Korean Chem. Eng. Res., 61(4), 523-541 (2023)
https://doi.org/10.9713/kcer.2023.61.4.523
PISSN 0304-128X, EISSN 2233-9558

3D 프린팅 소재 화학물질의 독성 예측을 위한 Data-centric XAI 기반 분자 구조 Data Imputation과 QSAR 모델 개발

정찬혁‡ · 김상윤‡ · 허성구 · Shahzeb Tariq · 신민혁 · 유창규†

경희대학교 공과대학 환경응용과학과 융합공학전공
17104 경기도 용인시 기흥구 덕영대로 1732
(2023년 5월 23일 접수, 2023년 7월 5일 수정본 접수, 2023년 9월 1일 채택)

Data-centric XAI-driven Data Imputation of Molecular Structure and QSAR Model for Toxicity Prediction of 3D Printing Chemicals

ChanHyeok Jeong‡, SangYoun Kim‡, SungKu Heo, Shahzeb Tariq, MinHyeok Shin and ChangKyoo Yoo†

Integrated Engineering, Department of Environmental Science and Engineering College of Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-Gu, Yongin-si, Gyeonggi-do, 17104, Korea
(Received 23 May 2023; Received in revised from 5 July 2023; Accepted 1 September 2023)

요 약

3D 프린터의 활용이 높아짐에 따라 발생하는 화학물질에 대한 노출 빈도가 증가하고 있다. 그러나 3D 프린팅 발생 화학물질의 독성 및 유해성에 대한 연구는 미비하며, 분자 구조 데이터의 결측치로 인해 in silico 기법을 사용한 독성 예측 연구는 저조한 실정이다. 본 연구에서는 화학물질의 분자구조 정보를 나타내는 주요 분자표현자의 결측치를 보간하여 3D 프린팅의 독성 및 유해성을 예측한 Data-centric QSAR 모델을 개발하였다. 먼저 MissForest 알고리즘을 사용해 3D 프린팅으로 발생되는 유해물질의 분자표현자 결측치를 보완하였으며, 서로 다른 4가지 기계학습 모델(결정트리, 랜덤포레스트, XGBoost, SVM)을 기반으로 Data-centric QSAR 모델을 개발하여 생물 농축 계수(Log BCF)와 옥탄올-공기분배계수(Log Koa), 분배계수(Log P)를 예측하였다. 또한, 설명 가능한 인공지능(XAI) 방법론 중 Tree-SHAP (SHapley Additive exPlanations) 기법을 활용하여 Data-centric QSAR 모델의 신뢰성을 입증하였다. MissForest 알고리즘 기반 결측지 보간 기법은, 기존 분자구조 데이터에 비하여 약 2.5배 많은 분자구조 데이터를 확보할 수 있었다. 이를 바탕으로 개발된 Data-centric QSAR 모델의 성능은 Log BCF, Log Koa와 Log P를 각각 73%, 76%, 92% 의 예측 성능으로 예측할 수 있었다. 마지막으로 Tree-SHAP 분석결과 개발된 Data-centric QSAR 모델은 각 독성치와 물리적으로 상관성이 높은 분자표현자를 통하여 선택함을 설명할 수 있었고 독성 정보에 대한 높은 예측 성능을 확보할 수 있었다. 본 연구에서 개발한 방법론은 다른 프린팅 소재나 화학공정, 그리고 반도체/디스플레이 공정에서 발생 가능한 오염물질의 독성 및 인체 위해성 평가에 활용될 수 있을 것으로 사료된다.

Abstract – As accessibility to 3D printers increases, there is a growing frequency of exposure to chemicals associated with 3D printing. However, research on the toxicity and harmfulness of chemicals generated by 3D printing is insufficient, and the performance of toxicity prediction using in silico techniques is limited due to missing molecular structure data. In this study, quantitative structure-activity relationship (QSAR) model based on data-centric AI approach was developed to predict the toxicity of new 3D printing materials by imputing missing values in molecular descriptors. First, MissForest algorithm was utilized to impute missing values in molecular descriptors of hazardous 3D printing materials. Then, based on four different machine learning models (decision tree, random forest, XGBoost, SVM), a machine learning (ML)-based QSAR model was developed to predict the bioconcentration factor (Log BCF), octanol-air partition coefficient (Log Koa), and partition coefficient (Log P). Furthermore, the reliability of the data-centric QSAR model was validated through the Tree-SHAP (SHapley Additive exPlanations) method, which is one of explainable artificial intelligence (XAI) techniques. The proposed imputation method based on the MissForest enlarged approximately 2.5 times more molecular structure data compared to the existing data. Based on the imputed dataset of molecular descriptor,

† To whom correspondence should be addressed.
E-mail: ckyoo@khu.ac.kr
‡ The first and second authors have identical collaboration in this research paper.

the developed data-centric QSAR model achieved approximately 73%, 76% and 92% of prediction performance for Log BCF, Log Koa, and Log P, respectively. Lastly, Tree-SHAP analysis demonstrated that the data-centric-based QSAR model achieved high prediction performance for toxicity information by identifying key molecular descriptors highly correlated with toxicity indices. Therefore, the proposed QSAR model based on the data-centric XAI approach can be extended to predict the toxicity of potential pollutants in emerging printing chemicals, chemical process, semiconductor or display process.

Key words: 3D printing, Data imputation, Explainable AI (XAI), Quantitative-structure activity relationship (QSAR), Data-centric AI, Computational toxicology

1. 서 론

4차 산업혁명(Industrial Revolution, IR 4.0)이 대두되면서 전세계적으로 제조업 분야의 혁신이 이뤄지고 있다[1]. 인더스트리 4.0으로 알려진 제조업 분야의 혁신은 스마트 팩토리(smart factory) 기술을 중심으로 이뤄지고 있으며, 스마트 팩토리의 핵심 기술인 ICT (Information and Communication Technology)와 3D 프린팅 기술의 중요성도 함께 증대되고 있다[2]. 기존 제조업의 조형(Prototyping) 공정은 디자인, 목업, 금형, 사출 과정을 거치며, 제품의 크기와 재질에 따라 통상적으로 6주 이상이 소요되었지만, 3D 프린팅 기술을 이용한 신속조형기술(Rapid Prototyping; RP)은 조형 공정을 2일 내외로 단축시켜, 사용자 맞춤형 생산과 유통, 물류 서비스에 보다 신속한 대응으로 주목받고 있다[3]. 이와 같이, 3D프린팅은 4차 산업혁명 시대에 전자, 자동차, 항공우주, 의료 등의 다양한 첨단 제조산업 분야에 활용될 것으로 예상된다[4].

3D 프린터 관련 산업의 세계 시장 규모는 2020년부터 시작된 COVID-19에 의한 산업 분야의 침체에도 불구하고 꾸준히 성장하고 있다. 2017년 3D 프린팅 관련 산업의 세계 시장 규모는 73.4 억 달러로 전년 대비 21% 증가하였고, 2023년까지 최대 273억 달러에 이를 것으로 예측된다[5]. 3D 프린터의 활용 범위 역시, 과거에는 일부 산업에 국한된 것에 비해 최근에는 3D 프린터의 대표 기술인 수지 압출 적층 조형 기술(Fused Deposition Modeling, FDM)의 특허가 만료되어 보급형 3D 프린터가 개발되었고, 다양한 산업 분야에서 저렴한 비용으로 3D 프린터를 활용할 수 있게 되었다. 이에 따라, 2020-2021년간 전세계에서3D 프린터를 산업 현장에서 사용하는 경우는 전년대비 68% 증가하였으며, 2022년 기준 3D 프린팅을 통해 10개 이상의 부품을 생산하는 산업은 전체의 49%로 전년 대비 13%가 넘는 성장률을 보였다[6].

3D 프린터 시장의 성장과 보급형 3D프린터가 활용됨에 따라, 3D 프린팅 과정에서 발생하는 화학물질의 성분과 조성에 대한 연구도 활발하게 이루어지고 있다. Stabile 등은 나무 재질, 구리 분말, 대나무 섬유 등 다양한 내부 충진 물질을 포함한 PLA(Polylactic acid)를 이용하여 3D프린팅 실험을 수행하였으며, 재료에 따라 초미세먼지의 발생량은 최대 10배까지 차이를 보였다[7]. Kim 등은 ABS(Acrylonitrile butadiene styrene) 및 PLA소재의 3D프린팅 과정 중 유기화합물과 알데하이드류가 발생되는 것으로 보고하였다[8]. 이때 ABS 소재를 이용한 3D프린팅 과정에서 약 150 μg/m^{3}의 TVOC (Total volatile organic carbon)와 약 70 ppb의 포름알데히드가 발생된 반면, PLA소재를 이용한 3D프린팅 과정에서는 약 160 ppb의 포름알데이드가 발생되는 것으로 보고하였다. Azimi 등은 3D 프린팅 소재에 따른 발생물질의 공기 중 성분을 정량적으로 측정하였는데, 3D프린팅 소재의 종류에 따라 최대 180 μg/m^{3}의 카프로락탐(caprolactam)과 스티렌(styrene)이 검출되었으며, 특히 PLA 소재에서는 약 5 μg/m^{3}의 락티드(Lactide)가 검출된 것으로 보고하였다[9]. Steinle은 ABS를 이용하여 3D프린팅할 때 발생하는 물질로 스티렌, 에틸벤젠(ethylbenzene), 사이클로헥사논(cyclohexanone), 메틸메타크릴레이트(methyl-methacrylate)와 노말부탄올(n-butanol)을 지목하였다[10].

3D 프린팅에 이용되는 화학물질의 성분과 조성은 다양해지고 있지만, 이와 관련한 화학물질의 독성 측정 및 인체 위해성에 대한 연구는 부족한 실정이다. 전통적으로, 화학물질의 독성 및 인체위해성 정도는 생물체 내부/생체 외 에서 In vivo/In vitro 임상시험을 통해 수행된다. 그러나 동물에 대한 윤리의식 확산과 동물실험 결과의 인체 적용에 한계가 있어, 최근에는 In silico 방식의 생물대체시험법을 이용하는 추세이다. 이 중에서, QSAR (Quantitative Structure-Activity Relationship) 방법론은 분자의 특징적인 구조와 예측하고자 하는 물성치 사이의 상관관계를 수학적으로 표현하는 기법으로, 다양한 모델을 통해 분자 구조와 생물학적, 물리화학적 특성 및 독성 지표 간의 상관성을 계산할 수 있다[11]. QSAR를 활용한 다양한 독성 예측 연구가 수행되었는데, Ding 등은 hierarchical support vector regression (HSVR) 및 PLS (Partial least square) 기반 QSAR 모델을 사용하여 방향족 질소화합물(Nitroaromatic compounds; NACs)의 TA98 균주의 변이원성을 예측하였다[12]. Kobayashi 등은 Log BCF에 대해 높은 영향력을 갖는 분자표현자를 분석하기 위해 Gradient Boosting Decision tree (GBDT) 기반 QSAR 모델을 개발하였다[13]. 또한, Pandit 등은 ABS 소재를 이용해 3D 프린팅 과정에서 발생하는 물질의 흡입독성(inhalation toxicity), 발암성(carcinogenicity), 간세포독성(hepatotoxicity) 및 기형유발성(teratogenicity)등의 독성 지표를 QSAR 기법으로 분석하였다[14].

QSAR는 빅데이터(big data)와 기계학습(machine learning) 기법의 발전에 힘입어 화학물질의 독성 예측 연구에 활용되어 다양한 성과를 보였지만, 이를 성공적으로 수행하기 위해서는 충분히 많은 데이터 수와 양질의 데이터를 필요로 한다. 하지만 3D 프린팅 분야의 경우, 3D 프린팅 과정에서 발생하는 유해 물질의 종류가 매우 많아 일부 유해한 물질만을 대상으로 하거나, TVOCs의 농도로 대표하여 제시하고 있어, 유해 물질의 조성 및 성분 별 데이터가 부족하다[6,15]. 또한, 3D 프린팅에서 사용되는 화학물질이 신규 물질인 경우 분자표현자 등 관련 데이터 내에 결측치(missing data)가 포함되어 있을 수 있어, 예측 결과가 편향될 수 있다[16]. 이러한 문제는 기존 QSAR 모델의 구조 개선 및 최신 기계학습 모델 적용을 통한 모델 중심(Model-Centric) 접근법으로는 극복하기 어렵다[17].

이를 해결하기 위하여 최근에는 의학 등 다양한 분야에서 Data-Centric 기반 접근법 중 회귀모델(Regression model)이나 분류 모델(Classification model)을 활용하여 결측치를 보간(Imputation)하는 연구가 활발히 진행되고 있다. Yang 등 은 부정맥 기계에서 측정 오류 또는 기기 결함으로 인해 발생한 결측치를 보간하기 위해

Table 1. Summary of previous studies related to 3D printing, QSAR, and missing data imputation

Research field	Description	Ref.
Toxicity experiment of 3D printing chemicals	- Evaluated particle emission and the main influencing parameters from a low-cost 3D printer	[7]
	- Evaluated the emission of particulate matter and gaseous materials during FDM 3D printing	[8]
	- Evaluated emission rates of particles and a broad range of specific VOCs during 3D printing	[9]
	- Identifed a broad range of substances while printing a standard object with both ABS and PLA	[10]
Missing data imputation	- Developed a LightGBM model for clinical time-series data analysis - Verified that the Light GBM model performs better than KNN, XGBoost, and randomforest	[18]
	- Proposed a KNN-based missing value imputation approach to deal with an unbalanced dataset related to arrhythmias	[19]
	- Missing value corresponding to the BDL among chemicals in house dust - Analyzed the correlation between chemical data and childhood carcinogenicity	[20]
	- Solved the toxicity data gap by replacing missing values in the ToxCast database with predictive toxicity data using the DNN model	[21]
QSAR	- Developed novel QSAR models combined with HSVR and PLS methods to predict TA98 strain mutagenicity of NACs	[12]
	- Evaluated the bioconcentration factor (BCF) with the gradient boosting decision tree (GBDT) model	[13]
	- Developed Multilayer perceptron models to predict AOP(Adverse Outcome Pathway) in ToxCast database.	[14]

KNN(K-Nearest Neighbor) 기법을 적용하였다[18]. Luo는 임상 시계열 데이터에 관한 결측치를 보간하기 위해, 혈액 측정 데이터 셋 중 임의의 데이터를 삭제한 후 LightGBM (Light Gradient-Boosting Machine, MICE (Multivariate Imputation by Chained Equations) 및 XGBoost (eXtreme Gradient Boosting)을 적용하였다[19]. Matthew 등은 집 먼지의 화학 물질 중 검출 한계에 해당하는 결측치를 Bayesian group index regression 기법을 적용해 보간하고, 화학물질 데이터와 소아 발암성(Carcinogenic)의 상관관계를 분석하였다[20]. Jeong 등은 DNN 모델을 활용한 예측독성자료로 ToxCast 데이터베이스의 결측치를 보간하였다[21]. 3D 프린팅과 관련한 선행연구에서는 단순히 유해화학물질의 농도 및 독성의 결측치를 정량적으로 보간하는데 그쳤으며, 분자표현자 등 3D 프린팅 유해화학물질의 물리화학적 성질을 보간한 연구는 미비하다. Table 1은 3D 프린팅 과정에서 발생하는 화학물질 및 QSAR 모델, 독성 결측치 보간 선행연구를 나타내었다.

