(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0077113
(43) 공개일자 2025년05월30일

(51) 국제특허분류(Int. Cl.)
G16C 20/70 (2019.01) *G06N 20/00* (2019.01)
G16C 20/30 (2019.01) *G16C 20/40* (2019.01)
(52) CPC특허분류
G16C 20/70 (2019.02)
G06N 20/00 (2021.08)
(21) 출원번호 10-2023-0164552
(22) 출원일자 2023년11월23일
심사청구일자 2023년11월23일

(71) 출원인
경희대학교 산학협력단
경기도 용인시 기흥구 덕영대로 1732 (서천동, 경희대학교 국제캠퍼스내)
(72) 발명자
유창규
경기도 수원시 영통구 웰빙타운로 20, 8308동 202호 (이의동, 호반가든하임)
정찬혁
경기도 용인시 기흥구 서그내로 23-8, 204호 (서천동, 태림)
(뒷면에 계속)
(74) 대리인
리앤목특허법인

전체 청구항 수 : 총 14 항

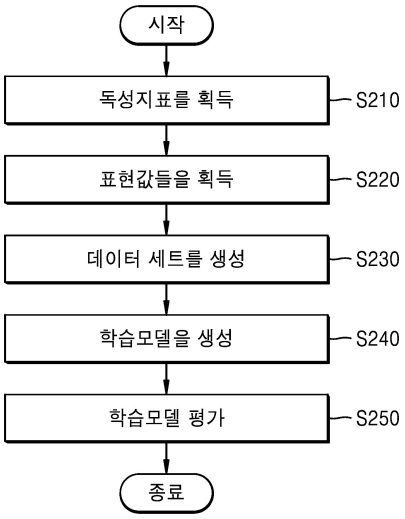
(54) 발명의 명칭 **화학물질의 독성 추정 모델 생성 방법 및 장치**

(57) 요 약

화학물질의 독성 추정 모델 생성 방법 및 장치에 관한 기술이다.

화학물질의 독성 추정 모델 생성 방법에 있어서, 복수개의 화학물질들 각각에 대응하는 구조 정보 및 독성지표 예측 도구에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 독성지표를 획득하는 단계; 상기 복수개의 화학물질들 각각에 대응하는 상기 구조 정보에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 복수개의 분자 표현자들의 값인 표현값들을 획득하는 단계; 상기 복수개의 분자 표현자들 및 상기 독성지표에 기반하여 데이터 세트를 생성하는 단계; 및 상기 데이터 세트에 기반하여 복수개의 분자 표현자로부터 독성지표를 추정하는 학습 모델을 생성하는 단계를 포함할 수 있다.

대 표 도 - 도2



(52) CPC특허분류
G16C 20/30 (2019.02)
G16C 20/40 (2019.02)
(72) 발명자
김상윤
경기도 수원시 영통구 덕영대로1681번길 16, 301호 (영통동)

허성구
서울특별시 강남구 선릉로72길 37 (대치동)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711181969
과제번호	2021R1A2C2007838
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	개인기초연구(과기정통부)
연구과제명	DeepAI 기반 하폐수처리장 스마트워터 자율운전 & 자율설계 원천기술 개발
기 여 율	1/1
과제수행기관명	경희대학교 산학협력단
연구기간	2021.03.01 ~ 2026.02.28

명 세 서

청구범위

청구항 1

화학물질의 독성 추정 모델 생성 방법에 있어서,

복수개의 화학물질들 각각에 대응하는 구조 정보 및 독성지표 예측 도구에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 독성지표를 획득하는 단계;

상기 복수개의 화학물질들 각각에 대응하는 상기 구조 정보에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 복수개의 분자 표현자들의 값인 표현값들을 획득하는 단계;

상기 복수개의 분자 표현자들 및 상기 독성지표에 기반하여 데이터 세트를 생성하는 단계; 및

상기 데이터 세트에 기반하여 복수개의 분자 표현자로부터 독성지표를 추정하는 학습모델을 생성하는 단계를 포함하는 방법.

청구항 2

제1 항에 있어서,

상기 표현값들을 획득하는 단계는,

제1 화학물질에 대응하는 제1 구조 정보에 기반하여 복수개의 제1 표현값들을 산출하는 단계;

제2 화학물질에 대응하는 제2 구조 정보에 기반하여 복수개의 제2 표현값들을 산출하는 단계; 및

산출된 상기 제1 표현값들, 상기 제2 표현값들 및 데이터 보간 알고리즘에 기반하여 결측값을 보간하는 단계를 포함하고,

상기 복수개의 화학물질들은 상기 제1 화학물질 및 상기 제2 화학물질을 포함하는 방법.

청구항 3

제2 항에 있어서,

상기 데이터 세트를 생성하는 단계는,

보간된 상기 제1 표현값들 및 상기 제2 표현값들 사이의 표준편차에 기반하여 상기 복수개의 분자 표현자들 중 적어도 하나를 포함하는 학습 분자 표현자들을 결정하는 단계; 및

상기 학습 분자 표현자들을 피처로하고, 상기 독성지표를 레이블로 하는 데이터 세트를 생성하는 단계를 포함하는 방법.

청구항 4

제2 항에 있어서,

상기 데이터 보간 알고리즘은,

MissForest 알고리즘인 방법.

청구항 5

제1 항에 있어서,

상기 학습모델을 생성하는 단계는,

상기 데이터 세트 및 XGBoost 알고리즘에 기반하여 학습모델을 생성하는 단계를 포함하는 방법.

청구항 6

제1 항에 있어서,

상기 독성지표는, Log P, Log BCF 및 Log Koa 중 적어도 하나를 포함하는 방법.

청구항 7

제1 항에 있어서,

상기 독성지표에 관한 상기 복수개의 분자 표현자의 기여도를 평가하는 단계를 더 포함하는 방법.

청구항 8

컴퓨터 장치에 있어서,

복수개의 화학물질들 각각에 대응하는 구조 정보 및 독성지표 예측 도구에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 독성지표를 획득하고, 상기 복수개의 화학물질들 각각에 대응하는 상기 구조 정보에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 복수개의 분자 표현자들의 값인 표현값들을 획득하며, 상기 복수개의 분자 표현자들 및 상기 독성지표에 기반하여 데이터 세트를 생성하고, 상기 데이터 세트에 기반하여 복수개의 분자 표현자로부터 독성지표를 추정하는 학습모델을 생성하는 프로세서를 포함하는 컴퓨터 장치.

