연합학습 기반

신약개발 가속화를 위한

데이터 공급 및 FDD/FAM 활성화 지원

연구책임자 – 심플렉스 대표 조성진

CIMPLRX

SIMPLIFY DRUG DISCOVERY

연합학습 기반 신약개발 가속화 프로젝트

(RFP 2-1) 데이터 활용 신약개발 협력 체계 구축

최종 목표

1. 연합학습 기반 신약개발 가속화를 위한 **데이터 공급**
2. **FDD/FAM 활성화**를 위하여 FDD 플랫폼 구축 기관/FAM 개발 기관과 협력

연구 내용

1단계 (2024~2026): 데이터 준비 및 공급 체계 구축
- ADMET 및 PK 파라미터 보유 데이터 현황 분석서 1건 작성
- 데이터 전처리를 위한 태스크 정의서 8건 작성
- 데이터 전처리 계획에 따른 기본 데이터 공급: 1,052건 공급 (527 (보유) + 525 (추가))
- CRO 활용 평가를 통한 신규 데이터 787건 매년 생성 및 공급 (2025~2026)
- 보유 데이터 유형 및 특성을 기술한 데이터 공급 계획서 8건 작성
- 연합학습 데이터 공급 보고서 (매년 1건, 공급 데이터량 및 데이터 품질 기준 명시)
- 로컬 학습 및 모델 개발/성능 분석

2단계 (2027~2028): 데이터 공급 지속성 확보 및 FDD 플랫폼 및 FAM 활용 검증
- 지속 데이터 공급 계획서 8건
- CRO 활용 평가를 통한 신규 데이터 787건 매년 생성 및 공급
- FDD 플랫폼 및 FAM 활용 보고서 1건
- FAM 예측 성능 분석서 1건
- 추가 데이터 활용하여 로컬 학습 및 모델 개발
- 비연합학습/연합학습 결과 비교 및 성능 분석

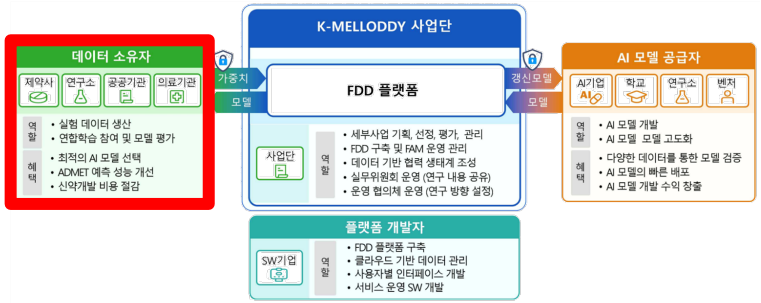


그림 3. 세부 사업구성 및 주요 역할

2024년도 제1차
연합학습 기반 신약개발 가속화 프로젝트
사업설명회 자료집

연구 내용 : 공급계획

매년 공급될 데이터의 양

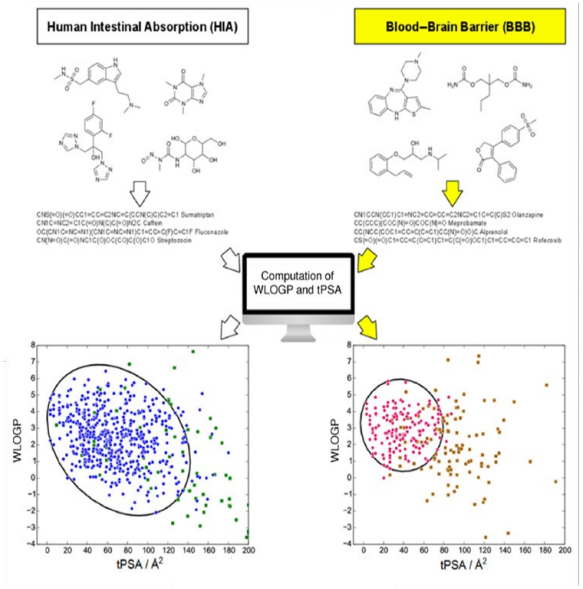
	유형	1년차	2년차	3년차	4년차	5년차	총계	%
PAMPA-BBB	A	310	500	500	500	500	2,310	55.0
Caco-2 Permeability	A	3	4	4	4	4	19	0.45
Plasma protein binding	D	15	10	10	10	10	55	1.31
Microsomal stability	M,E	29	50	50	50	50	229	5.45
Normal cell cytotoxicity	T	552	108	108	108	108	984	23.43
hERG	T	70	50	50	50	50	270	6.43
CYP inhibition	T	18	20	20	20	20	98	2.33
In vivo PK	PK	55	45	45	45	45	235	5.60
total		1,052	787	787	787	787	4,200	100