기존의 연구의 문제점을 극복하고자, 본 연구에서는 Data-centric 접근법으로 유해물질의 물리화학적 데이터의 결측치 보간 기법을 적용하고, 이를 바탕으로 3D 프린팅 과정에서 발생하는 화학물질의 독성 예측을 위한 QSAR 모델을 개발하고 XAI (eXplainable-Artificial Intelligence)를 적용하여 해석 가능한 인체 위해성 평가 연구를 수행하였다. 이를 위해 3D 프린팅 과정에서 발생하는 물질에 대한 분자 구조 및 독성 데이터를 수집하였으며, 이후 DRAGON 6 software를 이용하여 3D 프린팅 과정에서 발생하는 물질의 분자표현자를 계산하였다. 계산된 분자표현자의 결측치를 MissForest 기법을 적용하여 보간하였다. 보간된 데이터를 기반으로 3D 프린팅 과정에서 발생하는 물질의 세 가지 독성 지표 별로 기계학습 기반 QSAR 모델을 개발 및 예측 성능을 평가하였다. 마지막으로, XAI 기법 중 Tree-SHAP (SHapley Additive exPlanations)을 적용해 독성에 높은 영향력을 갖는 분자표현자를 분석하여 예측 신뢰성을 입증하였다.

2. 3D 프린팅 소재 별 유해물질 발생 메커니즘

3D 프린터에 사용되는 원재료는 플라스틱 폴리머, 천연재료, 세

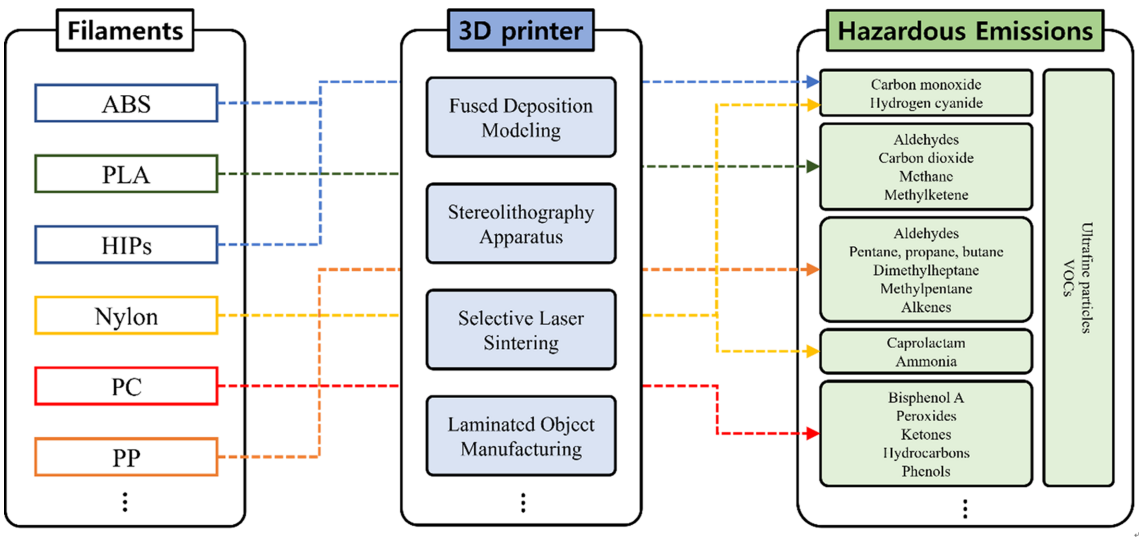
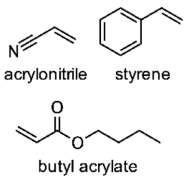
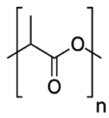
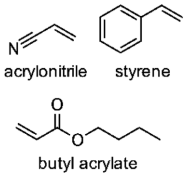
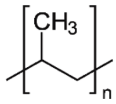
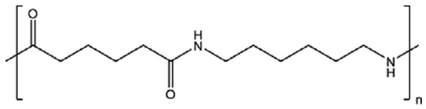
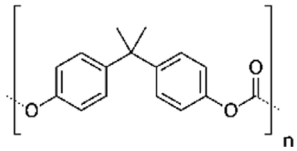


Fig. 1. Hazardous pollutants emitted from 3D printing processes according to filaments including Acrylonitrile butadiene styrene (ABS), Polylactic acid (PLA), High impact polystyrene (HIPs), Nylon, Polycarbonate (PC) and Polypropylene (PP).

Table 2. Monomer structure of 3D printing filaments and hazardous pollutants emitted from 3D printing operation

Filaments	Melting Point	Hazardous pollutants	Monomer
ABS	210-250 ℃	- Ultrafine particles - Aromatic VOCs - Carbon monoxide - Hydrogen cyanide	
PLA	180-220 ℃	- Ultrafine particles - Aldehydes - Carbon monoxide/dioxide - Methane - Methylketene	
HIPs	180-250 ℃	- Ultrafine particles - Aromatic VOCs - Carbon monoxide - Hydrogen cyanide	
PP	230-260 ℃	- Ultrafine particles - Aldehydes - Pentane, Propane, butane - Dimethylheptane - Methylpentane - Alkenes	
Nylon	240 ℃-	- Ultrafine particles - Caprolactam - Nitriles, aromatic VOCs - Ammonia - Hydrogen cyanide	
PC	250-320 ℃	- Ultrafine particles - Bisphenol A - Peroxides - Aldehydes - Ketones - Hydrocarbons - Phenols	

라믹 금속, 콘크리트 재질 등으로 다양하지만, 주로 플라스틱 중합체를 기반으로 하는 필라멘트를 이용한다. 플라스틱 중합체 기반 필라멘트는 고분자성 폴리머(polymer)에 플라스틱 가소재, 내연재, 안정재 등 첨가물이 함유되어 있으며, 이들이 프린팅 과정에서 열분해 되며 발생하는 물질은 사용자의 안전과 건강에 악영향을 줄 수 있다. Fig. 1은 대표적인 3D프린팅 소재인 Acrylonitrile butadiene styrene (ABS), Polylactic acid (PLA), High impact polystyrene (HIPs), Nylon, Polycarbonate (PC) and Polypropylene (PP) 및 3D프린팅 과정에서 발생하는 물질의 발생 경로를 나타내었으며, Table 2에는 3D 프린팅 과정에서 사용되는 주요 필라멘트의 작동 온도, 발생 가능한 화학물질의 종류 및 단량체 구조를 나타내었다. 선행연구를 통해 3D 프린팅 과정에서 사용된 필라멘트의 종류와 운전 조건에 따라 다양한 물질이 발생하는 것으로 나타났다.

2-1. Acrylonitrile butadiene styrene (ABS)

ABS는 acrylonitrile, 1,3-butadiene, styrene을 중합하여 얻어지는 중합체로 전자제품 및 자동차의 부품소재로 많이 사용되며, 보급형 3D 프린터의 재료 중에서 가장 많이 사용되는 소재이다. 일반 플라스틱과 유사한 강도를 보이지만 성형성, 내구성, 내열성, 전기적 성질 등이 우수한 것으로 평가된다[22]. 내구성은 매우 뛰어나지만 녹은 뒤 굳는 데까지 소요되는 시간이 짧아 완성 후에 갈라지거나 바닥에 달라붙는 현상이 발생할 수도 있다. 일반적으로 약 105 ℃에서 녹으며, 일관적인 압출을 위해서는 약 230~260 ℃의 고온을 유지해야 한다. 특히, 열 분해 과정에서 ABS를 구성하고 있는 단량체를 포함하여 휘발성 유기화합물, 일산화탄소, 시안화 수소 등 유해 물질이 발생되는 것으로 보고되었다[23].

2-2. Polylactic acid (PLA)

PLA는 옥수수 전분에서 추출한 원료로 만든 생분해성(biodegradable) 수지로 일반 수지가 열을 받았을 때 발생되는 환경호르몬, 중금속 물질 등 유해 물질이 검출되지 않아 상대적으로 안전한 재료로 평가된다. PLA는 일반 플라스틱과 동일한 특성을 가지나 폐기 시 미생물에 의해 생분해되는 장점을 가진다[24]. ABS와 더불어 3D 프린터 원료로 가장 많이 사용되는 원료로써 ABS가 가지는 단점을 대부분 해결한 재료이며, 3D프린팅 과정에서 ABS보다 낮은 온도인 150~210 ℃을 사용한다. 또한 열분해 과정에서 메탄, 메틸 케텐, 알데히드류 등 유해물질이 다른 소재에 비해 상대적으로 낮은 농도로 발생되는 것으로 알려져있다.

2-3. High Impact Polystyrene (HIPS)

HIPS는 기존 폴리스틸렌의 약점인 내충격성을 보강하기 위하여 rubber 성분을 보강한 수지이다. HIPS는 기계적 성질이 우수하고, 성형가공성의 용이하며, 전기적 특성이 뛰어난 재료로 평가된다. HIPS의 용도는 ABS수지의 용도와 유사하며, 재활용이 가능한 장점이 있어 3D 프린팅의 재료로 많이 사용된다. 3D 프린팅 과정에서 HIPS의 가열 온도는 180~250 ℃ 이며, 고온에 의해 열분해 될 때 알데하이드류와 유기 화합물, 알켄류 등 유해 물질이 발생된다.

2-4. Nylon

나일론은 가볍고 층간 결합이 강하여 강도가 높으며 내구성이 좋은 소재이다. 순수한 열가소성 수지이므로 재활용이 가능하며, 강한 산을 제외한 보통의 유기용제 또는 알칼리에 대한 강한 내화학성을 가진다. ABS와 PLA와 같이 공기 중 습기를 흡수하려는 성질이 가지고 있기 때문에 습도가 높은 환경에서는 인쇄 품질에 악영향을 받을 수 있다. 선행 연구 결과에 의하면 3D 프린팅 과정에서 나일론으로부터 발생되는 유해물질은 카프로락탐(caprolactam) 단량체, 니트릴, 케톤류, 방향족 유기화합물, 암모니아, 시안화 수소가 있다.

2-5. Polycarbonate (PC)

PC는 가볍고 투명한 형태의 열가소성 수지로, 응용 범위가 매우 넓다. 고온에서 열 변형이 쉽게 일어나며 유연성이 뛰어나 변형이 생겼을 때 원래 모양으로 돌아가려는 성질을 가지고 있다. PLA와 ABS보다 높은 밀도와 인장 강도를 가지고 있어 높은 강도와 유연성을 가진 투명 제품을 가공할 때 이상적인 재료이다. 3D 프린팅 과정에서 PC는 보통 260~300 ℃을 사용하며, 일반적으로 300 ℃에 가까울수록 인쇄 품질이 높은 것으로 알려져 있다. PC의 단량체는 비스페놀 A와 포스겐으로 구성되어 있어 고온에서 열분해가 이루어질 때, 비스페놀 A를 비롯하여 알데하이드류, 유기화합물, 페놀류 등이 발생될 수 있다.

2-6. Polypropylene (PP)

PP는 상온에서 지방, 유기용제 등 화학물질에 대한 내화학성을 가지며, 일반 생활 제품의 주 원료로 사용되는 플라스틱의 일종이다. 재활용이 가능한 장점이 있어 3D 프린팅의 재료로 많이 사용되고 있으며, 에틸렌과 중합화하여 강도를 높일 수 있어 산업용으로도 많이 사용된다. 그러나 고온 환경에서 가공될 때 쉽게 변형되는 성질이 있으며, 우수한 전기 절연적 특성으로 인해 표면에 먼지를 비롯한 이물질이 붙는 단점이 있다. 3D 프린팅 과정에서 PP의 가열 온도는 230~260 ℃이며, 고온에 의해 열분해 될 때 알데하이드류와 유기 화합물, 알켄류 등 유해 물질이 발생된다.

3. 연구 방법

Fig. 2는 인체 위해성 평가를 위한 XAI 기반 Data-centric QSAR 모델 개발을 위한 순서도를 나타낸다. 3D 프린팅 과정에서 발생되는 화학물질의 분자 구조 및 독성(Log P, Log Koa, Log BCF) 데이터를 참고문헌을 조사하여 수집하였으며, DRAGON 6 software를 사용하여 분자표현자를 계산하였으며, 누락된 결측치는 MissForest를 이용하여 계산된 보간하였다. 보간된 분자표현자를 기반으로 기계학습 기법 중 결정트리(Decision Tree; DT), 랜덤포레스트(Randomforest; RF), XGboost, SVM (Support Vector Machine) 를 사용하여 독성 예측을 위한 ML-QSAR 모델을 개발하였으며, 독성 예측 성능은 R^2, RMSE (Root mean squared error), MAE (Mean absolute error)를 이용하여 평가하였다. 마지막으로 XAI 기법 중 Tree-SHAP을 통해

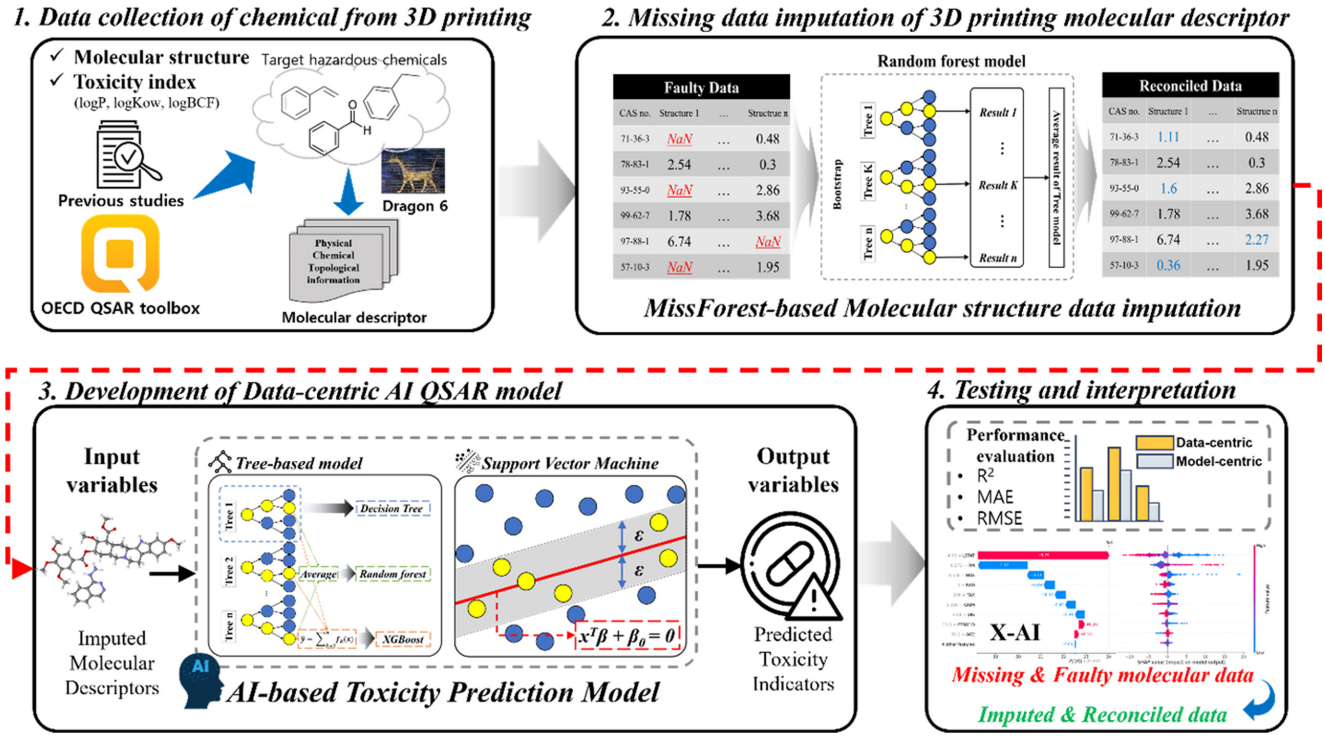


Fig. 2. Research framework of molecular descriptors imputation and data-centric XAI-driven QSAR model for toxicity prediction of 3D printing chemicals.