청구항 9

제8 항에 있어서,

상기 프로세서는,

제1 화학물질에 대응하는 제1 구조 정보에 기반하여 복수개의 제1 표현값들을 산출하고, 제2 화학물질에 대응하는 제2 구조 정보에 기반하여 복수개의 제2 표현값들을 산출하며, 산출된 상기 제1 표현값들, 상기 제2 표현값들 및 데이터 보간 알고리즘에 기반하여 결측값을 보간하고,

상기 복수개의 화학물질들은 상기 제1 화학물질 및 상기 제2 화학물질을 포함하는 컴퓨터 장치.

청구항 10

제9 항에 있어서,

상기 프로세서는,

보간된 상기 제1 표현값들 및 상기 제2 표현값들 사이의 표준편차에 기반하여 상기 복수개의 분자 표현자들 중 적어도 하나를 포함하는 학습 분자 표현자들을 결정하고, 상기 학습 분자 표현자들을 피처로하고, 상기 독성지표를 레이블로 하는 데이터 세트를 생성하는 컴퓨터 장치.

청구항 11

제9 항에 있어서,

상기 데이터 보간 알고리즘은,

MissForest 알고리즘인 컴퓨터 장치.

청구항 12

제8 항에 있어서,

상기 프로세서는,

상기 데이터 세트 및 XGBoost 알고리즘에 기반하여 학습모델을 생성하는 컴퓨터 장치.

청구항 13

제8 항에 있어서,

상기 독성지표는, Log P, Log BCF 및 Log Koa 중 적어도 하나를 포함하는 컴퓨터 장치.

청구항 14

제8 항에 있어서,

상기 독성지표에 관한 상기 복수개의 분자 표현자의 기여도를 평가하는 컴퓨터 장치.

발명의 설명

기 술 분 야

[0001] 화학물질의 독성 추정 모델 생성 방법 및 장치에 관한 기술이다.

배 경 기 술

[0002] 과거의 제품 조형 과정은 다양한 단계를 거쳐야 했고, 비교적 오랜 시간에 걸쳐 이루어지곤 했다. 그러나 3D 프린팅의 등장과 그에 따른 신속조형기술의 도입으로, 이러한 시간은 혁신적으로 줄어들었다. 이를 통해 사용자 중심의 제조, 유통 및 물류 서비스가 가능해지면서 3D 프린팅 기술은 전자, 자동차, 항공우주, 의료 등 여러 산업 분야에서 그 중요성이 높아지고 있다.

[0003] 이러한 3D 프린팅에서 사용되는 필라멘트는 다양한 화학물질을 포함하고 있으며, 프린팅 과정 중에 여러 화학물질을 방출할 수 있으나, 이에 대한 구체적인 정보와 안전 지침은 아직 충분하지 않다.

발명의 내용

해결하려는 과제

[0004] 본 명세서의 실시 예는 상술한 문제점을 해결하기 위하여 제안된 것으로 화학물질의 독성 추정 모델 생성 방법 및 장치를 제공한다.

과제의 해결 수단

[0005] 상술한 과제를 달성하기 위한 화학물질의 독성 추정 모델 생성 방법은, 복수개의 화학물질들 각각에 대응하는 구조 정보 및 독성지표 예측 도구에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 독성지표를 획득하는 단계; 상기 복수개의 화학물질들 각각에 대응하는 상기 구조 정보에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 복수개의 분자 표현자들의 값인 표현값들을 획득하는 단계; 상기 복수개의 분자 표현자들 및 상기 독성지표에 기반하여 데이터 세트를 생성하는 단계; 및 상기 데이터 세트에 기반하여 복수개의 분자 표현자로부터 독성지표를 추정하는 학습모델을 생성하는 단계를 포함할 수 있다.

[0006] 상기 표현값들을 획득하는 단계는, 제1 화학물질에 대응하는 제1 구조 정보에 기반하여 복수개의 제1 표현값들을 산출하는 단계; 제2 화학물질에 대응하는 제2 구조 정보에 기반하여 복수개의 제2 표현값들을 산출하는 단계; 및 산출된 상기 제1 표현값들, 상기 제2 표현값들 및 데이터 보간 알고리즘에 기반하여 결측값을 보간하는 단계를 포함하고, 상기 복수개의 화학물질들은 상기 제1 화학물질 및 상기 제2 화학물질을 포함할 수 있다.

[0007] 상기 데이터 세트를 생성하는 단계는, 보간된 상기 제1 표현값들 및 상기 제2 표현값들 사이의 표준편차에 기반하여 상기 복수개의 분자 표현자들 중 적어도 하나를 포함하는 학습 분자 표현자들을 결정하는 단계; 및 상기 학습 분자 표현자들을 피처로하고, 상기 독성지표를 레이블로 하는 데이터 세트를 생성하는 단계를 포함할 수 있다.

[0008] 상기 데이터 보간 알고리즘은, MissForest 알고리즘일 수 있다.

[0009] 상기 학습모델을 생성하는 단계는, 상기 데이터 세트 및 XGBoost 알고리즘에 기반하여 학습모델을 생성하는 단계를 포함할 수 있다.

[0010] 상기 독성지표는, Log P, Log BCF 및 Log Koa 중 적어도 하나를 포함할 수 있다.

[0011] 상기 방법은 상기 독성지표에 관한 상기 복수개의 분자 표현자의 기여도를 평가하는 단계를 더 포함할 수 있다.

[0012] 상술한 과제를 달성하기 위한 컴퓨터 장치는, 복수개의 화학물질들 각각에 대응하는 구조 정보 및 독성지표 예

측 도구에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 독성지표를 획득하고, 상기 복수개의 화학물질들 각각에 대응하는 상기 구조 정보에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 복수개의 분자 표현자들의 값인 표현값들을 획득하며, 상기 복수개의 분자 표현자들 및 상기 독성지표에 기반하여 데이터 세트를 생성하고, 상기 데이터 세트에 기반하여 복수개의 분자 표현자로부터 독성지표를 추정하는 학습모델을 생성하는 프로세서를 포함할 수 있다.

