韓國環境保健學會誌
Journal of Environmental Health Sciences
Original Article

pISSN 1738-4087 eISSN 2233-8616
J Environ Health Sci. 2021; 47(5): 462-471
https://doi.org/10.5668/JEHS.2021.47.5.462

화학물질 독성 빅데이터와 심층학습 모델을 활용한 내분비계 장애물질 선별 방법- 세정제품과 세탁제품을 중심으로

이인혜[1], 이수진[2], 지경희[2]*

[1]용인대학교 자연과학연구소, [2]용인대학교 일반대학원 환경보건학과

A Screening Method to Identify Potential Endocrine Disruptors Using Chemical Toxicity Big Data and a Deep Learning Model with a Focus on Cleaning and Laundry Products

Inhye Lee[1], Sujin Lee[2], and Kyunghee Ji[2]*

[1]Institute of Natural Science, Yongin University, [2]Department of Environmental Health, Graduate School at Yongin University

ABSTRACT

Background: The number of synthesized chemicals has rapidly increased over the past decade. For many chemicals, there is a lack of information on toxicity. With the current movement toward reducing animal testing, the use of toxicity big data and deep learning could be a promising tool to screen potential toxicants.

Objectives: This study identified potential chemicals related to reproductive and estrogen receptor (ER)-mediated toxicities for 1135 cleaning products and 886 laundry products.

Methods: We listed chemicals contained in cleaning and laundry products from a publicly available database. Then, chemicals that potentially exhibited reproductive and ER-mediated toxicities were identified using the European Union Classification, Labeling and Packaging classification and ToxCast database, respectively. For chemicals absent from the ToxCast database, ER activity was predicted using deep learning models.

Results: Among the 783 listed chemicals, there were 53 with potential reproductive toxicity and 310 with potential ER-mediated toxicity. Among the 473 chemicals not tested with ToxCast assays, deep learning models indicated that 42 chemicals exhibited ER-mediated toxicity. A total of 13 chemicals were identified as causing reproductive toxicity by reacting with the ER.

Conclusions: We demonstrated a screening method to identify potential chemicals related to reproductive and ER-mediated toxicities utilizing chemical toxicity big data and deep learning. Integrating toxicity data from *in vivo*, *in vitro*, and deep learning models may contribute to screening chemicals in consumer products.

Key words: Big data, chemical products, deep learning, predictive toxicity, ToxCast

Received August 25, 2021
Revised October 6, 2021
Accepted October 6, 2021

Highlights:

- 783 chemicals contained in cleaning and laundry products were listed from Ecolife database.
- Chemicals with reproductive and ER-mediated toxicities were identified from EU CLP and ToxCast.
- The reactivity with the ER of ToxCast untested chemicals was predicted using deep learning models.
- A total of 13 chemicals were identified as causing reproductive toxicity by reacting with the ER.

***Corresponding author:**
Department of Occupational and Environmental Health, Yongin University, Yongindaehak-ro 134, Yongin 17092, Republic of Korea
Tel: +82-31-8020-2747
Fax: +82-31-8020-2886
E-mail: kyungheeji@yongin.ac.kr

I. 서 론

생활화학제품에는 제품의 용도에 부합하는 기능을 발휘하기 위해 수많은 화학물질이 원료로 함유되어 있으며, 소비자는 다양한 생활화학제품을 사용하면서 제품에 포함된 화학물질에 노출되고 있다.[1] 생활화학제품 사용으로 인한 인체의 화학물질 부하가 매년 높아지고 있으며,[2] 이러한 화학물질이 사람과 생태계에 미치는 영향에 대한 우려도 커지고 있다.[3] 「화학물질의 등록 및 평가 등에 관한 법률(약칭: 화학물질등록평가법)」이 개정(2018년 3월)되고 「생활화학제품 및 살생물제의 안전관리에 관한 법률(약칭: 화학제품안전법)」이 제정(2019년 1월 시행)되면서 안전확인대상 생활화학제품의 관리대상이 가정, 사무실, 다중이용시설에서 사용하는 제품까지로 확대되었다.[4] 「화학물질등록평가법」에 따른 위해우려제품 또는 「화학제품안전법」에 따른 안전확인대상 생활화학제품에 대한 연구는 주로 제품 내 함유된 유해물질을 조사하거나[5,6] 일부 물질의 잠재적인 유해성을 평가한[7] 것이 대부분이었다. 방향제나 탈취제의 사용현황[1], 곰팡이 제거제의 노출평가를 진행한 연구[8]도 있다. 그러나 현재까지 진행된 연구는 특정 생활화학제품에 초점을 맞추어 극히 일부 물질군에 대해 제한적으로 조사되었다는 한계가 있다. 또한 생활화학제품에 함유된 화학물질의 독성 및 위해성 평가 자료가 매우 부족하고, 동물을 이용한 전통적인 독성시험(*in vivo*)에는 시간과 비용이 많이 소요되어 제품의 안전한 사용이 어려운 실정이다.