각 독성에 대해 높은 영향력을 갖는 분자표현자를 분석하여 QSAR 모델의 신뢰성을 평가하였다.

3-1. 3D 프린팅 발생 유해물질의 독성 데이터 수집

3D 프린팅 과정에서 발생하는 유해물질의 Data-centric 결측치 보간 및 인체 위해성 평가 QSAR 모델 개발을 위해, 3D 프린팅 과정에서 발생하는 유해물질에 대한 5개의 선행 연구를 조사하여, 3D 프린팅 소재와 그 때 발생하는 유해물질의 종류에 대한 데이터를 수집하였다[25-29]. 3D 프린팅 과정에서 발생하는 물질의 독성지표(Log BCF, Log Koa, Log P)는 물질안전보건자료(MSDS, Material Safety Data Sheet)와 화학물질 오픈 데이터베이스인 Pubchem (https://pubchem.ncbi.nlm.nih.gov/)에서 수집하였으며, 독성 실험 데이터가 없는 경우 OECD QSAR Toolbox에서 제공되는 값을 사용하였다.

3-1-1. OECD QSAR Toolbox

OECD (Organization for economic cooperation and development)가 제공하는 OECD QSAR Toolbox 프로그램은 화학물질에 대한 국제적인 법적 규제를 위해 개발한 독성 예측 프로그램으로 다양한 화학물질이 가지는 종말점(end point)을 일괄적으로 계산할 수 있다[30]. OECD QSAR Toolbox의 데이터베이스는 산업체와 규제당국에서 보고된 독성 정보를 기반으로 하고 있으며, 매년 신규 화학물질의 종말점 데이터를 갱신하여 제공하고 있다. OECD QSAR Toolbox는 대상 화학물질의 구조나 발생 메커니즘이 유사한 화학물질끼리 분류할 수 있으며, 분류된 화학물질 별로 물리화학적 특성 정보를 확인할 수 있다. 이를 통해 유사한 구조적 특징을 갖는 분자들의 정보를 이용하여 새로운 화학물질의 종말점 또는 물리적 특성을 예측할 수 있다는 장점이 있다. 본 연구에서는 3D 프린팅 발생 물질의 독성 지표인 Log BCF, logKOA와 Log P를수집하기 위해, 3D 프린팅 발생 물질의 구조적인 특징 중 작용기를 기반으로 3D 프린팅 발생 물질과 유사한 화학물질 정보를 각 독성 지표별로 199개, 205개, 209개의 데이터를 수집하였다.

3-1-2. 3D 프린팅 발생 화학물질의 분자표현자 계산 및 결측치 분석

3D 프린팅 발생 화학물질의 분자 구조 및 물성치를 정량화하기 위해 DRAGON 6 software를 사용하였다. Dragon 6 software는 질량, 반 데르 발스 부피, 샌더슨 전기 음성도 및 분극성 등의 주요 원자 매개변수를 원자 가중치로 분자 표현자를 계산하며[31], 화학물질의 정량적구조활성관계에 대한 다양한 연구에 사용되고 있다[15,32]. Table 3는 독성 평가 ML-QSAR 모델 개발을 위하여 DRAGON 6 software를 이용하여 계산한 3D 프린팅 발생 화학물질의 분자표현자의 종류와 개수를 나타낸다. 분자표현자는 분자량 및 특정 원소(C, O 등)의 수와 같은 분자의 기본적인 정보를 나타내는 Constitutional indices group과, 분자의 위상학적 구조에 기반한 Topological indices group등을 포함한 29개의 그룹으로 구분할 수 있으며, 분자표현자의 총 개수는 4,885 개이다.

DRAGON 6 software를 이용하여 계산한 분자표현자 중에는 NaN값으로 나타난 결측치가 있으며, Table 4는 참고문헌을 통해 조사한 Log BCF, Log Koa, Log P에 따른 3D프린팅 발생 화학물질의 수와, 분자표현자의 결측치 비율을 나타낸다. Log BCF에 대하여 총 199개의 3D 프린팅 발생 화학물질의 종말점 데이터가 수집되었으며, 해당 화학물질의 분자표현자 중 23.93%가 결측치를 포함하고 있다. Log Koa의 종말점 데이터는 총 205개의 3D프린팅 발생 화학물질에 대해 수집되었으며, 분자표현자 중 23.57%가 결측치로 나타났다. Log P는 209개의 3D 프린팅 발생 화학물질의 종말점 데이터를 수집하였으며, 분자표현자 중 23.56%가 결측치이었다. 세 가지 독성 지표에 대해 평균적으로 23.68%의 분자표현자가 결측치인 것으로 분석되었다. 위상학적 분자구조를 기반으로 계산되는 2D matrix-based descriptor 그룹 중 VE1D/Dt, VE2D/Dt, VE3D/Dt 분자표현자는 65% 이상의 화학물질에 대하여 결측치에 해당되었다. 이는 3D 프린팅 발생 화학물질은 위상학적 구조에 관련한 분자표현자 데이터가 미비함을 나타내며, 정확한 QSAR 모델링을 위

Table 4. Toxicity indices and the number of 3D printing chemicals and percentage of those missing molecular data

Toxicity	No. of 3D printing chemicals	Percentage of missing molecular descriptors (%)
Log BCF	199	23.93
Log Koa	205	23.57
Log P	209	23.56

Table 3. The molecular descriptors in each group calculated by DRAGON 6 software [15]

Group name	Numbers	Group name	Numbers
Constitutional indices	43	Ring descriptors	32
Topological indices	75	Walk and path counts	46
Connectivity indices	37	Information indices	48
2D matrix-based descriptors	550	2D autocorrelations	213
Burden eigenvalues	96	P_VSA-like descriptors	45
ETA indices	23	Edge adjacency indices	324
Geometrical descriptors	38	3D matrix-based descriptors	90
3D autocorrelations	80	RDF descriptors	210
3D-MoRES descriptors	224	WHIM descriptors	114
GETAWAY descriptors	273	Randic molecular profiles	41
Functional group counts	154	Atom-Centered fragments	115
Atom-type E-state indices	170	CATS 2D	150
2D Atom Pairs	1596	3D Atom Pairs	36
Charge descriptors	15	Molecular properties	20
Drug-like indices	27	Total	4885

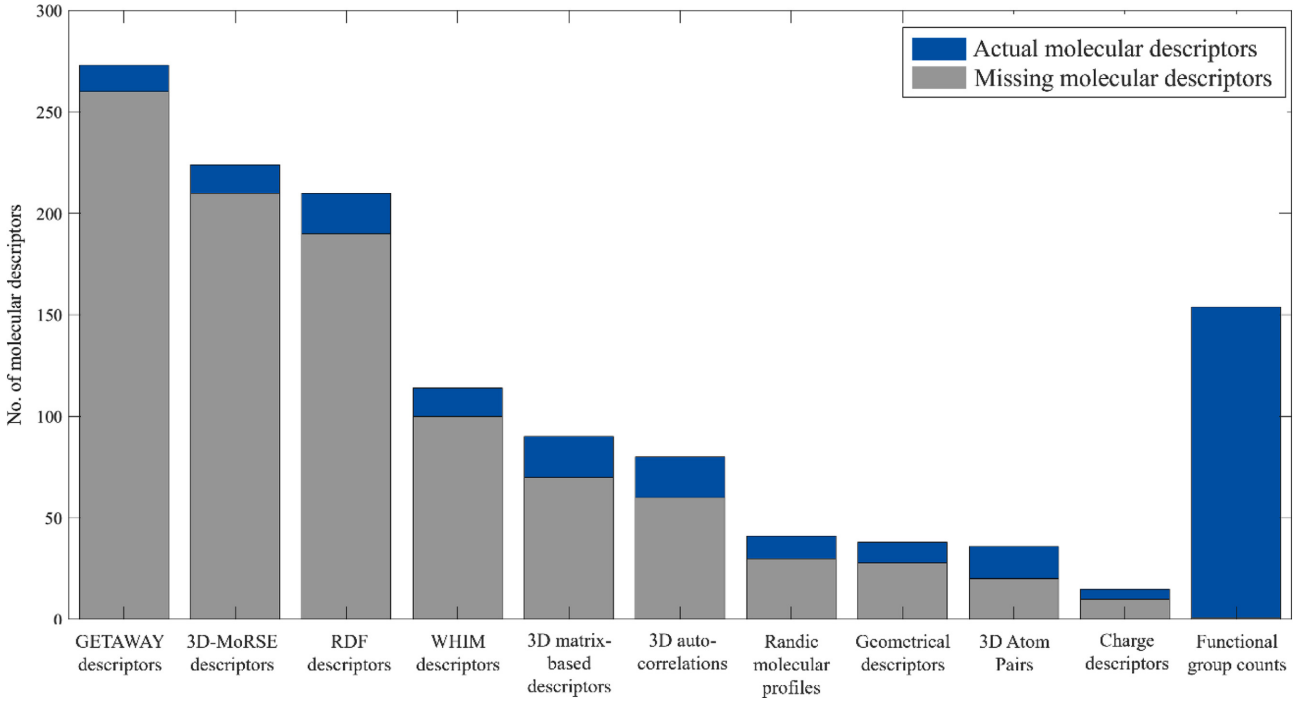


Fig. 3. Balance of actual and missing molecular descriptors in missing-data-included groups.

하여 다양한 물리화학적 구조데이터의 결측치 보간을 수행해야 함을 시사한다.

각 독성지표별 수집한 3D 프린팅 발생 화학물질에 대하여 전체 분자표현자 중 약 22.97%인 1,122개가 결측치로 나타났다. 이와 같이, Fig. 3은 분자표현자 Group별로 모든 3D 프린팅 발생 화학물질에 대하여 결측치에 해당하는 분자표현자 개수를 나타낸다. 분자표현자 Group별로, GETAWAY descriptors group (273개), 3D-MoRSE descriptors group (224개), RDF descriptors group (210개) 순이며, GETAWAY descriptors는 Molecular Influence Matrix에 따른 화학구조의 특성을 나타내고[33], 3D-MoRSE descriptors는 전자회절(Electron diffraction)에 기초한 분자 구조물의 3차원 형태를[34], RDF descriptors는 분자를 구성하고 있는 원자쌍 사이의 최인접 평균거리를 나타내는 반경분포함수(RDF, radial distribution function)을 통해 계산되는 분자표현자이다[15].

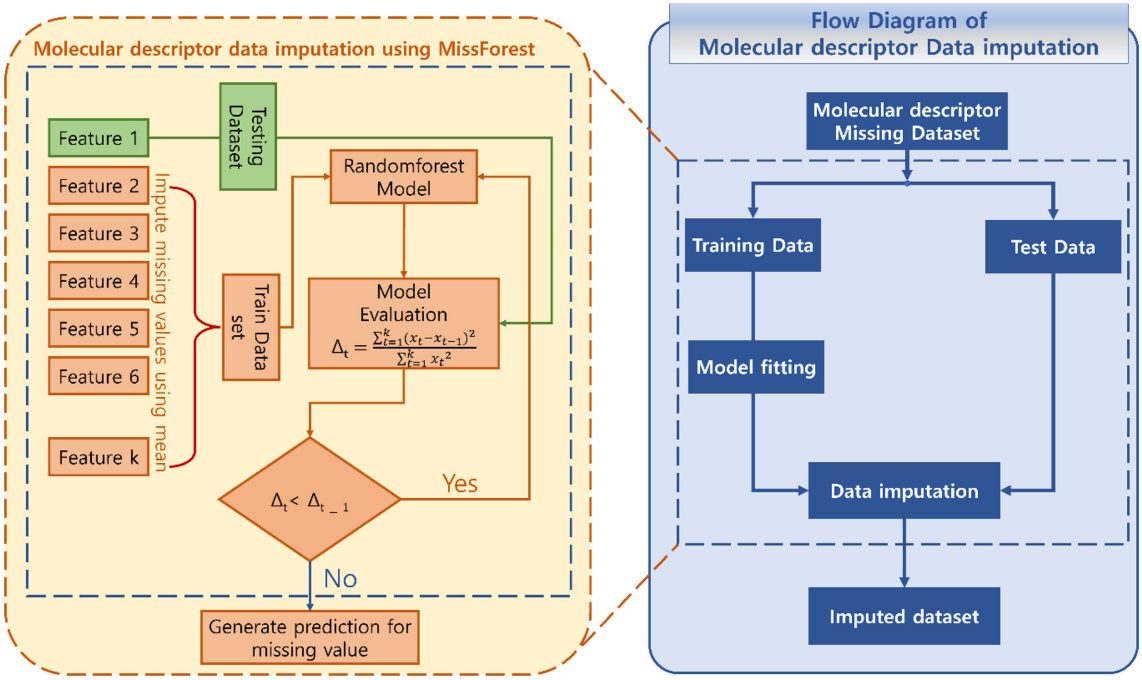


Fig. 4. MissForest algorithm workflow for the missing imputation of molecular descriptors.

3-2. MissForest 알고리즘을 이용한 Data-centric 분자표현자 결측치 보간

분자표현자 값의 결측치를 보간하기 위하여 적용된 MissForest 알고리즘은 랜덤포레스트 모델을 기반으로 한 예측 알고리즘이다. Fig. 4는 MissForest 알고리즘의 결측치 보간 방법에 대한 흐름도를 나타내었다. 먼저 종속변수를 결측치가 있는 분자표현자들 중 하나로 설정하고, 이를 제외한 나머지 분자표현자를 독립변수로 하여 MissForest 내부의 랜덤포레스트 모델을 훈련시키고, 예측값으로 결측치를 보간하였다. MissForest 알고리즘은 손실 함수를 사용한 기존의 렌덤포레스트 알고리즘과는 달리, 결측치에 대한 보간값의 변화량이 최소화되도록 식 (1)과 같이 알고리즘을 학습시킨다[33].

$$\Delta_N = \frac{\sum_{j\in N}(X_{imp}^{(t)} - X_{imp}^{(t-1)})^2}{\sum_{j\in N}(X_{imp}^{(t)})^2} \quad (1)$$

이 때, Δ_N은 N번째 종속변수의 결측치에 대한 보간값의 변화량을 나타내고, $X_{imp}^{(t)}$ t 번째 iteration에서 보간값 행렬을 나타낸다. 식 (1)에서, iteration이 진행됨에 따라 새로운 보간값과 이전 보간값의 변화량인 Δ_N을 기준으로, Δ_N가 더 이상 감소하지 않는 iteration까지 MissForest 알고리즘을 학습시킨다[35]. MissForest 알고리즘은 기존 보간 방법들과 달리, 수치형 및 범주형 변수가 혼합되어 있는 데이터에 적용할 수 있어 기존의 보간 방법들보다 더 다양한 분야에 적용이 가능하다. 또한, 랜덤 포레스트 알고리즘의 구조적 특성으로 인하여 서로 다른 변수들로 구성된 각각의 트리로부터 강건한 보간 결과를 얻을 수 있다는 장점이 있다.