[0013] 상기 프로세서는, 제1 화학물질에 대응하는 제1 구조 정보에 기반하여 복수개의 제1 표현값들을 산출하고, 제2 화학물질에 대응하는 제2 구조 정보에 기반하여 복수개의 제2 표현값들을 산출하며, 산출된 상기 제1 표현값들, 상기 제2 표현값들 및 데이터 보간 알고리즘에 기반하여 결측값을 보간하고, 상기 복수개의 화학물질들은 상기 제1 화학물질 및 상기 제2 화학물질을 포함할 수 있다.

[0014] 상기 프로세서는, 보간된 상기 제1 표현값들 및 상기 제2 표현값들 사이의 표준편차에 기반하여 상기 복수개의 분자 표현자들 중 적어도 하나를 포함하는 학습 분자 표현자들을 결정하고, 상기 학습 분자 표현자들을 피처로 하고, 상기 독성지표를 레이블로 하는 데이터 세트를 생성할 수 있다.

[0015] 상기 데이터 보간 알고리즘은, MissForest 알고리즘일 수 있다.

[0016] 상기 프로세서는, 상기 데이터 세트 및 XGBoost 알고리즘에 기반하여 학습모델을 생성할 수 있다.

[0017] 상기 독성지표는, Log P, Log BCF 및 Log Koa 중 적어도 하나를 포함할 수 있다.

[0018] 상기 독성지표에 관한 상기 복수개의 분자 표현자의 기여도를 평가할 수 있다.

발명의 효과

[0019] 본 발명의 실시 예에 따르면 화학물질의 분자 표현자에 기반하여 독성지표를 추정할 수 있다.

도면의 간단한 설명

[0020] 도 1은 3D 프린팅 소재 별 유해물질 발생 메커니즘을 개략적으로 설명하기 위한 도면이다.

도 2는 본 발명의 실시 예에 따른 컴퓨터 장치의 독성 예측 모델을 생성하는 동작에 관한 순서도다.

도 3은 본 발명의 실시 예에 따른 컴퓨터 장치의 표현값을 획득하는 동작에 관한 순서도다.

도 4는 본 발명의 실시 예에 따른 컴퓨터 장치의 결측값을 보간하는 동작을 설명하기 위한 도면이다.

도 5는 본 발명의 실시 예에 따른 컴퓨터 장치의 데이터 세트를 생성하는 동작에 관한 순서도이다.

도 6a 내지 도 6c는 본 발명의 실시 예에 따른 분자 표현자의 기여도를 설명하기 위한 도면이다.

도 7은 본 발명의 실시 예에 따른 컴퓨터 장치의 블록 구성을 개략적으로 나타내는 블록도이다.

발명을 실시하기 위한 구체적인 내용

[0021] 본 발명은 다양한 변환을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세한 설명에 상세하게 설명하고자 한다. 본 발명의 효과 및 특징, 그리고 그것들을 달성하는 방법은 도면과 함께 상세하게 후술되어 있는 실시예들을 참조하면 명확해질 것이다. 그러나 본 발명은 이하에서 개시되는 실시예들에 한정되는 것이 아니라 다양한 형태로 구현될 수 있다.

[0022] 이하, 첨부된 도면을 참조하여 본 발명의 실시예들을 상세히 설명하기로 하며, 도면을 참조하여 설명할 때 동일하거나 대응하는 구성 요소는 동일한 도면부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다.

[0023] 이하의 실시예에서, 제1, 제2 등의 용어는 한정적인 의미가 아니라 하나의 구성 요소를 다른 구성 요소와 구별하는 목적으로 사용되었다.

[0024] 이하의 실시예에서, 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다.

[0025] 이하의 실시예에서, 포함하다 또는 가지다 등의 용어는 명세서상에 기재된 특징, 또는 구성요소가 존재함을 의미하는 것이고, 하나 이상의 다른 특징들 또는 구성요소가 부가될 가능성을 미리 배제하는 것은 아니다.

[0026] 도면에서는 설명의 편의를 위하여 구성 요소들이 그 크기가 과장 또는 축소될 수 있다. 예컨대, 도면에서 나타난 각 구성의 크기 및 두께는 설명의 편의를 위해 임의로 나타내었으므로, 본 발명이 반드시 도시된 바에 한정

되지 않는다.

[0027] 어떤 실시예가 달리 구현 가능한 경우에 특정한 공정 순서는 설명되는 순서와 다르게 수행될 수도 있다. 예를 들어, 연속하여 설명되는 두 공정이 실질적으로 동시에 수행될 수도 있고, 설명되는 순서와 반대의 순서로 진행될 수 있다. 이때, 본 실시예에서 사용되는 '~부'라는 용어는 소프트웨어 또는 하드웨어 구성요소를 의미하며, '~부'는 어떤 역할들을 수행한다. 그렇지만 '~부'는 소프트웨어 또는 하드웨어에 한정되는 의미는 아니다. '~부'는 어드레싱할 수 있는 저장 매체에 있도록 구성될 수도 있고 하나 또는 그 이상의 프로세서들을 재생시키도록 구성될 수도 있다.

[0028] 하기에서 본 개시를 설명함에 있어 관련된 공지 기능 또는 구성에 대한 구체적인 설명이 본 개시의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는 그 상세한 설명을 생략할 것이다.

[0029] 본 개시에 따른 '독성지표'란 특정 물질이 생물체나 환경에 얼마나 유해한지를 수치적, 또는 질적으로 평가하기 위한 지표를 지시할 수 있다. 예컨대 독성지표는 물질의 지방성과 수성 간의 분배 계수를 로그 스케일로 나타낸 Log P(Octanol-water Partition coefficient), 물생물에 의해 흡수된 물질의 농도와 물 중의 물질 농도 간의 비율을 로그 스케일로 나타낸 Log BCF(Bioconcentration Factor) 및 화합물의 지방성과 공기 중의 분배 계수를 로그 스케일로 나타낸 Log Koa(Octanol-air partition coefficient)중 적어도 하나를 포함할 수 있다.

