도메인(Prior) 지식을 반영한 태스크 관계 모델링으로 데이터 부족 현상 극복 및 모델 신뢰도 제고
- 로컬 모델과 다른 연합학습 환경에서 최적화할 수 있는 기술 확보

구축방법론

다양한 멀티태스크 학습과 연합학습 기술을 적용해 최적의 성능 모델 확보

멀티태스크 학습

- Hard/Soft 공유표현학습
- 태스크 간 관계 모델링
- 불균형 데이터 처리

연합학습 최적화

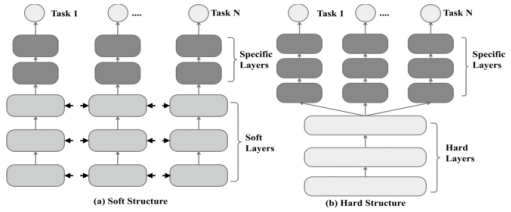
- 차등 프라이버시
- 데이터참여자 기여도 평가
- 보안 기반 성능 모니터링

연합학습 기반
ADMET 예측 AI 모델

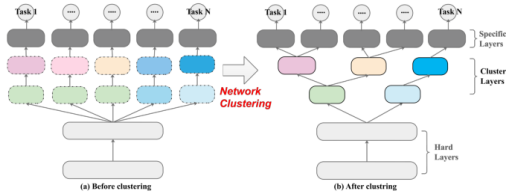
Gao, Dehong, et al. "Network clustering for multi-task learning." *arXiv preprint arXiv:2101.09018* (2021).

멀티태스크 학습

다양한 멀티태스크 학습 기술 적용을 통해 여러 ADMET 속성에 효과적인 모델 구축



Soft/Hard 공유표현 예시



태스크 클러스터링을 통한 성능 제고

보안을 고려한 연합학습 최적화

모델 사이드에서 연합학습의 데이터 유출을 방지하고 최적화된 모델 구축 방법 확보

학습 데이터 유출 방지를 위한 Model-side 기술

기술	주요 내용
Secure Batch Normalization	• Local 배치의 statistics에 noise 추가 • ε-differential privacy 보장
Secure Gradient Update	• Local gradient 에 noise 추가 • Gradient clipping • Encrypted gradient 전송
Secure Model Aggregation	• Encrypted model parameter 집계 • Aggregation 결과에 noise 추가 • Homomorphic encryption 활용

연합학습 최적화를 위한 Model-side 기술

기술	주요 내용
Communication Efficiency	• Gradient compression • Sparse update
Participant Contribution Evaluation	• Secure contribution measurement • Incentive mechanism 설계
Performance Monitoring	• Secure validation set 평가 • Privacy-utility trade-off 분석
Continuous Learning	• Incremental learning 방법 개발 • Model versioning