3-3. 기계학습 기반 3D프린팅 화학물질의 독성 예측을 위한 data-centric QSAR 모델

QSAR 모델은 화학물질의 독성 및 물성치와 화학물질의 분자구조 사이의 관계를 수학적으로 모델링하는 기법이다. 화학물질의 독성을 QSAR를 이용하여 계산할 때, 화학물질의 분자구조는 분자표현자로 표현되며, QSAR 모델은 화학물질의 독성과 분자표현자 사이의 상관관계를 반응 함수로 나타내어, 반응 함수에 어떤 모델을 활용하는지에 따라 다양한 QSAR 모델을 개발할 수 있다. QSAR의 일반적인 수식형태는 식 (2)과 같이 나타낼 수 있다.

$$\text{Property or toxicity} = f(\text{Molecular descriptor}) + E_r \quad (2)$$

위 식에서 f는 분자표현자와 독성 사이의 반응 함수를, E_r은 예측된 독성과 측정값 간의 오차를 나타낸다. 기존의 물리화학적 방정식에 기반한 QSAR 모델은 이론적 · 실험적으로 알려진 일부 화학적 특성 간의 관계에 대해서만 모델링 할 수 있었으나, 기계 학습을 적용한 QSAR 모델은 물리화학적 방정식에 의존하는 모델에 비해 효율적이며 대량의 데이터를 처리할 수 있고, 화학물질의 독성과

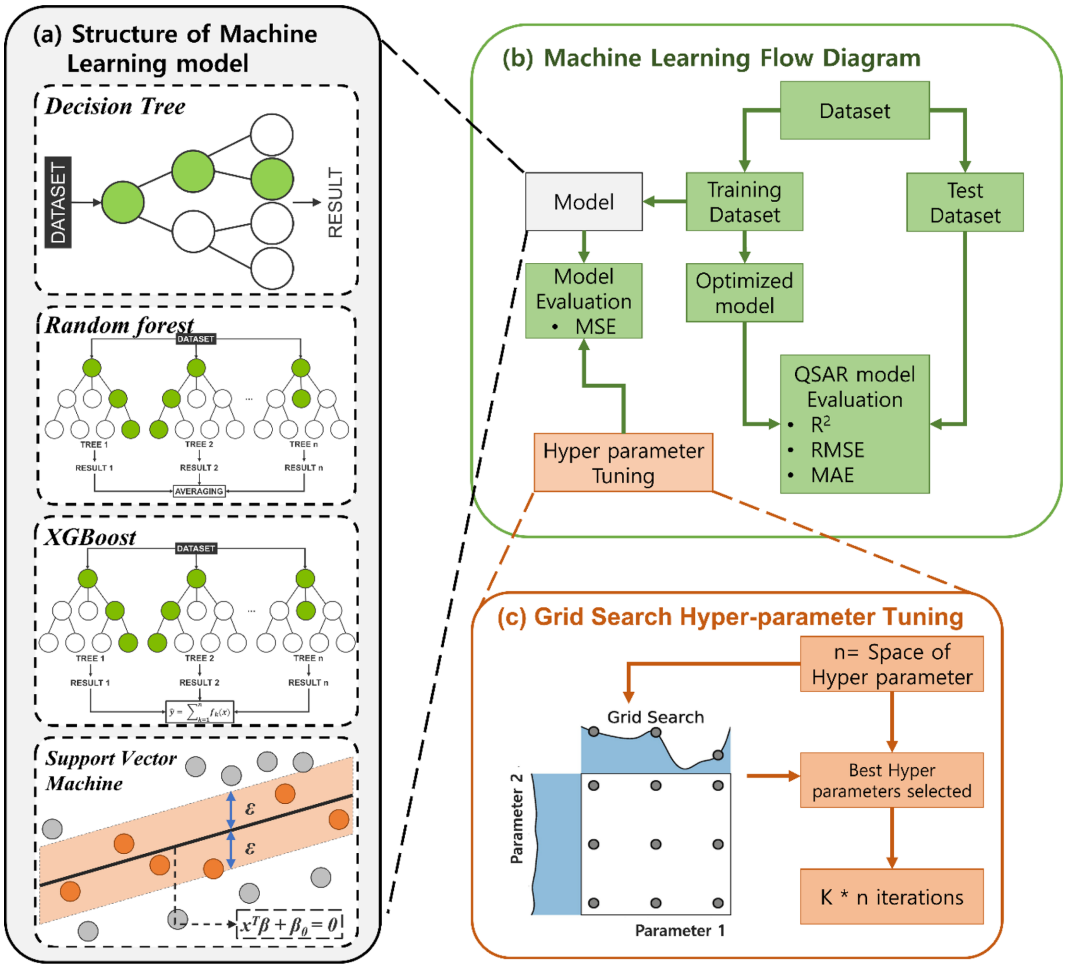


Fig. 5. Overall workflow of ML-QSAR model development: (a) Structure of machine learning models, (b) Model training procedure, and (c) Hyperparameter tuning using grid search.

분자표현자의 비선형적 관계 역시 모델링 할 수 있다는 장점이 있다. Fig. 5는 기계학습을 활용하여 3D 배출 화학물질의 독성 예측을 위한 ML-QSAR 모델의 개발 과정을 나타내었다. 본 연구에서는 Python 3.8의 scikit-learn 1.10.1, xgboost 1.7.3 (XGboost 실행을 위한 배경 라이브러리) 라이브러리를 활용한 결정 트리, 랜덤포레스트, XGBoost 및 서포트 벡터 머신 모델을 ML-QSAR 모델의 반응함수로 이용하였으며 각 기계학습 알고리즘의 구조적 형태는 Fig. 5(a)와 같다. Fig. 5(b)와 같이 수집된 3D 프린팅 발생 화학물질의 각 독성 지표별 종말점 데이터는 학습데이터(Training set)와 테스트데이터(Test set)를 7:3으로 나누어 ML-QSAR 예측 모델을 학습 및 검증하였다.

Fig. 5(c)는 GridSearch를 사용해 ML-QSAR 모델의 하이퍼 파라미터에 대한 최적값을 찾는 과정을 도식화한 것이다. GridSearch는 가능한 모든 하이퍼 파라미터의 조합을 확인하는 방법이다[36]. 본 연구에서는 각 하이퍼 파라미터의 설정에 대해서 모두 학습한 뒤, 예측 성능을 비교하여 최적의 하이퍼 파라미터의 설정값을 탐색하였다. 개발한 Data-centric QSAR 모델의 예측 성능을, 결측치를 제거한 데이터 기반의 Model-centric QSAR 모델의 예측 성능과 비교하였다. 개발된 ML-QSAR 모델의 예측 성능은 R^2 (Coefficient of determination)와 Pearson 상관계수(r), 그리고 MAE (Mean absolute error)를 이용하여 평가하였다. 평가 지표 중 R^2는 예측 결과의 분산을 기반으로 모델의 상대적인 성능을 나타낼 때 주로 사용된다. Pearson 상관 계수는 연속형 데이터의 선형 상관 관계를 정량화 한 수치이며[37], MAE는 모든 절대 오차(Error)의 평균을 나타낸다. R^2와 Pearson 상관계수, MAE는 각각 식 (3)에서 (5)와 같이 표현할 수 있다.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{measured} - y_{predicted})^2}{\sum_{i=1}^{n}(y_{measured} - y_{mean})^2} \tag{3}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{4}$$

$$\mathrm{MAE} = \sum_{i=1}^{n} |y_{measured} - y_{predicted}| \tag{5}$$

$y_{predicted}$은 QSAR 모델의 예측 값, $y_{measured}$은 실제 측정값, y_{mean}은 측정값의 평균을 나타내며, n은 데이터의 개수를 의미한다.

3-3-1. 결정 트리(Decision Tree)

기계학습 알고리즘 중 결정 트리는 훈련 데이터의 종속 변수를 가장 잘 설명하는 독립변수의 기준을 학습하는 알고리즘이다. 결정 트리 모델의 자식 노드는 학습 과정에서 독립변수의 불순도(Impurity)를 최소화하도록 생성된다. 불순도는 해당 범주 안에 서로 다른 데이터가 얼마나 섞여 있는지를 의미하며, 불순도를 정량화하기 위한 지표인 지니 지수(Gini Index)와 엔트로피 지수(Entropy Index)는 각각 식 (6)과 식 (7)로 나타낼 수 있다.

$$G = 1 - \sum_{j=1}^{c} P(j)^2 \tag{6}$$

$$E_{ntropy}(T) = \sum_{j=1}^{c} P_i \log^2 P_i \tag{7}$$

결정 트리의 회귀 모델은 식 (8)과 같이 오차 제곱합(Residual Sum of Squares; RSS)을 최소화하는 변수(Predictor)를 기준으로 분기를 만든다. 회귀 트리의 모든 독립변수 X_1, …, Xp에 대한 반평면(Half-planes) 쌍은 식 (9)과 같이 정의할 수 있으며, 이후 결정 트리 모델은 식 (10)를 최소로 하는 j 및 s값을 찾는다. 이 과정은 결정 트리의 회귀 모델이 정지기준(Stopping criterion)을 만족할 때까지 반복된다.

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j}) \tag{8}$$

$$R_1(j,s) = \{X | X_j < s\} \text{ and } R_2(j,s) = \{X | X_j < s\} \tag{9}$$

$$\sum_{i:x_i \in R_1} (y_i - \hat{y}_{R_j})^2 + \sum_{i:x_i \in R_2} (y_i - \hat{y}_{R_2})^2 \tag{10}$$

여기서 j는 분할된 공간의 개수를 나타내며, s는 각 독립변수의 절단점(Cutpoint)를 의미한다.

결정 트리 모델은 다른 기계 학습 모델에 비해 독립 변수와 종속변수 사이의 관계를 설명하기 용이하다는 장점이 있지만, 결정 트리는 종속변수가 연속형인 경우 일반적으로 예측력이 저하된다는 단점이 있으며, 고차원의 데이터일수록 해석력 또한 감소한다는 단점이 있다.

3-3-2. 랜덤포레스트(Random Forest)

랜덤포레스트는 결정 트리 모델을 여러 개 구축하여 앙상블 학습의 배깅(Bagging) 기법을 적용한 기계 학습 기법이다. 배깅 기법은 여러 결정 트리를 구축한 후, 전체 학습 데이터 셋으로부터 무작위 복원 추출된 부트스트랩 데이터셋에 대해 학습한 결과를 집계(Aggregation)하는 앙상블 기법의 일종이다. 회귀 모델의 배깅 기법의 집계 방식은 식 (11)과 같다.

$$\varnothing_B(x) = \frac{1}{B}\sum_{b=1}^{B} \varnothing(x, L_b) \tag{11}$$

여기서 L은 학습 데이터를 의미하며, B는 학습 데이터에 대해에 대해, 부트스트랩(Bootstrap) 샘플링을 수행하여 얻은 데이터 셋의 개수이다. 각 결정 트리마다 학습 데이터 셋이 다르기 때문에 구축된 결정 트리의 모델과 예측된 값이 모두 다르며, 서로 다른 결정 트리의 예측 값을 결합하여 랜덤포레스트의 예측 값을 결정한다.

랜덤포레스트 모델은 일반적으로 학습 데이터의 결측치를 처리하기 용이하며, 과적합(Overfitting)을 방지하여 높은 정확도를 보이는 장점이 있다. 그러나 변수의 개수가 적은 경우 낮은 무작위성으로 인해 성능이 저하되며, 결정 트리 모델 대비 낮은 해석력을 갖는다는 단점이 있다.

3-3-3. XGBoost

XGBoost는 결정 트리 모델에 앙상블 학습의 부스팅(Boosting) 기법을 적용한 알고리즘으로, 다른 Tree-based ensemble 모델과 달리 CART(Classification and regression trees)라는 모델로 구성된다. CART에서 각 노드는 데이터가 아닌 하나의 값이 할당되며, 각각의 CART는 해당 트리의 구조와 노드의 예측 값을 포함하고 있다.

XGBoost는 각 모델을 결합하는 방식으로 부스팅 기법을 적용한다. 부스팅 기법은 주어진 데이터를 약한 분류기를 통해서 학습한 후 학습된 결과에서 나타나는 오차를 또 다른 약한 분류기에서 학습시켜 오차를 줄여 나간다. 즉 첫 번째 학습을 통해 생성된 모델에서 오류를 발생시키는 데이터들을 다음 모델을 이용하여 오류를 줄이고 또 다시 발생된 오류 데이터들을 그 다음 모델을 이용하여 오류를 줄이는 방식을 순차적으로 적용하게 된다. 이후 부스팅 기법을 통해 각 모델별로 서로 다른 가중치를 부여하여 중요도가 높은 모델에 높은

점수를 부여한다. 첫 번째 학습의 예측 값을 $\hat{y}_i = 0$ 으로 설정했을 때, 부스팅 기법에 따른 매 단계의 예측 값은 식 (12)와 같다.

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{12}$$

여기서 $\hat{y}_i^{(t)}$는 t 번째의 데이터 x_i에 대한 모델의 예측 값을 나타내며, f는 CART 모델을 나타낸다. XGboost의 t번째 모델이 가지는 가중치는 t-1번째의 오차에 따라서 결정된다. 이때, t단계에서 훈련시킬 목적 함수는 식 (13)와 같이 정의할 수 있다.

$$obj^{(t)} = \sum_{k=1}^{t} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_i)$$

$$= \sum_{k=1}^{t} l(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)) + \Omega(f_i) + \text{constant} \tag{13}$$

XGBoost 모델의 손실함수는 측정 값과 예측 값 사이의 오차와 정규화 항 $\Omega(f_i)$ 으로 구성되며, XGBoost모델은 이러한 손실함수를 최소화하는 가중치를 구하여 모델을 학습한다.

3-3-4. 서포트 벡터 머신(Support vector machine)

SVM (Support Vector Machine) 기법은 구조적 위험 최소화(structural risk minimization)원리에 기반하여 분류(classification)를 하기 위하여 개발된 학습 방법이다[37]. 추후 SVM기법은 ε-무감도 손실 함수를 도입함으로써 회귀 예측에도 적용 가능한 SVR(Support Vector Regression)으로 확장되었다. SVR은 저차원 입력공간에 있는 비선형 회귀 문제를 고차원 형상공간(feature space)에 사상(mapped)시킨다. SVR의 선형 추정함수는 식 (14)와 같다.

$$f(x) = \omega^T x + b \tag{14}$$

ε-무감도 손실함수 L_ε은 SVR에 사용되는 비용함수로, 식 (15)와 같이 나타낼 수 있다.

$$L_\varepsilon = (F(x), q) = \max(0, |f(x) - q - \varepsilon) \tag{15}$$

여기서 ε는 회귀 모델 f(x)의 주변에 위치한 튜브의 반지름을 나타내는 정밀모수(Precision parameter)이다.

SVR의 선형 추정함수의 Weighted Vector (v)와 상수는 식 (16)의 정규화 된 위험함수로 추정된다.

$$R(C) = C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon f(x_i, q_i) + \frac{1}{2}\|\omega\|^2 \tag{16}$$

여기서 $\frac{1}{2}\|\omega\|^2$는 SVR의 정규화 항이며, C는 경험적 위험과 정규화 항의 균형을 맞추는 데 사용되는 정규화 된 상수이다[37].

SVR모델은 실제 값 q_i와 예측 값 f(x)의 값을 가능한 ε 범위 안에서 유지하면서, 마진을 최대화하게 된다. 최종적으로 SVR의 손실함수는 식 (17)과 같이 나타낼 수 있다.

$$L_{SVM} = \min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{17}$$

$$s.t. \begin{cases} (\omega^T x_i + b) - y_i \le \varepsilon + \xi_i \\ y_i - (\omega^T x_i + b) \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

SVR의 손실함수에 Lagrangian 승수와 Karush-Kuhn-Tucker 조건을 적용한 SVR 기반 회귀함수의 일반적 형태는 식 (18)와 같이 나타낼 수 있다.

$$f(x, v) = f(x, a, a^*)$$

$$= \sum_{i=1}^{n}(a_i + a_i^*)K(x_i + x_j) + b \tag{18}$$

이때 $K(x_i, x_j)$는 커널 함수를 의미한다. SVR은 커널 함수를 사용하여 학습 데이터를 특징공간의 점으로 변환시킨 후, 특징 공간에서 모델을 훈련시키게 된다. SVM 회귀 모델은 이상치(Outlier)에 대한 영향이 적고, 커널 함수를 통해 비선형 데이터에 대한 회귀 분석이 가능하며 높은 예측 정확도를 보이는 특징이 있다.