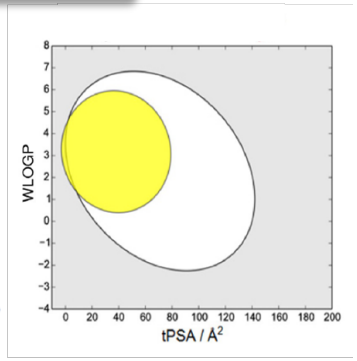
- 1차년도에는 총 1,052개 데이터 공급 예정 (당사 보유 데이터 527개 + 추가 확보 예정 525개)
- 본 과제 수행 기간 동안 총 4,200개의 데이터를 공급할 계획

연구 내용 : PAMPA-BBB & BOILED-Egg 모델

CNS focused/모델 개발 특화된 라이브러리 (약 3천 종)으로 PAMPA-BBB 모델 고도화

BOILED-Egg 모델

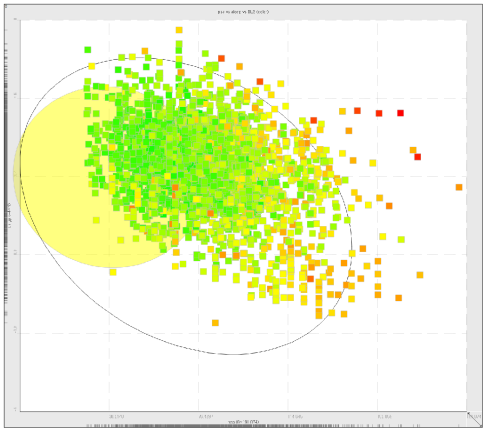




- **하얀색 영역 (흰자)**
  위장관에서 흡수될 확률이 가장 높은 분자들의 물리화학적 공간
- **노란색 영역 (노른자)**
  뇌로 침투될 확률이 가장 높은 분자들의 물리화학적 공간

ChemMedChem. 2016, 11, 1117.

약물 유사성(DLS) 모델 결과와 유사



- 3,000여 종의 당사 물질을 BOILED-Egg 모델로 분석한 결과
- 화합물은
  - 녹색 (약물 유사성 높음)과
  - 빨간색 (약물 유사성 낮음)으로 표시

연구내용 : 정상 세포 독성

고품질 데이터로 정상 세포 독성 예측 모델 고도화

	유형	1년차	2년차	3년차	4년차	5년차	총계	%
PAMPA-BBB	A	310	500	500	500	500	2,310	55.0
Caco-2 Permeability	A	3	4	4	4	4	19	0.45
Plasma protein binding	D	15	10	10	10	10	55	1.31
Microsomal stability	M,E	29	50	50	50	50	229	5.45
Normal cell cytotoxicity	T	552	108	108	108	108	984	23.43
hERG	T	70	50	50	50	50	270	6.43
CYP inhibition	T	18	20	20	20	20	98	2.33
In vivo PK	PK	55	45	45	45	45	235	5.60
total		1,052	787	787	787	787	4,200	100

Normal cell toxicity 비율 (23.43%)이 높은 이유는,

NCATS에서 공개한 약 5,200종의 물질로 생성된 고품질 데이터를 보완하여, 개선된 연합학습 모델을 개발하기 위함

Predictive Models For Estimating Cytotoxicity On The Basis Of Chemical Structures

Hongmao Sun*, Yuhong Wang, Dorian M. Cheff, Matthew D. Hall, Min Shen*
National Center for Advancing Translational Sciences (NCATS), 9800 Medical Center Dr. Rockville, MD 20850