내분비계 장애물질(endocrine disrupting chemical)은 호르몬 수용체(예: 에스트로겐 수용체(estrogen receptor, ER), 안드로겐 수용체(androgen receptor) 등)와의 상호작용을 통해 신체의 정상적인 호르몬 기능에 영향을 주는 체외 화학물질이다.[9] 세정제, 표백제, 합성세제 등 안전확인대상 생활화학제품에서도 프탈레이트류, 알킬페놀류,[10] 이소티아졸리논류[6] 등의 내분비계 장애물질이 검출된 바 있다. 실험동물의 윤리적인 측면에서 동물을 이용한 실험(*in vivo*)을 금지하는 추세와 더불어 내분비계 장애물질을 고속대량으로 스크리닝(high-throughput screening)하는 분석법들이 개발되고 있으며,[11] 그 대표적인 예가 ToxCast[12] 및 Tox21[13]이다. 미국 환경보호청(US Environmental Protection Agency)에서는 에스트로겐성 화학물질을 판단하기 위한 설치류 동물실험의 대체방법으로 18개 시험관 내(*in vitro*) 분석 자료를 통합한 ToxCast ER 모델을 사용하고 있다.[14] 또한 유럽화학물질청(European Chemical Agency)과 유럽식품안전청(European Food Safety Authority)에서는 농약과 살생물제에서 내분비계 장애물질을 확인하기 위해 고속대량 스크리닝 자료를 적극 활용하는 평가 전략을 제시했다.[15]

시험관 내(*in vitro*) 대체시험 방법을 이용한 내분비계 독성평가가 증가하고 있으나, CompTox Chemicals Dashboard에 등록된 88만종 가량의 화학물질 중에서 ToxCast ER 시험자료는 1만종이 채 되지 않는다. 종종 위해성 평가 과정에서 독성자료가 없는 물질을 제외하기도 하는데, 화학물질의 독성 잠재력을 배제하기 때문에 의사 결정이 편향될 수 있다.[16] 이러한 제한점을 극복하기 위해 기계학습(machine-learning), 심층학습(deep-learning)과 같은 인공지능 기법으로 기존의 자료를 학습한 후 부족한 자료를 예측하고, 이를 화학물질 관리에 활용하고 있다.[16-18] 예를 들어, ToxCast 자료가 없을 경우 심층학습 모델을 활용한 예측독성 자료로 독성 데이터 갭을 해결한 연구가 있다.[19,20]

본 연구에서는 사람에 대한 역학연구나 동물 독성시험(*in vivo*)을 토대로 설정되는 독성 분류 시스템에 대체시험 자료(시험관 내(*in vitro*) 독성 빅데이터 자료와 인공지능을 활용한 예측독성 자료)를 추가하여 분자수준의 초기현상부터 최종 독성영향까지 고려한 화학물질 선별 시스템의 구동 가능성을 살펴보고자 하였다. 안전확인대상 생활화학제품 중 일반 소비자들이 가장 빈번하게 노출되는 세정제품(세정제, 제거제)과 세탁제품(합성세제, 표백제, 섬유유연제)에 함유된 성분들을 중심으로 *in vivo-in vitro-in silico* 자료를 활용하여 에스트로겐 수용체와 반응하여 생식·수유독성을 유발할 수 있는 화학물질들을 선별하였다.