[0030] 본 개시에 따른 분자 표현자(molecular descriptor)란 분자의 구조적, 물리적, 화학적 특성을 수치적, 기호적, 혹은 그래픽적으로 표현하는 정보를 지시할 수 있다. 예컨대 분자량, 원자 수 등의 전체 분자에 대한 정보, 특정 유형의 원자나 결합의 수, 분자 내의 특정 원자 간의 최단 경로, 원자 쌍 간의 최단 경로 길이, 특정 원자 쌍 간의 최대 경로 길이, 3D 모델링에서 파생된 정보, 분자의 형태와 관련된 파라미터, 분자의 표면 영역, 전하 분포, 및 다양한 원자 간의 거리 및 각도 중 적어도 하나를 포함할 수 있다.

[0031] 본 발명에서 언급되는 '학습 모델'은 데이터로부터 특정 패턴이나 구조를 학습하거나 이해하기 위해 사용되는 모든 형태의 알고리즘 또는 방법론을 포함할 수 있다. 이는 전통적인 통계 모델, 머신 러닝 모델, 딥러닝 모델, 강화학습 모델, 그리고 그외의 학습 방법론을 포함하며, 이에 한정되지 않는다.

[0032] 즉 학습 모델은 회귀 모델, 의사결정트리, 랜덤 포레스트, 서포트 벡터 머신, K-최근접 이웃, 나이브 베이즈, 클러스터링 알고리즘 등의 머신 러닝 모델뿐만 아니라, 신경망, 컨볼루션 신경망, 순환 신경망, Transformer 기반 신경망, GANs(Generative Adversarial Networks), 오토인코더 등의 딥러닝 모델도 포함할 수 있다. '학습 모델'은 특정 입력에 대한 출력을 예측하거나 분류하는 데 사용되는 학습된 파라미터나 가중치 집합을 지시할 수 있으며. 이 모델은 지도 학습, 비지도 학습, 반지도 학습, 강화 학습 등의 방법을 통해 학습될 수 있다. 또한, 이는 단일 모델 뿐 아니라 앙상블 모델, 멀티-모달 모델, 트랜스퍼 학습을 통한 모델 등 다양한 학습 방식과 구조를 포함할 수도 있다. 이러한 학습 모델은 입력에 대한 출력을 예측하는 컴퓨터 장치와 별개의 컴퓨터 장치에서 사전 학습되어 타 컴퓨터 장치에서 사용될 수도 있다.

[0033] 도 1은 3D 프린팅 소재 별 유해물질 발생 메커니즘을 개략적으로 설명하기 위한 도면이다.

[0034] 3D 프린터에 사용되는 원재료는 플라스틱 폴리머, 천연재료, 세라믹 금속, 콘크리트 재질 등으로 다양 하지만, 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 열분해되며 발생하는 화학물질은 사용자의 안전과 건강에 악영향을 줄 수 있다.

[0035] 도 1을 참고하면 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 3D 프린팅 과정에서 사용된 필라멘트의 종류와 운전 조건에 따라 다양하고 신규한 화학물질이 발생할 수 있다.

[0036] 이러한 화학물질에 대해서는 물질의 조성 및 성분 별 데이터가 부족하고, 독성지표 또한 불분명하게 알려져 있다.

[0037] 도 2는 본 발명의 실시 예에 따른 컴퓨터 장치의 독성 예측 모델을 생성하는 동작에 관한 순서도다.

[0038] 도 2를 참고하면, 컴퓨터 장치는 S210 단계에서, 복수개의 화학물질들 각각에 대응하는 구조 정보 및 독성지표 예측 도구에 기반하여 복수개의 화학물질들 각각에 대응하는 독성지표를 획득할 수 있다. 예컨대 컴퓨터 장치는 MOL, SDF와 같은 구조 정보 및 OECD QSAR(Quantitative Structure-Activity Relationship) Toolbox와 같은 예

측 도구를 이용해서 복수개의 화학물질 각각의 독성지표를 추정할 수 있다.

[0039] 본 발명의 일 실시 예에 따른 컴퓨터 장치는 S220 단계에서, 복수개의 화학물질들 각각에 대응하는 구조 정보에 기반하여 복수개의 화학물질들 각각에 대응하는 복수개의 분자 표현자들의 값인 표현값들을 획득할 수 있다. 예컨대 컴퓨터 장치는 분자 표현자를 산출하기 위해 DRAGON 6 software를 이용하여 구조 정보로부터 해당 화학물질의 다양한 화학적, 물리적 특성을 나타내는 분자 표현자의 값들을 산출할 수 있다.

[0040] 본 발명의 일 실시 예에 따른 컴퓨터 장치는 S230 단계에서, 복수개의 분자 표현자들 및 독성지표에 기반하여 데이터 세트를 생성할 수 있다. 컴퓨터 장치는 분자 표현자들의 결측값을 보간하고, 복수개의 분자 표현자들 중에서 학습에 이용하기 위한 학습 분자 표현자들을 결정하여 테이터 세트를 생성할 수 있다.

[0041] 본 발명의 일 실시 예에 따른 컴퓨터 장치는 S240 단계에서 데이터 세트에 기반하여 복수개의 분자 표현자로부터 독성지표를 추정하는 학습모델을 생성할 수 있다. 컴퓨터 장치는 생성된 데이터 세트를 활용하여 머신러닝 알고리즘을 학습시켜 독성지표를 추정하는 모델을 생성할 수 있다. 예컨대 컴퓨터 장치는 훈련 데이터와 테스트 데이터로 데이터 세트를 구분하고, 선택된 알고리즘을 사용하여 훈련 데이터로 모델을 학습할 수 있다. 이때 컴퓨터 장치는 학습 모델의 하이퍼파라미터 조정하는 방식으로 학습 모델을 생성하고, 테스트 데이터를 사용하여 모델의 일반화 성능을 평가할 수 있다.