Abstract

Cytotoxicity is a critical property in determining the fate of a small molecule in the drug discovery pipeline. Cytotoxic compounds are identified and triaged in both target-based and cell-based phenotypic approaches due to their off-target toxicity or on-target and on-mechanism toxicity for oncology and neurodegenerative targets. It is critical that chemical-induced cytotoxicity be reliably predicted before drug candidates advance to the late stage of development, or more ideally, before compounds are synthesized. In this study, we assessed the cell-based cytotoxicity of nearly 10,000 compounds in NCATS annotated libraries against four 'normal' cell lines (HEK 293, NIH 3T3, CRL-7250 and HaCat) using CellTiter-Glo (CTG) technology and constructed highly predictive models to estimate cytotoxicity from chemical structures. There are 5,241 non-redundant compounds having unambiguous activities in the four different cell lines, among which 11.8% compounds exhibited cytotoxicity in two or more cell lines and are thus labelled cytotoxic. The support vector classification (SVC) models trained with 80% randomly selected molecules achieved the area under the receiver operating characteristic curve (AUC-ROC) of 0.88 on average for the remaining 20% compounds in the test sets in 10 repeating experiments. Application of under-sampling rebalancing method further improved the averaged AUC-ROC to 0.90. Analysis of structural features shared by cytotoxic compounds may offer medicinal chemists heuristic design ideas to eliminate undesirable cytotoxicity. The profiling of cytotoxicity of drug-like molecules with annotated primary mechanism of action (MOA) will inform on the roles played by different targets or pathways in cellular viability. The predictive models for cytotoxicity (accessible at https://tripod.nih.gov/web_adme/cytotox.html) provide the scientific community a fast yet reliable way to prioritize molecules with little or no cytotoxicity for downstream development.

정상 세포 독성 공개 데이터
(Bioorg Med Chem. 2020, 28, 10.)

- 미국 정부기관인 NCATS에서 공개
- 동일 실험 조건으로 생성된 데이터
- 예측력이 높은 모델 개발에 큰 기여

- NCATS에서 공개한 데이터에 당사에서 제공할 데이터를 추가하여 모델을 만든다면 고품질 데이터를 사용한 모델 생성에 기여할 수 있음

동일한 분석 조건을 사용하여 생성된 고품질 데이터

	Normal Cell Lines			
	CRL-7250	NIH 3T3	HEK 298	HaCat
Cytotoxic compounds	558	559	733	570
Out of 5241 (%)	10.6	10.7	14.0	10.9

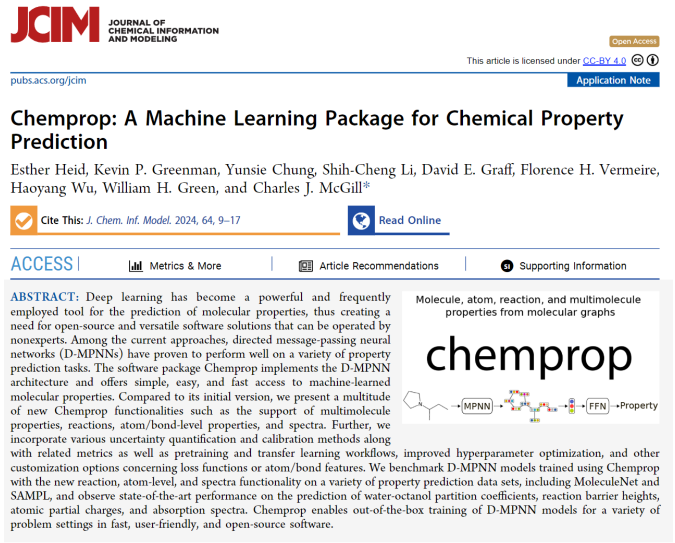
고품질 데이터로 고품질 모델 생성 가능

Method	Split Method		N	R^2	AUROC
RF	All		5,241	0.810	0.995
	Random	Train	4,192	0.805	0.995
		Test	1,049	0.478	0.910
	Sphere Exclusion	Train	4,192	0.823	0.996
		Test	1,049	0.274	0.836

연구내용 : 공개된 모델로 지표 생성

Chemprop 모델이나 Deep-PK로 결과를 비교하여 지표 생성

- Original Chemprop 모델과 Deep-PK로 예측결과를 비교할 계획



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Open Access

This article is licensed under CC-BY 4.0

pubs.acs.org/jcim Application Note

Chemprop: A Machine Learning Package for Chemical Property Prediction

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill*

Cite This: *J. Chem. Inf. Model.* 2024, 64, 9−17 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