3-4. Tree-SHAP을 이용한 독성에 대한 분자표현자의 영향력 분석

일반적으로 독성 예측을 위한 기계학습 모델의 경우 사용된 데이터를 그대로 이용하는 데이터 기반 모델(Data-centric QSAR 모델)로서 input-output 데이터를 근사하는 black-box 모델 구조로 되어서 3D 프린팅 발생물질의 독성과 분자 표현자 사이의 관계에 대한 설명이 부족하다는 한계점이 존재한다. 따라서 본 연구에서는 결정트리 기반 앙상블 학습 모델에서 계산되는 특징 중요도(Feature or Variable Importance)[38]뿐만 아니라 입력변수의 중요도와 영향력까지 분석 가능한 SHAP(SHapley Additive exPlanations) 기법을 이용하여 QSAR 모델의 독성 예측 결과에 관한 신뢰성을 data-centric 관점에서 설명하였다. SHAP 기법은 학습된 모델에서 각 입력변수가 종속변수에 얼마나 영향을 미치는지를 나타내는지에 대한 샤플리 값(Shapley Value)을 계산하여 입력변수의 기여도를 해석하는 기법이다[39].

SHAP 기법은 각 입력변수에 대한 샤플리 값을 구하여 입력변수 중 특정 데이터에 관해 설명할 수 있으며, 이를 집계한 결과를 통해 입력변수의 전역적인 해석도 가능하다는 점에서 기계학습 모델의 결과에 대한 입력변수의 영향력을 더욱 정확하게 파악할 수 있다는 장점이 있다[39]. 샤플리 값은 음수와 양수의 값으로 표현하여, 각 입력변수의 기여도에 관해 정량적으로 계산한다. 예를 들어, 종속변수인 독성에 대해 입력변수인 분자표현자 중 분자량의 샤플리 값이 양수로 나타난다면, 분자량이 높을수록 분자가 높은 독성을 갖게 됨을 의미한다. SHAP에서 입력 X와 훈련된 모델 f가 주어지면, SHAP는간소화된 입력 X'를 매핑 함수 X = hX(X)로 정의한다. $Z' \approx X'$ 인 조건을 만족하는 경우 설명 가능한 모델은 g(Z') = f(hX(Z')) 로 학습되며, 이때 모델 $g(Z')$는 아래의 식 (19)과 같이 표현할 수 있다. 각 특성 값의 기여도를 쉽게 설명할 수 있는 간단한 모델 g로 f를 근사화한다.

$$g(Z') = \varnothing_0 + \sum_{i=1}^{p} \varnothing_i Z_i' \tag{19}$$

여기서 p는 각 화학물질 별 분자표현자의 개수를 나타내며, $\varnothing_0 = f(\text{hX}(0))$은 모델의 예측 평균값이다. $\varnothing_i$ 는 각 분자표현자가 모델에 미치는 기여도를 나타내는 샤플리 값을 의미한다. 본 연구에서는 ML-QSAR모델에 Tree-SHAP을 적용하여 각 분자표현자가 독성 예측에 미치는 영향을 분석하여, Data-centric 기반 QSAR 모델의 예측 성능을 설명하였다.

4. 결과 및 고찰

4-1. MissForest 기반 3D 프린팅 발생물질의 Data-centric 분자표현자 결측치 보간 결과

먼저 MissForest 알고리즘을 이용하여 3D 프린팅 발생 화학물질의

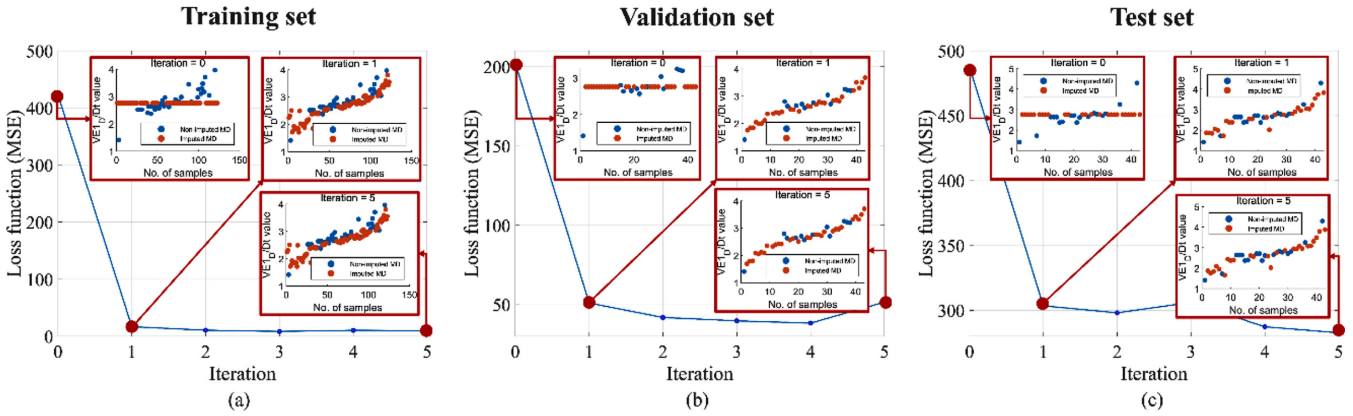


Fig. 6. Training result of MissForest algorithm for missing molecular descriptor data imputation: (a) Training dataset (b) Validation dataset, and (c) Test dataset.

분자표현자 결측치 보간을 수행하였다. 모든 3D 프린팅 발생 유해 물질에 대하여 결측치인 1,122개의 분자표현자를 제외한 나머지 분자표현자의 결측치를 보간하였다. Fig. 6은 본 연구에서 개발된 MissForest 알고리즘의 Log P 데이터 셋에 대한 학습 결과와 VE1D/Dt 에 대한 보간 결과를 보여준다. Log P 데이터는 6:2:2으로 나누어 각각 MissForest 모델의 학습데이터 (Training set), 검증 데이터(Validation set) 및 테스트데이터(Test set)로 활용되었다. Fig. 6에 나타난 것과 같이, VE1D/Dt는 학습 데이터, 검증 데이터 및 테스트데이터에 대해 특정 값으로 치우치지 않고 고르게 퍼져 있음을 확인 할 수 있다. MissForest 알고리즘은 최종적으로 5회의 iteration 동안 학습되었으며, 매 iteration 별로 실제 값과 보간된 값과의 오차를 손실함수로 계산하였다. 손실함수로는 MSE (Mean Squared Error)가 사용되었다. 학습이 진행됨에 따라 MissForest 알고리즘은 2D matrix-based descriptor group의 VE1D/Dt 분자표현자의 물리화학적 특성을 학습하여 데이터 내 결측치를 보간하였다.

학습이 진행됨에 따라 손실함수는 유의미하게 감소하였으며, 학습 데이터셋의 MSE는 최종적으로 9.95 로 수렴되었다. Iteration 횟수가 증가함에 따라 보간된 VE1D/Dt 분자표현자의 경향이 측정된 VE1D/Dt의 경향과 유사해지는 것을 확인할 수 있다. MissForest 알고리즘 학습의 초기 단계(즉, 0 itertation)에서는 MissForest 알고리즘이 결측치가 적은 분자표현자를 우선적으로 보간하여, 65%의 결측치를 가진 VE1D/Dt은 평균값으로 보간되었다. 이후 1번의 iteration 후에는 MissForest 알고리즘을 통해 보간된 VE1D/Dt의 경향이 DRAGON 6 software를 통해 계산된 VE1D/Dt 경향과 유사해졌다. 마지막으로, 5번의 iteration에서의 Δ5가 4번째 iteration 에서의 Δ4와 비교하여 더 이상 감소하지 않았으며, 보간된 VE1D/Dt 데이터는 기존 DRAGON 6 software를 통해 계산된 VE1D/Dt 데이터의 경향을 따르는 것을 확인할 수 있다. 이는 MissForest 알고리즘을 이용한 결측치 보간 방법이 3D 프린팅 발생 화학물질의 물리화학적 특성 및 위상학적 특성을 학습하여 결측치를 보간하였음을 나타낸다.

Fig. 7(a)는 MissForest 알고리즘을 이용해 보간된 VE1D/Dt 분자표현자와 Dragon 6 software를 통해 계산된 VE1D/Dt의 분포 비교 결과를 나타낸다. MissForest를 통해 보간된 VE1D/Dt 데이터의 평균 및 분산은 각각 2.43 및 0.60이고, DRAGON 6 software를 통해 계산된 VE1D/Dt 데이터의 평균 및 분산은 2.79와 0.40으로 유사한 값을 보였다. Fig. 7(b)는 Log P 데이터 셋 중 65% 이상의 결측치를 가진 6 개의 분자표현자에 대한 MissForest 알고리즘의 보간 결과를 나타내며, 보라색 점선은 95% 신뢰 구간을 의미한다. VE1D/Dt, VE2D/Dt, VE3D/Dt, VR1D/Dt, VR2D/Dt, VR3D/Dt 분자표현자들은 모두 위상학적 분자 구조를 기반으로 계산되는 2D matrix-based descriptor group에 속한다. Fig. 7(b)의 보간 전후에 데이터를 비교시, MissForest 알고리즘을 이용해 보간된 36개의 VE1D/Dt 값은 DRAGON 6 software를 이용해 계산된 VE1D/Dt 데이터 셋과 유사한 경향을 보이며, 모두 95% 신뢰구간내에 위치한다. 다른 분자표현자의 보간값의 경우에도 92% 이상이 95% 신뢰구간 내에서 보간된 것을 확인할 수 있다. 따라서 MissForest 알고리즘으로 보간된 분자표현자 데이터는 기존 DRAGON 6 software를 통해 계산된 분자표현자 데이터의 경향을 따르며, 유사한 데이터 분포를 나타내므로 MissForest 알고리즘이 3D 프린팅 발생 화학물질의 분자구조 및 물리화학적 특성을 반영하여 결측치를 보간하였음을 시사한다.

Fig. 8은 MissForest 알고리즘을 이용하여 결측치를 보간한 후 분자표현자의 수와 기존 DRAGON 6 software를 이용하여 계산된 분자표현자의 수의 비교를 나타낸다. MissForest 알고리즘을 이용하여 결측치를 보간한 결과, 보간한 분자표현자 데이터에 대해 모든 값이 0으로 상수인 분자표현자와 같이 물리화학적으로 무의미한 데이터들을 제외한, 독성 예측에 활용할 수 있는 분자표현자의 수는 DRAGON 6 software를 이용하여 계산된 분자표현자 수에 비해 최소 2.3 배에서 최대 2.47 배까지 증가하였다. 이를 통해, MissForest 알고리즘을 이용한 결측치 보간 기법은 기존 방법과 비교하여 3D 프린팅 발생 화학물질의 물리화학적 특성을 학습함으로써 분자표현자의 수를 증가시켜 QSAR 모델의 예측 성능 향상에 기여할 수 있을 것으로 사료된다.

4-2. 3D 프린팅 발생 화학물질의 독성 예측을 위한 Data-centric QSAR 모델 개발 결과

3D 프린팅 발생 화학물질의 독성 예측을 위한 Data-centric QSAR 모델을 개발하기 위해, 결정 트리(Decision Tree, DT), 랜덤 포레스트(Random Forest, RF), XGBoost , 서포트 벡터 머신(Support Vector Machine, SVM) 4가지 기계 학습 모델을 활용하였다. Data-centric QSAR 모델의 예측 성능은 R^2, RMSE, MAE를 이용하여

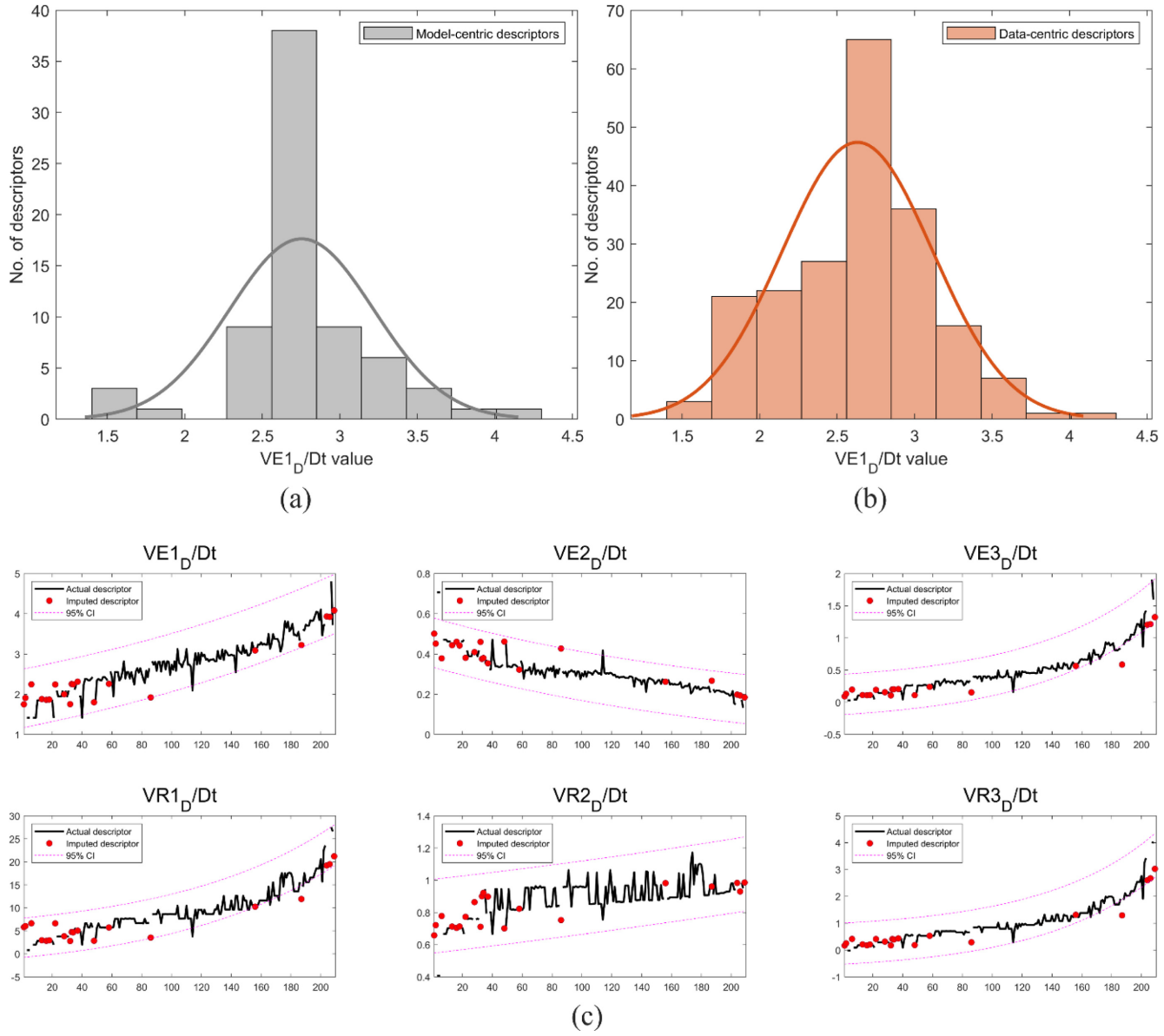


Fig. 7. Distribution of (a) original and (b) imputed molecular descriptor data, and (c) imputation results for molecular descriptors in 2D matrix-based descriptor group by MissForest algorithm.

평가하였다. Table 5는 GridSearch 방법을 이용하여 최적화한 각 독성 지표별 ML-QSAR 모델의 하이퍼 파라미터의 값을 나타낸다.