II. 연구방법

1. 연구대상 제품 및 성분 자료 수집

2020년 7월 기준 생활환경안전정보시스템 초록누리 홈페이지[21]에서 안전확인대상 생활화학제품 3,135종의 생활화학제품명과 성분 자료를 수집하였다. 수집된 제품 중 영업비밀 차원에서 성분이 공개되지 않은 제품은 추후 분석에서 제외하였다. 또한 연구대상 제품의 성분수가 부풀려지는 것을 방지하기 위해 동일 회사에서 출고한 제품의 명칭이 동일하거나(정확한 중복) 크기만 다르게 판매되는 제품(부분 중복)은 제외하였다. 성분정보에 관련이 없는 문구(예: 유기농)나 용도에 관련된 문구(예: 향료, 용제 등)가 기재된 것도 추후 분석에서 제외하였다.

2. 성분명과 화학물질 식별자 연결

제품 안에 포함된 성분명을 정리하는 과정에서 동일한 화학물질이 서로 다른 명칭(국문 또는 영문)으로 나타남을 확인하였고, 화학물질의 이명을 하나로 통일하는 작업을 진행하였다. 이 과정에서 국립환경과학원의 화학물질정보시스템,[22] 유럽화학물질청의 Information on Chemicals,[23] 미국 환경보호청의 CompTox Chemistry Dashboard,[24] 미국 국립의학도서관의 PubChem[25]을 활용하였다. 각각의 데이터베이스에는 많은 화학물질의 동의어 목록이 있으며, CAS 등록번호를 포함한 다양한 화학물질 식별자(예: IUPAC International Chemical

Identifiers (InChI), simplified molecular-input line-entry system (SMILES), PubChem CID, DTXSID 등)가 수록되어 있다. 독성자료를 수집하거나 심층학습 모델 구동 시 활용하기 위해 화학물질의 성분명과 CAS 등록번호, PubChem CID, SMILES, DTXSID를 연결하여 정리하였다.

3. 독성자료 수집

연구대상 물질 중 생식독성을 유발하는 물질을 선별하기 위해 EU CL Inventory[23]에서 생식독성(reproductive toxicity category 1A, 1B, 2) 또는 수유독성(lactation)으로 분류된 물질을 확인하였다(Fig. 1). 대상물질 중에서 에스트로겐 수용체와 결합(binding)하거나 활성화(activation)하는 물질을 확인하기 위해 CompTox Chemicals Dashboard[24]에서 ToxCast *in vitro* assay자료를 확인하였다. 각 물질을 검색한 후 bioactivity 항목에서 18개 ER assay에 대한 Hitcall (activity) 자료를 살펴보고 1개 이상의 assay에서 active로 표시된 물질을 선별하였다(Fig. 1).

4. 심층학습 모델 구축 및 예측독성

ToxCast *in vitro* assay 자료를 확인할 수 없는 물질에 대해 독성자료의 공백을 채우고자 기존의 연구[19,20]를 참고하여 ToxCast의 ER assay를 토대로 심층학습 인공 신경망 모델을 구축하였다. 이 때 18개 ToxCast assay 중 대상물질에 대한 assay의 활성/비활성 자료가 없는 1개 assay (ATG_ERβ_TRANS2_up)를 제외하였다. ToxCast는 활성(active)과 비활성(inactive)의 비율이 매우 차이가 나며(highly-imbalanced data) 이러할 경우 모델은 소수 클래스(minor class)의 분포를 제대로 학습하지 못하고 어떠한 데이터가 들어와도 다수 클래스(major class)로 분류하는 문제가 발생한다.[26,27] 클래스 불균형을 조정하고 학습의 예측 정확도를 향상시키기 위해 언더 샘플링과 오버 샘플링을 결합한 synthetic minority over-sampling technique and edited nearest neighbor algorithm (SMOTEENN) 기법을 사용하였다.[28] 모델에는 각 화학물질의 성분명, SMILES 코드, 해당 화학물질의 ER 활성 여부를 입력하였으며, 화학물질의 구조는 RDKit[29]를 사용하여 Morgan Fingerprint로 변환하였다. 본 연구에서는 파이썬(Python3) 프로그램의 케라스(Keras)와 텐서 플로우(Tensorflow) 모듈로 심층학습 모델(다층 퍼셉트론(multilayer perceptron) 인공 신경망 모델)을 구축하였다.[30] 이 모델은 단일 입력층(input layer), 리키 렐루(Leaky rectified linear unit, LReLU) 활성화 함수(activation function)를 사용한 은닉층(hidden layer), 시그모이드(sigmoid) 함수를 사용한 단일 출력층(output layer)으로 구성된다. 구축한 모델의 성능은 사