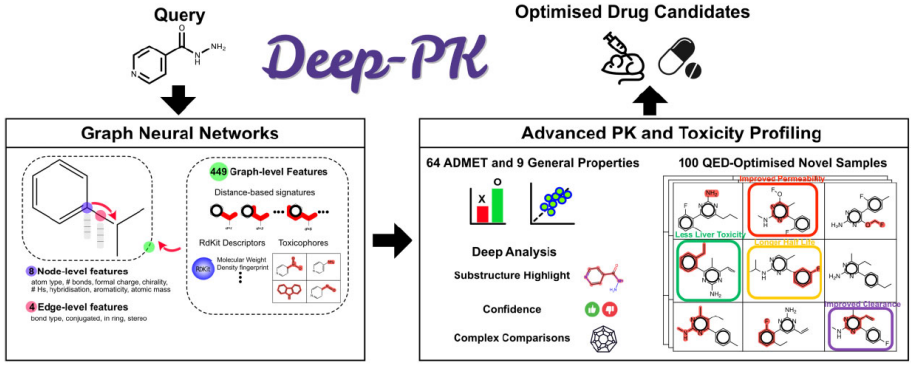
ABSTRACT: Deep learning has become a powerful and frequently employed tool for the prediction of molecular properties, thus creating a need for open-source and versatile software solutions that can be operated by nonexperts. Among the current approaches, directed message-passing neural networks (D-MPNNs) have proven to perform well on a variety of property prediction tasks. The software package Chemprop implements the D-MPNN architecture and offers simple, easy, and fast access to machine-learned molecular properties. Compared to its initial version, we present a multitude of new Chemprop functionalities such as the support of multimolecule properties, reactions, atom/bond-level properties, and spectra. Further, we incorporate various uncertainty quantification and calibration methods along with related metrics as well as pretraining and transfer learning workflows, improved hyperparameter optimization, and other customization options concerning loss functions or atom/bond features. We benchmark D-MPNN models trained using Chemprop with the new reaction, atom-level, and spectra functionality on a variety of property prediction data sets, including MoleculeNet and SAMPL, and observe state-of-the-art performance on the prediction of water-octanol partition coefficients, reaction barrier heights, atomic partial charges, and absorption spectra. Chemprop enables out-of-the-box training of D-MPNN models for a variety of problem settings in fast, user-friendly, and open-source software.

Molecule, atom, reaction, and multimolecule properties from molecular graphs

chemprop

MPNN → FFN → Property

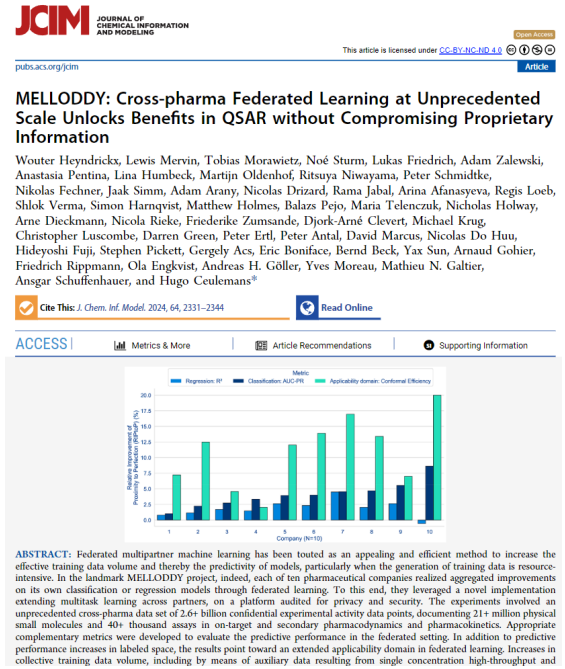
J. Chem. Inf. Model. 2024, 64, 1, 9.
github.com/chemprop/chemprop