[0042] 예를들면 컴퓨터 장치는 그래디언트 부스팅 방법에 기반한 XGBoost를 이용하여, 결정 트리를 연속적으로 학습해 나가면서 학습 모델의 성능을 개선할 수 있다. 컴퓨터 장치는 학습률(learning rate), 트리의 최대 깊이(max depth), 사용될 트리의 수(n_estimators)의 하이퍼파라미터를 설정하고, 훈련 데이터를 사용하여 학습 모델을 생성할 수 있다. 이러한 과정을 통해 컴퓨터 장치는 각 트리가 이전 트리의 잔차(residual) 또는 오차를 학습하도록 할 수 있다.

[0043] 본 발명의 일 실시 예에 따른 컴퓨터 장치는 S250 단계에서 Tree-SHAP을 활용하여 독성 예측 모델의 결과를 해석하고, 각 분자 표현자가 독성지표 예측에 얼마나 큰 영향을 미치는지에 관한 기여도를 평가할 수 있다.

[0044] 도 3은 본 발명의 실시 예에 따른 컴퓨터 장치의 표현값을 획득하는 동작에 관한 순서도이고, 도 4는 본 발명의 실시 예에 따른 컴퓨터 장치의 결측값을 보간하는 동작을 설명하기 위한 도면이다. 도 3에서의 동작은 도 2에서의 S220 단계에서의 동작과 대응한다.

[0045] 도 3을 참고하면, 컴퓨터 장치는 S310 단계에서, 제1 화학물질에 대응하는 제1 구조 정보에 기반하여 복수개의 제1 표현값들을 산출할 수 있다. 예컨대 컴퓨터 장치는 DRAGON 6 software를 이용하여 제1 구조 정보로부터 제1 화학물질의 분자 표현자들의 값들인 제1 표현값들을 산출할 수 있다.

[0046] 본 발명의 일 실시 예에 따른 컴퓨터 장치는 S320 단계에서, 제2 화학물질에 대응하는 제2 구조 정보에 기반하여 복수개의 제2 표현값들을 산출할 수 있다. 예컨대 컴퓨터 장치는 DRAGON 6 software를 이용하여 제2 구조 정보로부터 제2 화학물질의 분자 표현자들의 값들인 제2 표현값들을 산출할 수 있다.

[0047] 본 발명의 일 실시 예에 따른 컴퓨터 장치는 S330 단계에서, 산출된 제1 표현값들, 제2 표현값들 및 데이터 보간 알고리즘에 기반하여 결측값을 보간할 수 있다. 이러한 데이터 보간 알고리즘은 비모수적 결측치 보간 알고리즘으로서, 예컨대 MissForest 알고리즘일 수 있다. 도 4에서 도시된 바와 같이 표현값들에 관한 데이터는, 구조 정보를 행(410)으로 하고, 복수개의 분자 표현자들을 열(420)로 하는 표로 나타낼 수 있다. 컴퓨터 장치는 결측값(431)을 그 변수의 평균 또는 최빈값으로 대체하고, 각 변수에 대하여 순서대로 진행하여, 그 변수의 결측값을 반응 변수로 하고, 나머지 변수들을 예측 변수로 하여 무작위 포레스트 모델을 학습할 수 있다. 컴퓨터 장치는 모든 변수에 대해 반복을 진행하여 결측값(431)을 업데이트하고, 업데이트 된 결측값(431)이 변동이 없거나 임계값 이하의 변동을 가지는 경우 결측값 보간을 종료할 수 있다.

[0048] 도 5는 본 발명의 실시 예에 따른 컴퓨터 장치의 데이터 세트를 생성하는 동작에 관한 순서도이다. 도 5에서의 동작은 도 2에서의 S230 단계에서의 동작과 대응한다.

[0049] 도 5를 참고하면 컴퓨터 장치는 S510 단계에서 보간된 상기 제1 표현값들 및 제2 표현값들 사이의 표준편차에 기반하여 복수개의 분자 표현자들 중 적어도 하나를 포함하는 학습 분자 표현자들을 결정할 수 있다. 예컨대 컴퓨터 장치는 표현값이 미리 설정된 임계값 이상의 표준편차를 가지는 분자 표현자를 학습 분자 표현자로 결정하여, 복수개의 분자 표현자들 중에서 학습 분자 표현자들을 결정할 수 있다.

[0050] 본 발명의 일 실시 예에 따른 컴퓨터 장치는 S520 단계에서, 학습 분자 표현자들을 피처로하고, 독성지표를 레이블로 하는 데이터 세트를 생성할 수 있다. 컴퓨터 장치는 이후 생성된 데이터 세트에 기반하여 복수개의 분자

표현자로부터 독성지표를 추정하는 학습모델을 생성할 수 있다.

[0051] 도 6a 내지 도 6c는 본 발명의 실시 예에 따른 분자 표현자의 기여도를 설명하기 위한 도면이다. 구체적으로 도 6a 내지 도 6c은 Log BCF, Log Koa 및 Log P와 같은 독성지표에 대한 분자표현자의 섀플리값을 도시한다.

[0052] 도 6a을 참고하면 Log BCF 데이터에서 ALOGP와 CIC0 분자표현자가 독성 예측에 높은 영향력를 보였으며, 두 분자표현자 모두 섀플리 값이 양수이므로 ALOGP와 CIC0 증가함에 따라 Log BCF 도 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. CIC0은 상호보완적인 정보 내용 지수(0차 이웃 대칭)를 나타내며, 이는 분자 내의 연결성과 작용기를 설명하며 분자 형태 및 녹는점과 연관된다.

[0053] 도 6b를 참고하면, GMTIV의 섀플리 값 분포가 가장 큰 것으로 확인되었고, SMTIV 역시 log KOA에 큰 영향을 주는 분자표현자임을 확인할 수 있다. GMTIV는 Gutman 분자 위상 지수이고, SMTIV는 Schultz 분자 위상 지수이다. 두 지표의 섀플리 값은 모두 음수로, 이들의 값이 높을수록 log KOA 값은 낮아진다.