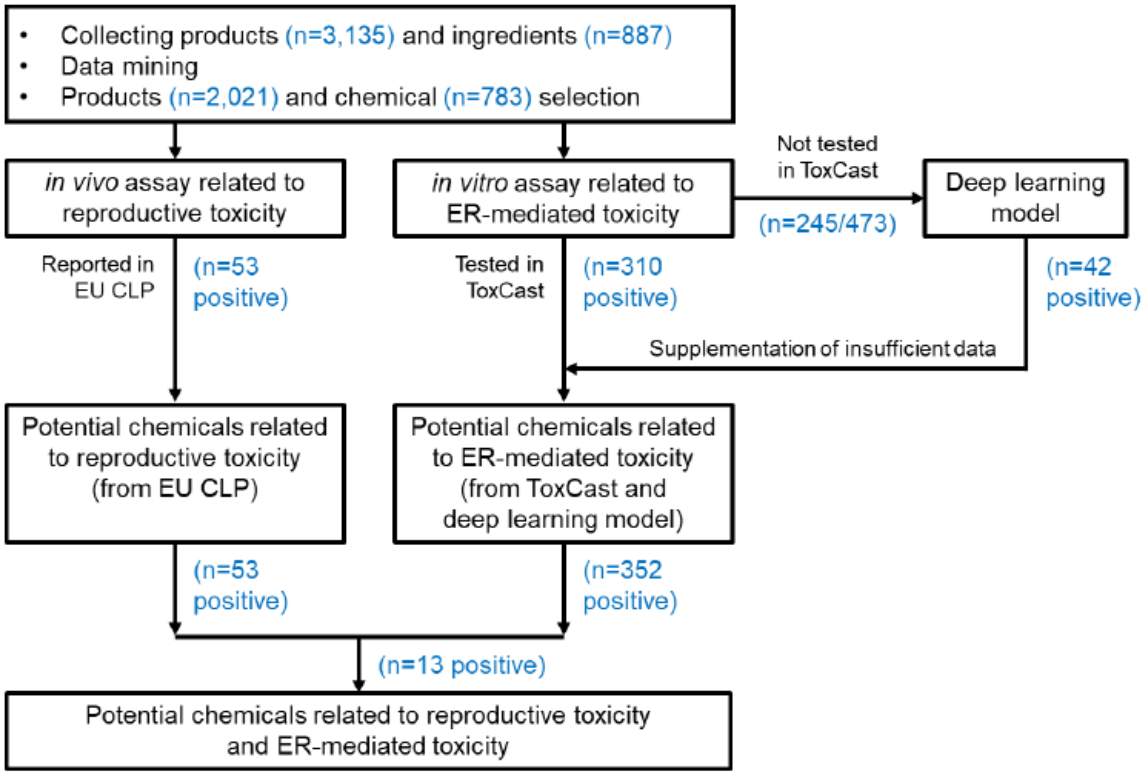

Fig. 1. Workflow for screening potential chemicals related to reproductive and estrogen receptor (ER)-mediated toxicities using EU CLP classification, ToxCast database, and deep learning models

이킷런(Scikit-learn)의 층화된 5중 교차 검증(stratified 5-fold cross validation)으로 평가하였다. 구축된 모델을 활용하여 ToxCast *in vitro* assay 자료를 확인할 수 없는 473개 대상 물질 중 PubChem CID와 SMILES 코드가 있는 245개 물질에 대해 에스트로겐 수용체와 반응할 수 있는지 확인하였다(Fig. 1).

선별된 화학물질의 에스트로겐 수용체 반응 활성(독성기전)이 생식·수유독성(최종 독성영향)으로 이어지는지 살펴보려면 *in vivo* 결과와 *in vitro* 결과의 연관성을 확인해야 한다. 생식·수유독성(유(1), 무(0))과 ToxCast의 hitcall 자료(활성(1), 비활성(0))는 모두 이분화 변수(명목척도)이므로, 본 연구에서는 IBM SPSS Statistics 통계프로그램(IBM Corp., New York, NY, USA)의 파이-상관계수(phi-correlation coefficient)를 조사하여 상관관계를 확인하였다. 파이 상관계수가 0.25보다 클 경우 매우 강한 상관관계, 0.15보다 클 경우 강한 상관관계, 0.1보다 클 경우 중간 상관관계, 0.05보다 클 경우 약한 상관관계가 있음을 나타낸다.[31]