Nucleic Acids Research. 2024, gkae254.
https://biosig.lab.uq.edu.au/deeppk/

연구내용 : 지표 정의

EU-MELLODDY 사업에서 개발한 RIPtoP 포뮬러 활용



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information

Wouter Heyndrickx, Lewis Mervin, Tobias Morawietz, Noé Sturm, Lukas Friedrich, Adam Zalewski, Anastasia Pentina, Lina Humbeck, Martijn Oldenhof, Ritsuya Niwayama, Peter Schmidtke, Nikolas Fechner, Jaak Simm, Adam Arany, Nicolas Drizard, Rama Jabal, Arina Afanasyeva, Regis Loeb, Shlok Verma, Simon Harnqvist, Matthew Holmes, Balazs Pejo, Maria Telenczuk, Nicholas Holway, Arne Dieckmann, Nicola Rieke, Friederike Zumsande, Djork-Arné Clevert, Michael Krug, Christopher Luscombe, Darren Green, Peter Ertl, Peter Antal, David Marcus, Nicolas Do Huu, Hideyoshi Fuji, Stephen Pickett, Gergely Acs, Eric Boniface, Bernd Beck, Yax Sun, Arnaud Gohier, Friedrich Rippmann, Ola Engkvist, Andreas H. Göller, Yves Moreau, Mathieu N. Galtier, Ansgar Schuffenhauer, and Hugo Ceulemans*

Cite This: J. Chem. Inf. Model. 2024, 64, 2331−2344 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Federated multipartner machine learning has been touted as an appealing and efficient method to increase the effective training data volume and thereby the predictivity of models, particularly when the generation of training data is resource-intensive. In the landmark MELLODDY project, indeed, each of ten pharmaceutical companies realized aggregated improvements on its own classification or regression models through federated learning. To this end, they leveraged a novel implementation extending multitask learning across partners, on a platform audited for privacy and security. The experiments involved an unprecedented cross-pharma data set of 2.6+ billion confidential experimental activity data points, documenting 21+ million physical small molecules and 40+ thousand assays in on-target and secondary pharmacodynamics and pharmacokinetics. Appropriate complementary metrics were developed to evaluate the predictive performance in the federated setting. In addition to predictive performance increases in labeled space, the results point toward an extended applicability domain in federated learning. Increases in collective training data volume, including by means of auxiliary data resulting from single concentration high-throughput and

J. Chem. Inf. Model. 2024, 64, 7, 2331.

$$\mathrm{RIPtoP}(\mathrm{metric}) = \frac{\mathrm{metric}_{\mathrm{MoI}} - \mathrm{metric}_{\mathrm{baseline}}}{\mathrm{metric}_{\mathrm{perfect}} - \mathrm{metric}_{\mathrm{baseline}}}$$

$$\mathrm{RIPtoP1}(\mathrm{metric}) = (\mathrm{metric1}_{\mathrm{MoI}} - \mathrm{metric}_{\mathrm{baseline}}) \,/\, (\mathrm{metric}_{\mathrm{perfect}} - \mathrm{metric}_{\mathrm{baseline}})$$

$$\mathrm{RIPtoP2}(\mathrm{metric}) = (\mathrm{metric2}_{\mathrm{MoI}} - \mathrm{metric}_{\mathrm{baseline}}) \,/\, (\mathrm{metric}_{\mathrm{perfect}} - \mathrm{metric}_{\mathrm{baseline}})$$

$$\mathrm{RIPtoP3}(\mathrm{metric}) = (\mathrm{metric3}_{\mathrm{MoI}} - \mathrm{metric}_{\mathrm{baseline}}) \,/\, (\mathrm{metric}_{\mathrm{perfect}} - \mathrm{metric}_{\mathrm{baseline}})$$

$$\mathrm{RIPtoP4}(\mathrm{metric}) = (\mathrm{metric4}_{\mathrm{MoI}} - \mathrm{metric}_{\mathrm{baseline}}) \,/\, (\mathrm{metric}_{\mathrm{perfect}} - \mathrm{metric}_{\mathrm{baseline}})$$

$\mathrm{metric}_{\mathrm{perfect}}$는 AUC-ROC이나 R^2라면 1 이며, RMSE라면 0.

$\mathrm{metric}_{\mathrm{baseline}}$는 공개 (ChEMBL) 데이터로만 비연합학습한 모델.

$\mathrm{metric1}_{\mathrm{MoI}}$는 모든 기관에서 보유하고 있는 비공개 데이터로 연합학습한 모델.

$\mathrm{metric2}_{\mathrm{MoI}}$는 모든 기관에서 보유하고 있는 비공개 데이터와 공개 (ChEMBL) 데이터로 연합학습한 모델.