[0054] 도 6c를 참고하면 Log P에 대한 분자표현자의 분포가 도시된다. ALOGP와 함께 MLOGP가 Log P와 높은 연관성을 보인다. 여기서 모든 섀플리 값은 양수로, 이는 각 분자표현자와 Log P의 양의 상관관계를 나타낸다. Log P 값을 추정하는데 영향이 큰 분자표현자로는 SpAD_B(s), SpMax6_Bh(s), SpMax2_Bh(s) 등이 있다. Log BCF와 비교하면, ALOGP의 섀플리 값은 두 지표 모두에서 높게 나타났으나, 각 분자표현자의 비율에서는 차이가 있다.

[0055] 도 7은 본 발명의 실시 예에 따른 컴퓨터 장치의 블록 구성을 개략적으로 나타내는 블록도이다.

[0056] 도 6을 참고하면 컴퓨터 장치(700)는 프로세서(720), 메모리(710)를 포함하는 것으로 도시하였으나 반드시 이에 한정되는 것이 아니다. 예를 들면 프로세서(720), 메모리(710)는 각각 물리적으로 독립한 하나의 구성부로서 존재할 수 있다.

[0057] 메모리(710)는 컴퓨터 장치(700)에서 프로세서(720)의 처리 또는 제어를 위한 프로그램 등 전반의 동작을 위한 다양한 데이터를 저장할 수 있다. 메모리(710)는 구동되는 다수의 응용 프로그램(Application program), 컴퓨터 장치(700)의 동작을 위한 데이터들, 명령어들을 저장할 수 있다. 메모리(710)는 프로세서(720)에 포함된 롬(ROM), 램(RAM) 등의 내부 메모리로 구현되거나, 프로세서(720)와 별도의 메모리로 구현될 수 있다.

[0058] 프로세서(720)는 컴퓨터 장치(700)를 전반적으로 제어하기 위한 구성일 수 있다. 구체적으로 프로세서(720)는 도 2내지 5에서의 컴퓨터 장치(700)의 동작을 제어할 수 있다. 이러한 프로세서(720)는 CPU, 램(RAM), 롬(ROM), 시스템 버스 등을 포함할 수 있다. 프로세서(720)는 단일 CPU 또는 복수의 CPU(또는 DSP, SoC)로 구현될 수 있다. 일 실시예로, 프로세서(720)는 디지털 신호를 처리하는 디지털 시그널 프로세서(digital signal processor(DSP), 마이크로프로세서(microprocessor), TCON(Time controller)으로 구현될 수 있다. 다만, 이에 한정되는 것은 아니며, 중앙처리장치(central processing unit(CPU)), MCU(Micro Controller Unit), MPU(micro processing unit), 컨트롤러(controller), 어플리케이션 프로세서(application processor(AP)), 또는 커뮤니케이션 프로세서(communication processor(CP)), ARM 프로세서 중 하나 또는 그 이상을 포함하거나, 해당 용어로 정의될 수 있다. 또한, 프로세서(720)는 프로세싱 알고리즘이 내장된 SoC(System on Chip), LSI(large scale integration)로 구현될 수도 있고, FPGA(Field Programmable gate array) 형태로 구현될 수도 있다.

[0059] 프로세서(720)는 예를 들면 복수개의 화학물질들 각각에 대응하는 구조 정보 및 독성지표 예측 도구에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 독성지표를 획득하고, 상기 복수개의 화학물질들 각각에 대응하는 상기 구조 정보에 기반하여 상기 복수개의 화학물질들 각각에 대응하는 복수개의 분자 표현자들의 값인 표현값들을 획득하며, 상기 복수개의 분자 표현자들 및 상기 독성지표에 기반하여 데이터 세트를 생성하고, 상기 데이터 세트에 기반하여 복수개의 분자 표현자로부터 독성지표를 추정하는 학습모델을 생성할 수 있다.

[0060] 한편, 본 명세서와 도면에는 본 발명의 바람직한 실시 예에 대하여 게시하였으며, 비록 특정 용어들이 사용되었으나, 이는 단지 본 발명의 기술 내용을 쉽게 설명하고 발명의 이해를 돕기 위한 일반적인 의미에서 사용된 것이지, 본 발명의 범위를 한정하고자 하는 것은 아니다. 여기에 개시된 실시 예 외에도 본 발명의 기술적 사상에 바탕을 둔 다른 변형 예들이 실시 가능하다는 것은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에게 자명한 것이다.

[0061] 예를 들면, 일련의 동작(act)을 설명하였지만, 본 발명의 원리와 일치하는 다른 실시예에서는 이 동작들의 순서가 변경될 수 있다. 또, 비종속적인(non-dependent) 동작들은 병렬로 실행될 수 있다.