Table 6는 Log BCF, Log Koa 및 Log P에 Data-centric QSAR 모델의 예측 성능과 기존의 결측치 제거 데이터를 기반으로 한 Model-centric QSAR 모델의 예측 성능을 나타낸다. XGboost를 이용한 Data-centric QSAR 모델의 예측 성능은, Log BCF의 경우 0.73의 R^2 값과 0.45의 MAE을 보인 반면, XGBoost를 이용한 Model-centric QSAR 모델의 예측 성능은 0.67의 R^2 값과 0.50의 MAE으로 Data-centric QSAR 모델이 기존의 결측치 제거 데이터 기반의 Model-centric QSAR 모델보다 더 좋은 예측 성능을 보였다. Log Koa의 경우 SVM를 이용한 Data-centric QSAR 모델의 예측 성능이 0.76의 R^2 값과 0.78의 MAE으로 가장 좋았으며, SVM을 이용한 Model-centric QSAR 모델의 예측 성능은 0.71의 R^2 값과 0.87의 MAE으로 Data-centric QSAR 모델의 예측 성능이 Model-centric QSAR 모델의 예측 성능보다 더 높은 것으로 나타났다.

Log P의 경우 XGboost를 이용한 Data-centric QSAR 모델의 예측 성능은, 0.92의 R^2 값과 0.44의 MAE을 보인 반면, XGboost를 이용한 Model-centric QSAR 모델의 예측 성능은 0.83의 R^2 값과 0.64의 MAE으로 Log BCF와 마찬가지로 Data-centric QSAR 모델이 기존의 Model-centric QSAR 모델보다 더 좋은 예측 성능을 보였다. 이는 기존 결측치 제거 데이터를 이용한 Model-centric QSAR 모델은 결측치가 없는 분자표현자만을 독립변수로 활용하는 데 반해, 개발된 Data-centric QSAR 모델은 3D 프린팅 발생 화학물질의 물리화학적 특성을 학습하고, 이를 바탕으로 분자표현자의 결측치를 보간하여 독립변수로 활용하였기 때문으로 판단된다.

DT, RF, XGboost, SVM을 이용한 Data-centric QSAR모델의 R^2와 Model-centric QSAR 모델의 R^2를 비교한 결과는 Fig. 9와 같다. Log BCF의 경우 결정 트리, 랜덤포레스트, XGBoost, SVM을 이용한 Data-centric QSAR 모델의 R^2 값이 Model-centric QSAR 모델에 비해 각각 0.26, 0.03, 0.06, 0.03 만큼 증가하였다. Log BCF 예측을

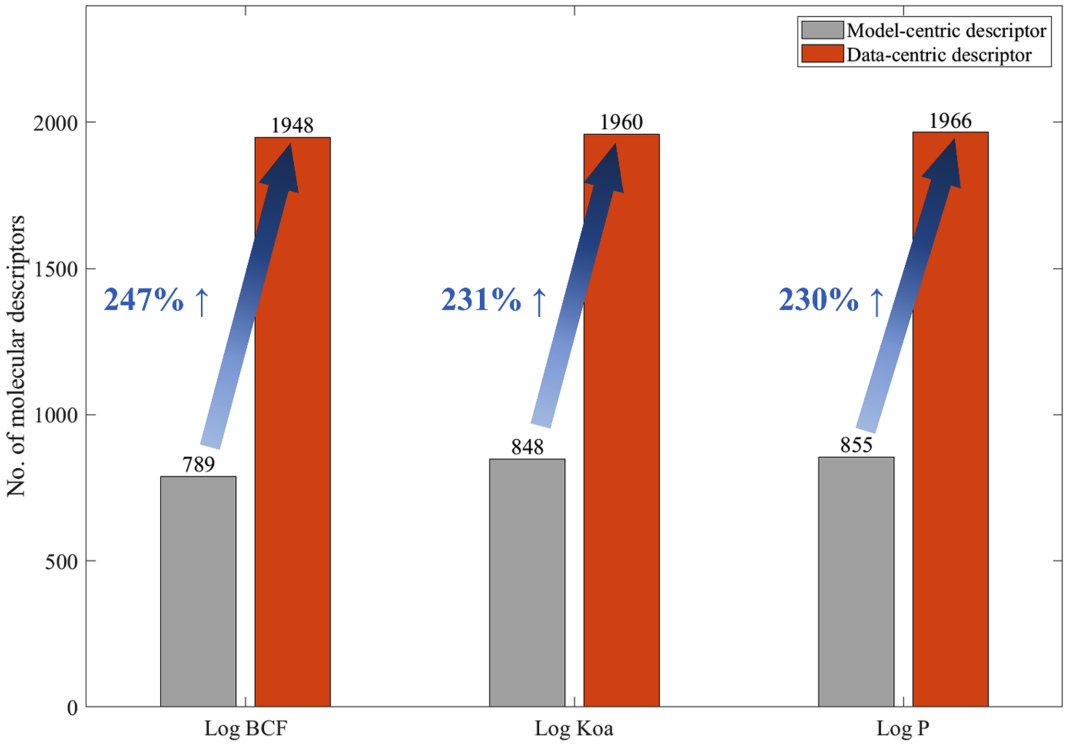


Fig. 8. Comparison of the number of molecular descriptors between imputed and non-imputed molecular descriptors dataset.

Table 5. Hyperparameters tuning results of the data-centric QSAR model by GridSearch

Model	Hyperparameter	Model	Hyperparameter
Decision Tree	Criterion: squared error	XGBoost	Number of estimators: 100
	Maximum depth: none		Subsample:0.8
	Minimum samples leaf: 1		Maximum depth: 3
	Minimum samples split: 2		Learning rate: 0.15
	Maximum leaf nodes: none		Gamma: 10^{-6}
Random forest	Bootstrap: True		Alpha: 0.3
	Criterion: mse		Lambda: 0.1
	Maximum depth: 10	Support Vector Machine	C: 1
	Minimum samples leaf: 1		Epsilon: 0.1
	Minimum samples split: 2		Kernal: RBF
	Number of estimators: 5		Gamma:10^{-3}

Table 6. Comparison of prediction performances in data-centric and model-centric QSAR model

Metric	Toxicity indices	Model-centric QSAR model				Data-centric QSAR model			
		DT	RF	XGBoost	SVM	DT	RF	XGBoost	SVM
R^2	Log BCF	0.28	0.64	0.67	0.68	0.54	0.67	0.74	0.73
	Log KOA	0.07	0.62	0.69	0.71	0.25	0.67	0.85	0.77
	Log P	0.60	0.80	0.83	0.82	0.58	0.83	0.92	0.85
r	Log BCF	0.67	0.82	0.83	0.83	0.74	0.83	0.86	0.85
	Log KOA	0.41	0.79	0.92	0.88	0.56	0.79	0.88	0.84
	Log P	0.78	0.91	0.92	0.90	0.77	0.94	0.92	0.92
MAE	Log BCF	0.68	0.50	0.50	0.49	0.56	0.48	0.45	0.47
	Log KOA	1.50	1.04	0.84	0.87	1.39	1.01	0.94	0.78
	Log P	1.16	0.74	0.64	0.69	1.17	0.72	0.44	0.67

위한 XGBoost 및 결정 트리의 경우 데이터 셋의 보간 여부에 따라 예측 성능의 변화가 큰 양상을 보여, 이러한 모델을 사용할 시에는 분자표현자의 결측치 보간 여부가 매우 중요한 것으로 판명되었다. 반면, 랜덤포레스트와 서포트 벡터 머신 모델의 경우 결측치를 보간하였을 때와 하지 않았을 때의 R^2값 차이가 0.03으로 작았으나, 역시 분자표현자 데이터의 보간 여부에 따라 예측 성능이 향상되는

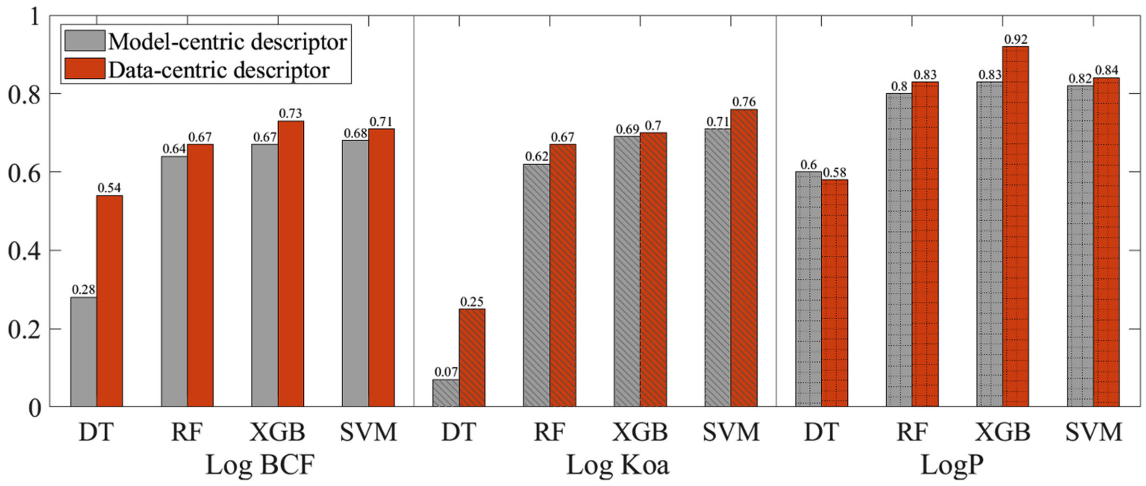


Fig. 9. R^2 score comparisons of the data-centric and model-centric QSAR models across toxicity indices.

경향을 보였다. Log BCF 값의 예측에 있어 가장 좋은 방법은 XGBoost를 이용한 Data-centric QSAR 모델을 이용하는 방법으로 사료된다.

Log Koa의 예측 성능은 결정 트리, 랜덤포레스트, XGBoost, SVM을 이용한 Data-centric QSAR 모델의 R^2 값이 Model-centric QSAR 모델에 비해 각각 0.18, 0.05, 0.01, 0.05 만큼 증가하였다. Log Koa 예측을 위해 결정 트리 모델을 사용했을 경우, 보간된 데이터 셋에 대해 257%만큼 향상된 예측 성능을 보였으며, 결정 트리 모델을 사용해 Log Koa를 예측할 시에는 분자표현자의 결측치 보간 여부가 예측 성능에 큰 영향을 미치는 것으로 나타났다. SVM과 XGBoost 모델의 경우 별도의 보간 절차 없이 결측치를 처리하여도 예측 성능이 크게 저하되지 않았으며, XGBoost를 이용한 Data-centric QSAR 모델은 Log Koa 예측을 위한 ML-QSAR 모델 중 가장 좋은 예측 성능을 보였다.

Log P의 경우 랜덤포레스트, XGboost, SVM을 이용한 Data-centric QSAR 모델의 R^2값이 Model-centric QSAR 모델에 비해 각각 0.03, 0.09, 0.02 만큼 증가하였다. Log P 예측을 위한 XGBoost, SVM, 랜덤포레스트 모델의 경우 분자표현자 데이터의 보간 여부에 따라 그 예측 성능이 3.75%, 10.84%, 2.44% 만큼 향상되었다. 반면, 결정 트리 모델의 경우 결측치를 보간했을 때 오히려 예측 성능이 0.02 만큼 떨어지는 결과를 보였다. Log P 예측을 위한 최적의 ML-QSAR 모델은 XGBoost를 이용한 Data-centric QSAR 모델로 판단된다.

Fig. 10은 Data-centric QSAR모델의 예측 결과와 Model-centric QSAR 모델의 예측 결과에 대해, 예측 모델 중 가장 좋은 성능을 보인 XGBoost 모델과 SVM 모델의 Q-Q plot을 나타낸다. XGBoost 기반 QSAR 모델은 모든 독성 지표인 Log P, Log Koa 및 Log BCF에 대하여 최적의 예측 모델로 나타났다. XGBoost는 모델 학습이 진행됨에 따라 잔차를 줄여 나가는 부스팅 기법을 사용하는 앙상블 모델로, 본 연구에서는 GridSearch를 통해 XGBoost의 8개의 Hyperparameter를 최적화하였으며, 이로 인해 타 모델 대비 뛰어난 성능을 보인 것으로 생각된다. SVM 모델은 고차원 데이터를 처리하는 것에 있어 장점을 보이며, 본 연구에서 사용된 4,885개의 feature를 처리하고 독성을 예측하는데 적합한 모델로 사료된다.

Data-centric QSAR 모델의 독성 예측 결과는 기존의 Model-centric QSAR 모델의 독성 예측 결과에 비교하였을 때, Fig. 10에서 측정값과 예측값이 동일한 45도 상의 점선에 근접하게 위치한 것을 확인할 수 있다. 이는 Data-centric QSAR 모델 개발을 위한 분자표현자 보간 과정에서, 보간된 분자표현자가 3D 프린팅 발생 화학물질의 물리화학적 특성을 반영한 채 보간되었기 때문인 것으로 사료된다. 또한 Fig. 10(c)의 붉은 원으로 표시된 것과 같이, 기존 Model-centric QSAR 모델로 예측된 Log P 값은 3D 프린팅 발생 유해물질의 종류와 무관하게 일정한 값을 나타내는 것을 확인할 수 있다. 이는 기존의 분자표현자 결측치 제거 과정에서 독성과 관련 있는 분자표현자가 제거되어, 독성을 예측하기 위한 분자표현자의 수가 부족하기 때문이며, 따라서 본 연구에서 개발한 기계학습 기반 Data-centric QSAR 모델은 기존 기계학습 기반 Model-centric QSAR 모델보다 향상된 예측 성능을 갖는 것으로 판단된다.

4-3. Tree-SHAP을 이용한 QSAR 모델의 독성 예측에 대한 분자표현자의 영향력 분석

본 연구는 QSAR모델의 독성 예측에 대한 분자표현자의 영향력을 해석하기 위해 Tree-SHAP을 이용하여 전체 분자표현자 중 각 독성지표별로 기여도가 높은 상위 10개를 산출하였으며, Fig. 11에 독성 예측에 기여도가 높은 상위 10개 분자표현자의 섀플리값을 나타내었다. Fig. 11(a)-(c)에는 각각 Log BCF, log KOA, Log P 데이터에서의 독성에 대한 분자표현자의 섀플리값을 나타내었다. 독성 예측에 우수한 예측 성능을 보인 XGBoost 기반 ML-QSAR 모델에 Tree-SHAP을 적용하였다.

Fig. 11(a)에서 확인할 수 있듯이, Log BCF 데이터에서 ALOGP와 CIC0 분자표현자가 독성 예측에 높은 영향력를 보였으며, 두 분자표현자 모두 섀플리 값이 양수이므로 ALOGP와 CIC0 증가함에 따라 Log BCF도 증가함을 알 수 있다. ALOGP(Atom-based method)는 화학물질의 소수성에 대한 원자 단위를 산술적으로 합산한 결과로, Log BCF와 양의 상관관계가 있는 것으로 보고된 지표이다[28]. CIC0은 상호보완적인 정보 내용 지수(0차 이웃 대칭)를 나타내며, 이는 분자 내의 연결성과 작용기를 설명하며 분자 형태 및 녹는점과 연관된다.