$\mathrm{metric3}_{\mathrm{MoI}}$는 당사가 보유하고 있는 비공개 데이터로 비연합학습한 모델.

$\mathrm{metric4}_{\mathrm{MoI}}$는 당사가 보유하고 있는 비공개 데이터와 공개 (ChEMBL) 데이터로 비연합학습한 모델.

연구 개발 역량 : 연구책임자

신약개발, 데이터 생산, 인공지능 모델 • 플랫폼 개발 경험

자체 파이프라인 발굴

- 신경병증성 통증치료체 후보물질 발굴 (신규 기전, 신규 모핵, 특허출원 완료)
- 자가면역질환치료제 유효물질 발굴
- 항암제 유효물질 발굴

인공지능 기반 활발한 신약개발 공동연구 연구 경험

- 동아ST 신약 공동연구개발 (퇴행성 뇌질환)
- SK케미칼 신약 공동연구개발 (항암제/호흡기 질환)
- 신풍제약 신약 공동연구개발 (심부전)
- 동화약품 신약 공동연구개발 (면역질환)
- 삼진제약 신약 공동연구개발 (항암제)
- 셀트리온제약 신약 공동연구개발
- 공동연구개발에서 도출한 신규 화합물 물질 특허 4건 출원 (2023년) 완료

정부 연구과제 수행 경험

- 2019~2021 인공지능 신약개발 플랫폼 (구축과제, 세부 연구책임자, 과기정통부)
- 2019~2021 인공지능 신약개발 플랫폼 (운영과제, 연구원, 과기정통부)
- 2022~2023 인공지능 활용 혁신 신약 발굴 사업 (과제 총괄 책임자, 과기정통부)
- 2023~2025 국가신약개발사업_신약 R&D 생태계 구축 연구 (세부 연구책임자, 국가신약개발사업단)
- 2024~2028 식품의약품안전처 출연연구개발사업 (세부 연구책임자, 식품의약품안전처)

심플렉스㈜ 조성진 대표

- 노스캐롤라이나 약학대학교 의약화학 박사
- GSK, BMS, Amgen, CHDI 등 세계적인 신약 개발 제약회사에서 25년 이상 신약 개발에 필요한 플랫폼 개발 경력
- 의약화학 기반 XAI 플랫폼 **CEEK-CURE** 개발
- AI startup 100 선정 (KT경제경영연구소/한경 AI경제연구소, 2023)

연구 개발 추진 체계

데이터 공급 및 FDD 플랫폼/FDD 활성화

전문가 자문

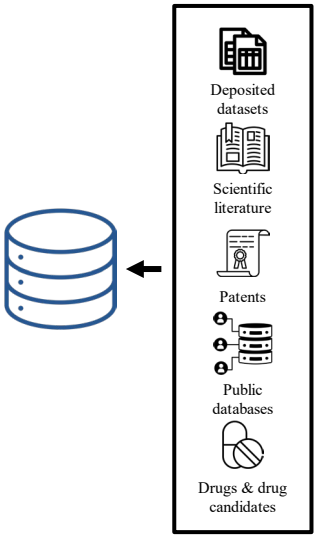
Dr. Yax Sun
- EU-MELLODDY 참여 (Amgen 대표 연구자)
- Molecular Structure 전문가

주관 연구책임자

CIMPLRX SIMPLIFY DRUG DISCOVERY

- 데이터 준비 및 공급체계 구축
- 로컬 학습 및 모델 개발/성능 분석
- 데이터 공급 지속성 확보
- 데이터 전처리 계획에 따른 기본 모델에 사용될 데이터 공급
- FDD 플랫폼/FAM 성능 분석

CIMPLRX	Wuxi	Sundia	Aragen	Eurofins
Normal Cell Cytotoxicity	PAMPA-BBB Caco-2 permeability (A)	Microsomal Stability CYP inhibition (M, E, T)	Plasma protein binding In vivo PK (D, PK)	hERG (T)



- 심플렉스의 내부 데이터와 CRO 활용 평가를 통한 지속적인 고품질 데이터 공급
- 인공지능 활용 신약개발 경험을 바탕으로 한 데이터에 대한 높은 이해도와 FDD 플랫폼/FAM 활용 및 검증 능력