부호의 설명

[0062] 410: 행

420: 열

431: 결측값

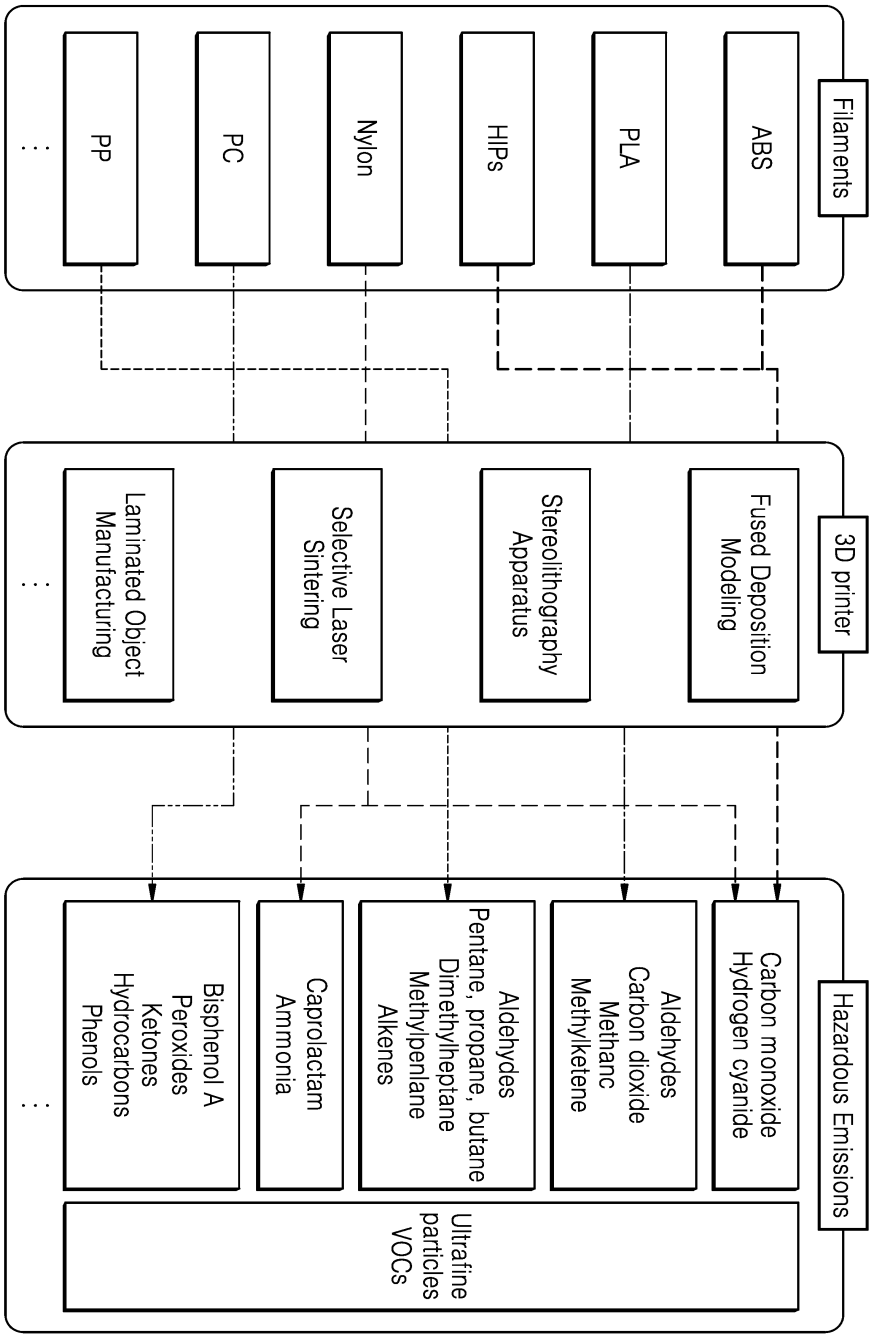
700: 컴퓨터 장치

710: 메모리

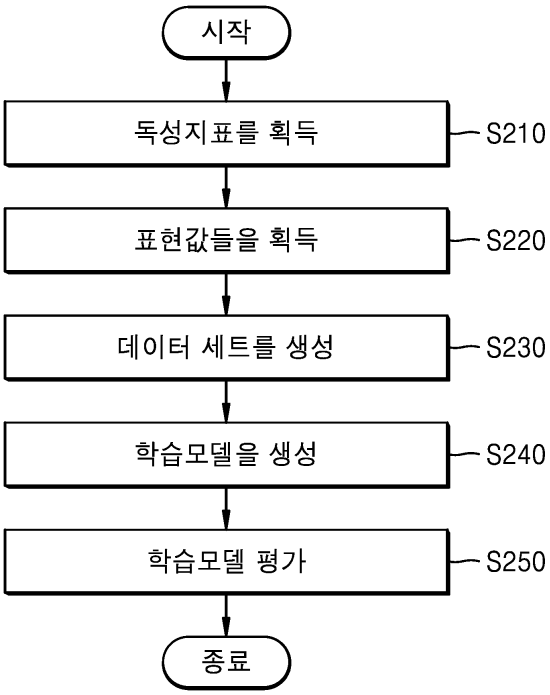
720: 프로세서

도면

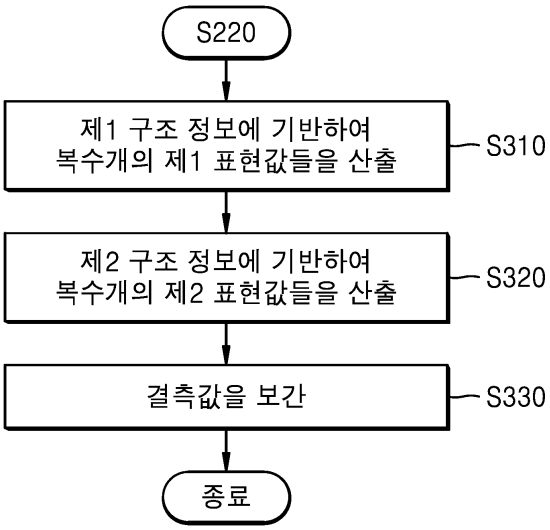
도면1



도면2


시작
독성지표를 획득
S210
표현값들을 획득
S220
데이터 세트를 생성
S230
학습모델을 생성
S240
학습모델 평가
S250
종료

도면3

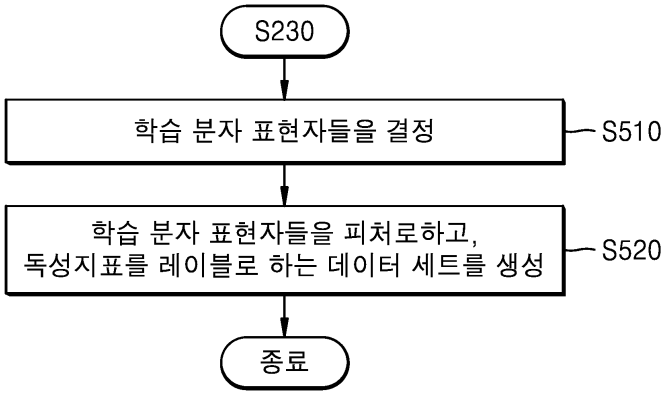

S220
제1 구조 정보에 기반하여
복수개의 제1 표현값들을 산출
S310
제2 구조 정보에 기반하여
복수개의 제2 표현값들을 산출
S320
결측값을 보간
S330
종료