Log KOA에 관해서는 Fig. 11(b)에 나타났듯이, GMTIV에 해당하는 섀플리 값의 분포가 가장 컸으며, SMTIV 역시 log KOA를

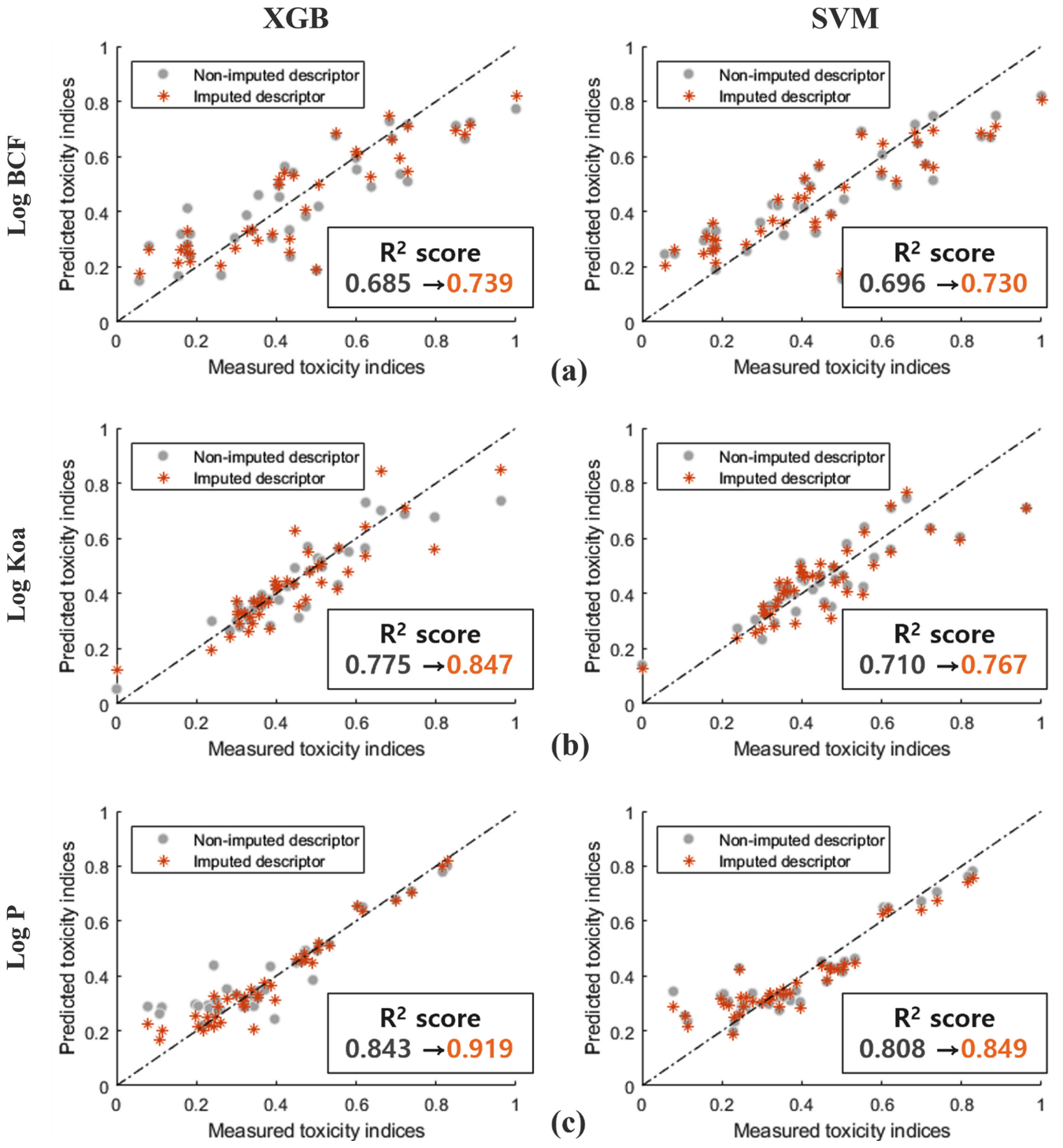


Fig. 10. Prediction results using ML-QASR models based on XGBoost and SVM for (a) Log BCF, (b) log KOA, and (c) Log P.

결정하는 중요한 변수임을 알 수 있었다. GMTIV는 가수 정점 차수에 기반한 Gutman 분자 위상 지수이며, SMTIV는 가수 정점 차수에 기반한 Schultz 분자 위상 지수로 정의되는 분자표현자이다. GMTIV와 SMTIV의 섀플리 값은 모두 음수로, 해당 분자표현자가 높을수록 log KOA 가 낮아짐을 알 수 있다. GMTIV의 경우, 분자의 활성 저해 능력과 관련된 IC50(half maximal inhibitory concentration, 반수 최대 억제 농도)와 양의 상관 관계를 가지는 것으로 보고되었는데[29], IC50의 수치가 낮을수록 분자의 활성 저해 능력이 낮으므로 생물학적 수용체와 결합하기 쉬고, 따라서 높은 log KOA 값을 의미한다. 그러므로 높은 log KOA를 가지는 3D 프린팅 발생 화학물질은 생물학적 수용체에 대한 높은 활성을 가지고 있으며, 이에 따라 IC50는 낮을 것임을 추정할 수 있다. 이러한 관점에서 GMTIV는 log KOA와 음의 상관관계를 가지고 있는 것으로 추정할 수 있다.

Log P의 경우, 앞서 Log BCF와 높은 연관성을 갖는 ALOGP를 포함하여, MLOGP에 대해 넓은 분포를 가진다. 이때 Log P 데이터

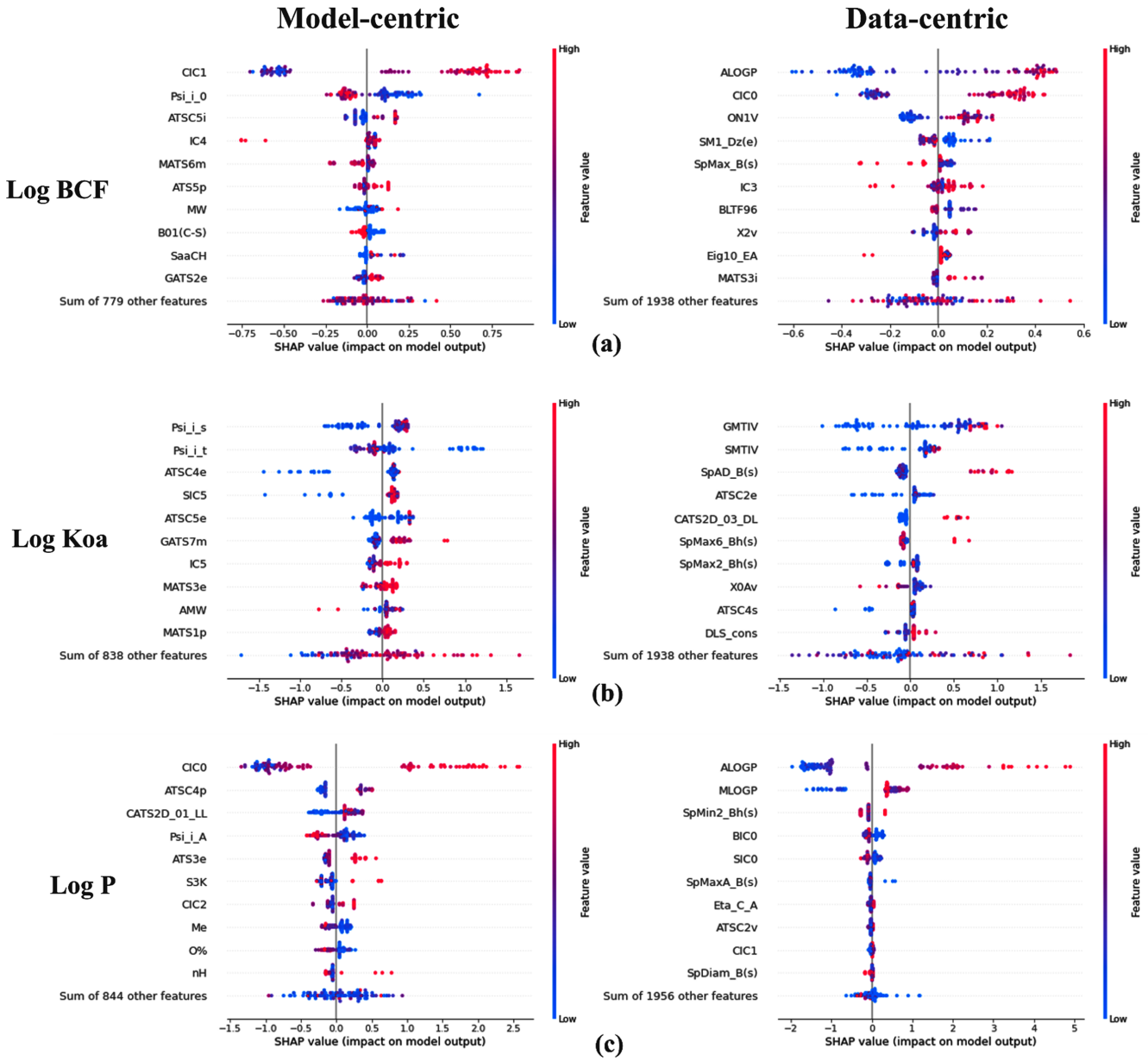


Fig. 11. Comparison of local explanation diagram of SHAP values for data-centric and model-centric QSAR models of (a) Log BCF (b) Log Koa, and (c) Log P.

에 섀플리 값은 모두 양수로, 각 분자표현자가 Log P와 양의 상관관계를 가지는 것으로 나타났다. 또한, 전반적으로 Log P 값을 추정하는데 SpAD_B(s), SpMax6_Bh(s), SpMax2_Bh(s)와 같이 위상학적 분자구조를 기반으로 계산되는 2D matrix-based descriptor에 포함된 분자표현자가 QSAR 모델에 대한 영향이 높다는 것을 확인할 수 있다. Log BCF와 비교했을 때, 두 독성 지표 모두 ALOGP의 섀플리값이 높게 나타났으나, 분자표현자간의 비율은 차이를 보였다. Log BCF의 경우, ALOGP와 CIC0가 비슷한 기여도를 보였고, 다른 분자표현자의 섀플리값도 ALOGP와 차이가 크지 않음을 확인할 수 있다. 그러나 Log P에서는 ALOGP와 MLOGP의 기여도가 다른 분자표현자의 기여도에 비해 높았으며, 이는 ALOGP와 MLOGP 모두 옥탄올/물 분배 계수를 나타내는 지표로 Log P와 큰 연관성을 가지기 때문으로 사료된다.

Fig. 12은 3D 프린팅 발생 화학물질의 Log P 및 log KOA에 대한 분포를 나타내며, 유사한 분포를 보이는 화학물질 그룹의 전역적 해석력(Global interpretablity)를 Waterfall SHAP plot으로 나타내었다. Waterfall SHAP plot에서의 X축은 QSAR 모델의 종속 변수에 해당하는 독성 지표이며, f(x)는 각 화학 물질에 대한 QSAR 모델의 독성 예측 값이고, E[f(x)]는 종속 변수에 대한 예측값의 평균이다. XAI 분석은 4가지 사례에 대해 수행되었으며, 각각 높은 log KOA (Fig. 12(a)), 낮은 log KOA(Fig. 12(b)), Log P(Fig. 12(c)), 높은 Log P (Fig. 12(d))를 보이는 화학물질들에 대해 독성 지표에 대한 분자표현자들의 기여도를 분석하였다. Fig. 12(a)에서, 분자표현자 SpMax7Bh(s), JD/Dt, Eig03EA(ri)가 증가함에 따라 log KOA가 증가하며, 최종적으로 log KOA는 7.158을 나타낸다. Fig. 12(b)에서, 대부분의 분자표현자가 높은 Log KOA를 형성하는데 기여하였

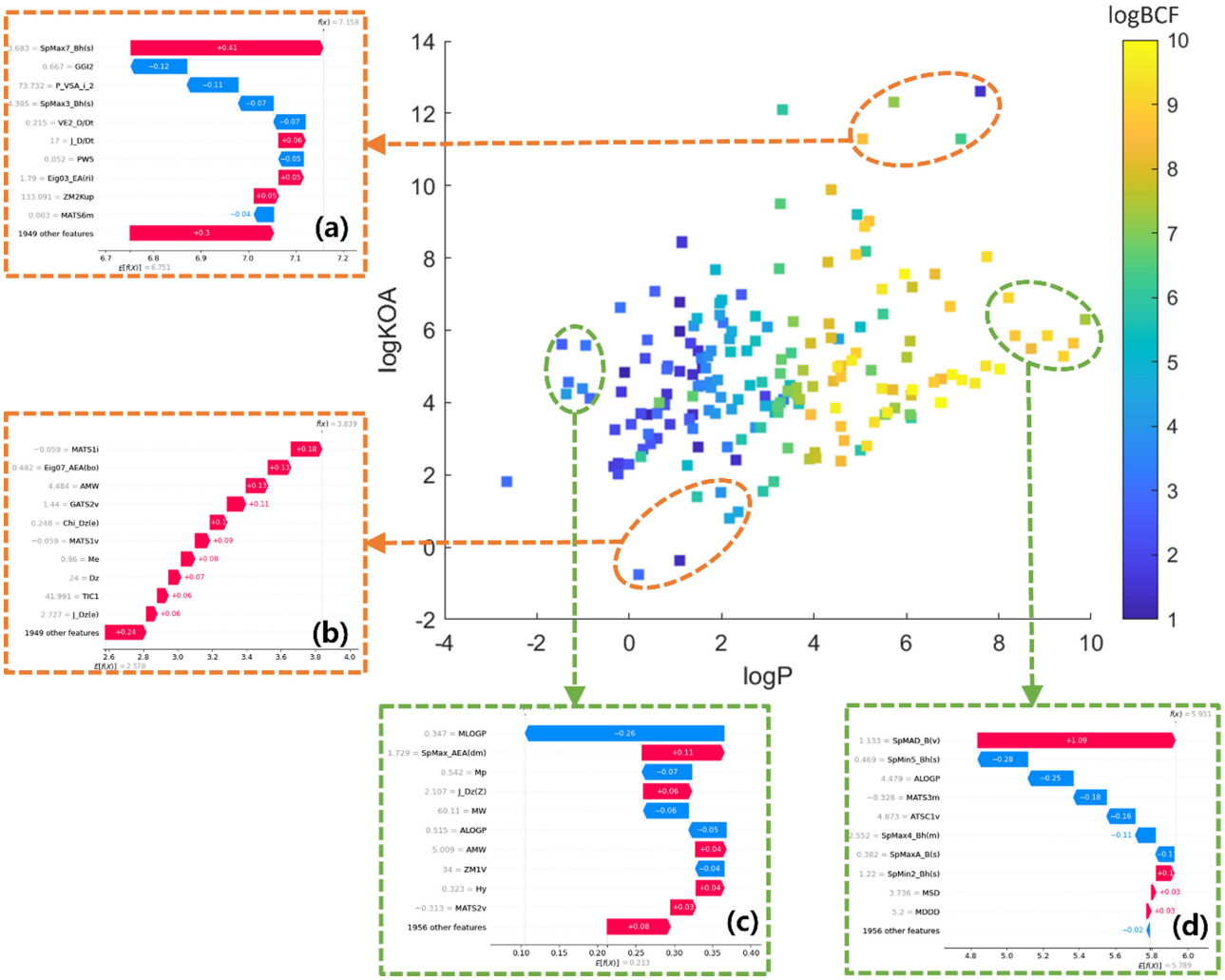


Fig. 12. Waterfall SHAP plots for (a) high Log Koa, (b) low Log Koa, (c) low Log P, and (d) high Log P 3D printing chemicals.

으나, log KOA값은 Fig. 12(a)에 비해 낮은 값을 보였다. Fig. 12(c)에 해당하는 낮은 Log P의 화학물질들의 경우, 양의 섀플리값과 음의 섀플리값이 비슷하여 Log P가 0.213의 낮은 값이 된다. Fig. 12(d)의 경우, 높은 Log P를 보이는 화학물질들에서는 SpMADB(v)가 주요하게 영향을 미치는 것으로 확인되었다. 본 연구에서 개발된 XAI 기반 Data-centric QSAR 모델은 실험 데이터가 부족한 화학물질의 독성을 예측하고, 독성에 영향을 미치는 분자표현자를 XAI를 이용하여 해석함으로써 3D 프린팅 발생 화학물질의 독성 평가를 위해 활용할 수 있다.