도면4

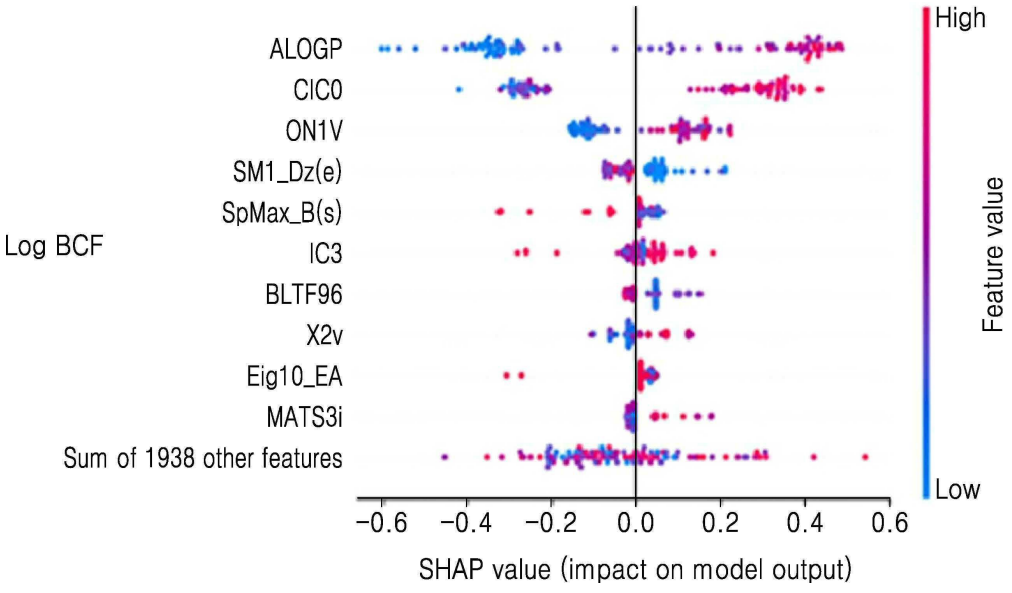
CAS no.	Structure 1	…	Structure n
71-36-3	NaN (431)	…	0.48
78-83-1	2.54	…	0.3
93-55-0	NaN (431)	…	2.86
…	…	…	…
97-88-1	6.74	…	NaN (431)
57-10-3	NaN (431)	…	1.95

410 (Structure 1 … Structure n)
420 (93-55-0)

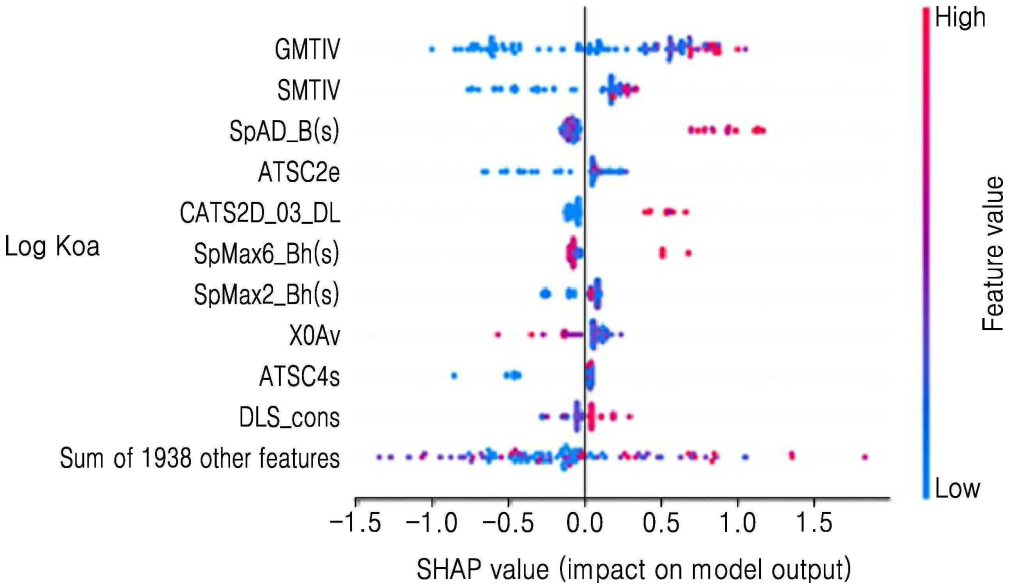
도면5



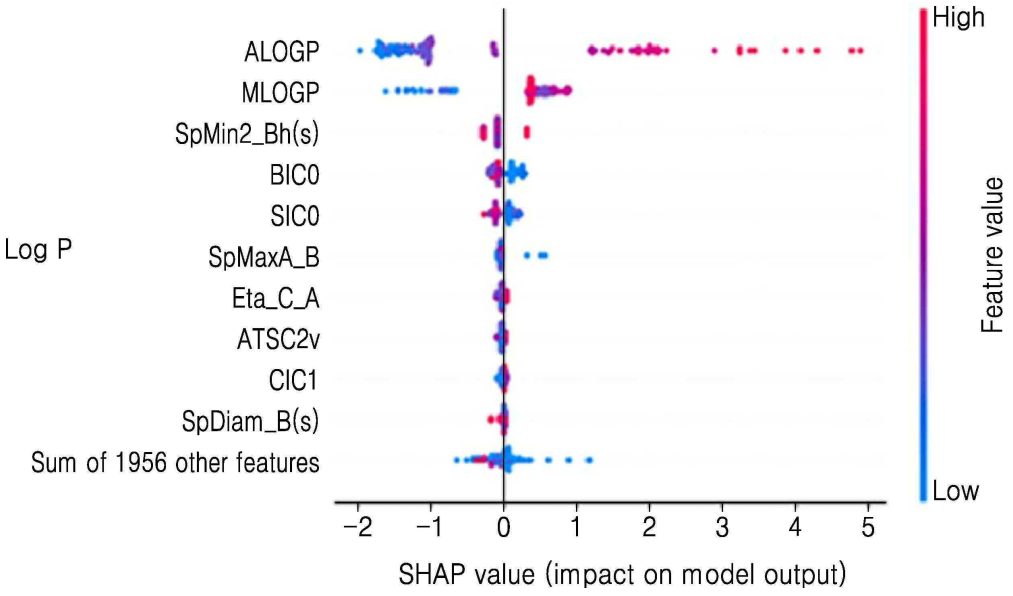
도면6a



도면6b



도면6c



도면7