Table 7은 본 연구에서 제안한 3D 프린팅 발생 화학물질의 Data-centric QSAR 모델을 이용한 독성 예측 및 XAI 기반 QSAR 모델 결과를 요약하였다. 분자표현자의 결측치를 보간하지 않고 제거했을 때, 독성 예측에 사용되는 분자표현자는 Log BCF, Log Koa, Log P에 대해 각각 789개, 848개, 그리고 855개로 나타났다. 해당 독성 데이터는 모든 분자표현자 결측치를 삭제하였으므로, 데이터의 편향이 일어난다. 따라서 3D 프린팅 발생 화학물질의 물리화학적 특성을 설명하기 위한 분자표현자 수의 부족으로 인해 Model-centric QSAR 모델이 낮은 예측 성능을 보인다. 특히, 분자표현자의 결측치를 MissForest를 이용하여 보간하였을 때, 독성 예측에 사용되는 분자표현자는 Log BCF, Log Koa, Log P에 대해 각각 1948개, 1960개, 1966개로 나타나 기존 방법 대비 약 2.47배 많은 분자표현자를 확보할 수 있었다. 보간된 데이터셋을 이용한 Data-centric QSAR 모델의 독성 예측 성능은 Log BCF, log KOA, log P에 대해 각각 0.73, 0.76, 0.92의 R^2를 보여 Model-centric QSAR model

Table 7. Summary of the data-centric and model-centric QSAR model for toxicity prediction of 3D printing chemicals

Technique		Descriptions		R^2 prediction performance (Improvement)
		Missing data imputation	Toxicity prediction	
Model-centric	Description	- Data bias due to elimination of missing values - Insufficient training data for QSAR model	- Low prediction accuracy - Lack of explainability of the QSAR model	- Log BCF: 0.57 - Log Koa: 0.54 - Log P: 0.76
Data-centric	Description	- Imputation more than 6000 molecular descriptors - Molecular descriptor dataset with high correlation with toxicity	- High prediction accuracy - Interpretation of QSAR model by analysis of influence of molecular desciptor	- Log BCF: 0.73 (16.74%↑) - Log Koa: 0.76 (13.88%↑) - Log P: 0.92 (3.93%↑)

의 예측 성능이 비해 높았다. 이러한 결과를 바탕으로 본 연구에서 제시한 Data-centric 기반 QSAR 모델은 다양한 3D 프린팅 소재 재료와 화학 및 석유화학공정, 그리고 반도체 및 디스플레이 공정 같은 독성 실험이 어렵고 많은 시간이 걸리는 유해 공정에서 발생 가능한 오염물질의 독성 및 인체 위해성 평가에 활용될 수 있을 것으로 사료된다.

4. 결 론

본 연구에서는 3D 프린팅에서 발생하는 화학물질의 독성을 예측하기 위하여 Data-centric 과 기계학습에 기반한 분자표현자 보간 기법과 QSAR 모델을 개발하였다. 먼저, 세 가지 독성 지표에 대하여 3D 프린팅 발생 화학물질 정보와 분자표현자 데이터를 수집하였으며, MissForest 모델을 적용하여 분자표현자의 결측치를 보간하였다. 보간된 분자표현자들을 바탕으로 기계학습 방법론 중 Decision Tree, 랜덤포레스트, SVM, 그리고 XGboost 알고리즘을 이용하여 ML-QSAR 모델을 개발하였다. 본 연구에서 제안된 Data-centric 기반 QSAR 모델 방법론은 분자표현자 결측치를 보간하여 기존 데이터 보다 2.47배 많은 분자표현자를 확보할 수 있었으며, 이를 통하여 3D 프린팅 발생 화학물질의 Log BCF, Log Koa 와 Log P를 각각 %, %, % 증가한 예측성능으로 정확히 예측할 수 있었다. 또한, 설명 가능한 인공지능(XAI) 기법 중 Tree-SHAP을 ML-QSAR 모델에 적용하여 각 독성 지표 예측에 주요한 영향을 미치는 분자표현자를 판별하여, 본 연구에서 개발한 Data-centric 기반 QSAR 모델의 예측 신뢰성을 입증하였다. 본 연구에서 개발된 Data-centric XAI QSAR 모델은 다양한 3D 프린팅 발생물질에 적용 가능하여, 새로운 화학물질의 분자표현자 데이터에 대한 결측치 추정과 독성실험에 대한 시간과 비용의 절약이 가능할 것으로 사료된다.

감 사

본 논문은 연구재단 4단계 BK21 사업과 과학기술정보통신부 재원으로 한국연구재단의 지원을 받아 수행된 연구로 이에 감사를 드립니다(No. 2021R1A2C2007838).

References

1. Kwon, K. M., Kim, H. G. and Moon, S. H., “The 4th Basic Plan for Material and Parts Development,” *Ministry of Trade, Industry and Energy* 16-28(2016).
2. Jang, Y. J. and Jeong, E. M.,, “Global Trends in the 4th Industrial Revolution and Strategies for the Response of Korean Industries,” *Korea Institute for Industrial Economics & Trade* 22-24(2017).
3. Park, S. H., “A Study on R&D Policy through 3D Printing Industry Trend Analysis,” *Science and Technology Policy* **24**(3), 93-104(2014).
4. An, K. C., “Trends and Implications of 3D Printing Industry in the 4th Industrial Revolution,” *Institute for Information & Communications Technology Promotion* 5-8(2018).
5. Zhou, Y., Kong, X., Chen, A. and Cao, S., “Investigation of Ultrafine Particle Emissions of Desktop 3D Printers in the Clean Room,” *Procedia Eng.*, **121**, 506-512(2015).
6. HUBS, *Additive Manufacturing Trend Report 2021* (2021).
7. Stabile, L., Scungio, M., Buonanno, G., Arpino, F. and Ficco, G., “Airborne Particle Emission of a Commercial 3D Printer: the Effect of Filament Material and Printing Temperature,” *Indoor Air*, **27**(2), 398-408(2017).
8. Kim, Y. N., Yoon, C. S., Ham, S. H., Park, J. H., Kim, S. H., Kwon, O. H. and Tsai, P. J., “Emissions of Nanoparticles and Gaseous Material from 3D Printer Operation,” *Environ Sci Technol.*, **49**(20), 12044-12053(2015).
9. Azimi, P., Zhao, D., Pouzet, C., Crain, N. E. and Stephens B., “Emissions of Ultrafine Particles and Volatile Organic Compounds from Commercially Available Desktop Three-Dimensional Printers with Multiple Filaments,” *Environ Sci Technol*., **50**(3), 1260-1268(2016).
10. Steinle, P., “Characterization of Emissions from a Desktop 3D Printer and Indoor Air Measurements in Office Settings,” *J. Occup. Environ. Hyg.*, **13**(2), 121-132(2016).
11. Kim G. H., Lyu K. G., Kim Y. J. and Kim H. C., “A Survey on Quantitative Structure-Activity Relationship(QSAR) Models,” *2008 Proceedings of the Korean Information Science Society Conference*, July, Pyungchang **35**(1), 43-44(2008).
12. Ding Y. L., Lyu Y. C. and Leong M. K., “In Silico Prediction of the Mutagenicity of Nitroaromatic Compounds Using a Novel Two-QSAR Approach,” Toxicology in Vitro **40**(1), 102-114(2017).
13. Kobayashi, Y. and Yoshida, K., “Development of QSAR Models for Prediction of Fish Bioconcentration Factors Using Physicochemical Properties and Molecular Descriptors with Machine Learning Algorithms,” *Ecol Inform* **63**(1), 2-9(2021).
14. Pandit, S., Singh, P., Sinha, M. and Parthasarathi, R., “Integrated QSAR and Adverse Outcome Pathway Analysis of Chemicals Released on 3D Printing Using Acrylonitrile Butadiene Styrene,” *Chem Res Toxicol* **34**(2), 355-364(2021).
15. Kim D. W., Lee S. C., Kim M. J., Lee E. J. and Yoo C. K., “Development of QSAR Model Based on the Key Molecular Descriptors Selection and Computational Toxicology for Prediction of Toxicity of PCBs,” *Korean Chemical Engineering Research* **54**(5), 621-629(2016).
16. To, K. T., Fry, R. C. and Reif, D. M., “Characterizing the Effects of Missing Data and Evaluating Imputation Methods for Chemical Prioritization Applications Using ToxPi,” *BioData Min* **11**(1), (2018).
17. Lee J. G., Shin G. J., Park C. Y. and Hwang U. J., "Robust, fair and scalable data-driven continuous learning," *Communications of the Korean Institute of Information Scientists and Engineers* **40**(11), 53-58(2022).
18. Yang, F., Du, J., Lang, J., Lu, W., Liu, L., Jin, C. and Kang, Q., “Missing Value Estimation Methods Research for Arrhythmia Classification Using the Modified Kernel Difference-Weighted KNN Algorithms,” *Biomed Res Int* **2020**(1), 1-9(2020).
19. Luo, Y., “Evaluating the State of the Art in Missing Data Imputation for Clinical Data,” *Brief Bioinform* **23**(1), 1-9(2022).
20. Carli, M., Ward, M. H., Metayer, C. and Wheeler, D. C., “Imputation of Below Detection Limit Missing Data in Chemical Mixture Analysis with Bayesian Group Index Regression,” *Int. J. Environ Res. Public Health* **19**(3), 2-14(2022).

21. Jeong, J. S., Garcia-Reyero, N., Burgoon, L., Perkins, E., Park, T. H., Kim, C. H., Roh, J. Y. and Choi, J. H., "Development of Adverse Outcome Pathway for PPARγ Antagonism Leading to Pulmonary Fibrosis and Chemical Selection for Its Validation: ToxCast Database and a Deep Learning Artificial Neural Network Model-Based Approach," *Chem Res Toxicol* **32**(6), 1212-1222(2019).
22. Tiganis, B. E., Burn, L. S., Davis, P. and Hill, A. J., "Thermal Degradation of Acrylonitrile-butadiene-styrene (ABS) Blends," *Polym Degrad Stab* **76**(1), 425-434(2002).
23. Rutkowski, J. V. and Levin, B. C., "Acrylonitrile-Butadiene-Styrene Copolymers (ABS): Pyrolysis and Combustion Products and their Toxicity-A Review of the Literature," *Fire Mater* **10**(1), 93-105(1986).
24. Wojtyła, S., Klama, P. and Baran, T., "Is 3D Printing Safe? Analysis of the Thermal Treatment of Thermoplastics: ABS, PLA, PET, and Nylon," *J. Occup. Environ. Hyg.*, **14**(6), 80-85(2017).
25. Davis, A. Y., Zhang, Q., Wong, J. P. S., Weber, R. J. and Black, M. S., "Characterization of Volatile Organic Compound Emissions from Consumer Level Material Extrusion 3D Printers," *Build Environ* **160**, 106209(2019).
26. Pandit, S., Singh, P., Sinha, M. and Parthasarathi, R., "Integrated QSAR and Adverse Outcome Pathway Analysis of Chemicals Released on 3D Printing Using Acrylonitrile Butadiene Styrene," *Chem Res Toxicol* **34**(2), 355-364(2021).
27. Park, J. H., Jeon, H. J., Oh, Y. S., Park, K. H. and Yoon, C. S., "Understanding Three-dimensional Printing Technology, Evaluation, and Control of Hazardous Exposure Agents," *Journal of Korean Society of Occupational and Environmental Hygiene* **28**(3), 241-256(2018).
28. Kim, S. H., Chung, E. K., Kim, S. D. and Kwon, J. W., "Assessment of Emitted Volatile Organic Compounds, Metals and Characteristic of Particle in Commercial 3D Printing Service Workplace," *Original Article Journal of Korean Society of Occupational and Environmental Hygiene* **30**(2), 153-162(2020).
29. Kim, S. H. and Chung, E. K., "A Study on the Types of Materials and Hazardous Substances Used in 3D Printers," *Korea Occupational Safety and Health Agency* 10-13(2019).
30. Hong, M. K., Jo, J. H., Choi, B. K. and Kim, K. W., *A Study on the Application of OECD Toolbox in Chemical Information* (2018).
31. Mauri, A., Srl, A., Consonni, V., Pavan, M. and Todeschini, R., *DRAGON SOFTWARE: AN EASY APPROACH TO MOLECULAR DESCRIPTOR CALCULATIONS* (n.d.).
32. Rahimi, R., Keshavarz, M. H., and Akbarzadeh, A. R., "Prediction of the Density of Energetic Materials on the Basis of their Molecular Structures," *Central European Journal of Energetic Materials* **13**(1), 73-101(2016).
33. Consonni, V., Todeschini, R. and Pavan, M., "Structure/response Correlations and Similarity/diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors," *J. Chem. Inf. Comput. Sci.*, **42**(3), 682-692(2002).
34. Devinyak, O., Havrylyuk, D. and Lesyk, R., "3D-MoRSE Descriptors Explained," *J. Mol. Graph. Model.*, **54**, 194-203(2014).
35. Stekhoven, D. J. and Bühlmann, P., "Missforest-Non-parametric Missing Value Imputation for Mixed-type Data," *Bioinformatics* **28**(1), 112-118(2012).
36. Marinov, D. and Karapetyan, D., "Hyperparameter Optimisation with Early Termination of Poor Performers," 2019 11th Computer Science and Electronic Engineering (CEEC), Colchester, UK 160-163(2019).
37. Choi, G. C., Kim, W. J. and Koo, J. M., "Investigating the Performance of Machine Learning Methods in Predicting Functional Properties of the Hydrogenase Variants," *Biotechnology and Bioprocess Engineering* **28**(1), 143-151(2023).
38. Moon, J. H., Park, S. W., Rho, S. M. and Hwang, E. J., "Interpretable Short-Term Electrical Load Forecasting Scheme Using Cubist," *Comput Intell Neurosci* **2022**(1), 2-19(2022).
39. Lundberg, S. M., Allen, P. G. and Lee, S. I., "A Unified Approach to Interpreting Model Predictions," 31st Conference on Neural Information Processing Systems (NIPS 2017), December, California 1-10(2017).

Authors

‡The first and second authors have identical collaboration in this research paper.

‡**ChanHyeok Jeong:** MSc. student, Department of Environmental, Science and Engineering College of Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-Gu, Yongin-si, Gyeonggi-do, 17104, Korea; ironwoman98@khu.ac.kr

‡**SangYoun Kim:** Ph.D student, Department of Environmental Science, and Engineering College of Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-Gu, Yongin-si, Gyeonggi-do, 17104, Korea; tkddbs5234@khu.ac.kr

SungKu Heo: Ph.D. student, Department of Environmental Science, and Engineering College of Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-Gu, Yongin-si, Gyeonggi-do, 17104, Korea; energizer@khu.ac.kr

Shahzeb Tariq: Ph.D. student, Department of Environmental Science, and Engineering College of Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-Gu, Yongin-si, Gyeonggi-do, 17104, Korea; shahzebtariq@gmail.com

MinHyeok Shin: Bachelor, Department of Environmental Science and Engineering College of Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-Gu, Yongin-si, Gyeonggi-do, 17104, Korea; smh8991@naver.com.

ChangKyoo Yoo: Professor, Department of Environmental Science and Engineering College of Engineering, Kyung Hee University, 1732, Deogyeong-daero, Giheung-Gu, Yongin-si, Gyeonggi-do, 17104, Korea; ckyoo@khu.ac.kr