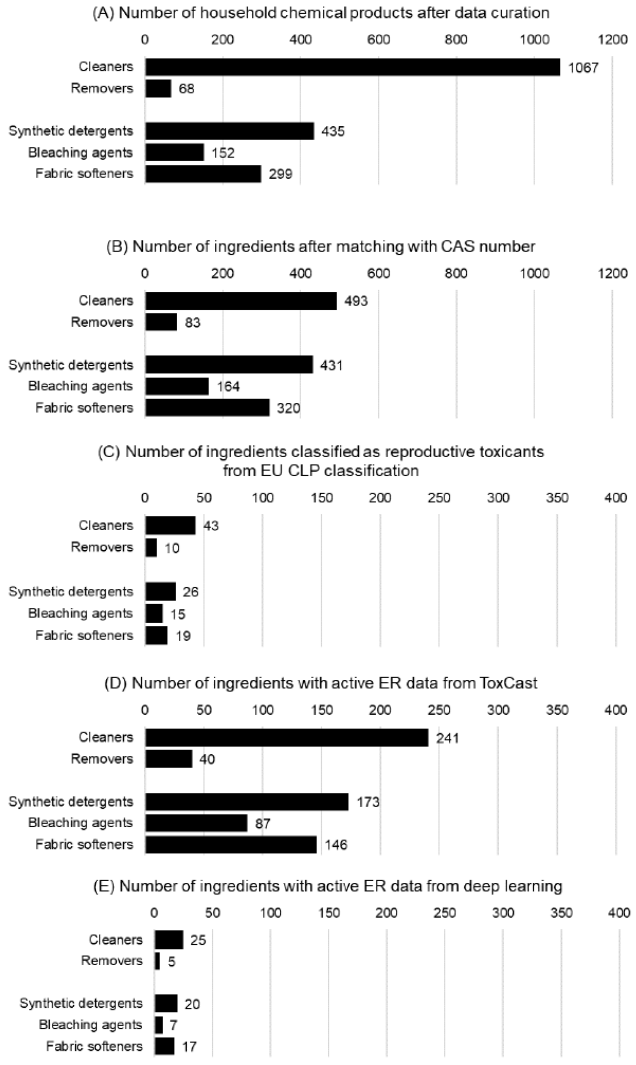

Fig. 2. Number of household chemical products and ingredients. (A) Number of cleaning and laundry products after data curation, (B) Number of ingredients after matching with CAS number, (C) Number of ingredients classified as reproductive toxicants from EU CLP classification, (D) Number of ingredients with active ER data from ToxCast, (E) Number of ingredients with active ER data from deep learning model

Ⅲ. 결 과

1. 연구대상 제품 및 성분 자료

자료를 수집한 3,135종의 안전확인대상 생활화학제품 중 성분이 제공된 제품은 총 2,021종으로, 중분류 별로 정리하면 세정제 1,067종, 제거제 68종, 합성세제 435종, 표백제 152종, 섬유유연제 299종이다(Fig. 2A). 성분을 공개하지 않거나 성분 정보에 관련이 없는 문구가 제시된 제품이 자료를 수집한 전체 제품의 35.5%에 해당되었다. 연구대상 제품에서 자료를 수집한 성분의 개수는 총 887개이며, 이 중에서 CAS 등록번호를 확인할 수 있는 물질은 총 783개(88.3%)로 확인되었다. 783개 물질 중에서 세정제, 제거제, 합성세제, 표백제, 섬유유연제에 함유된 물질은 각각 493개, 83개, 431개, 164개, 320개이었다(Fig. 2B). CAS 등록번호를 식별할 수 있는 물질 중에서 271개 물질(34.6%)은 2개 이상의 동의어로 제품에 기재되어 있음이 확인되었다. CAS 등록번호 뿐만 아니라 PubChem CID, SMILES, DTXSID가 모두 존재하는 물질은 538개(60.7%)이었다.

세정제품과 세탁제품에 다빈도로 함유된 15가지 물질과 제품에 표기된 동의어, 제품 내 물질의 용도를 Table 1에 제시하였다. 물을 제외하고 연구대상 제품에 가장 많은 빈도로 함유된 물질은 에탄올(CAS 등록번호: 64-17-5; 19.4%)이었으며, 에톡실화된 도데실-1-올(CAS 등록번호: 9002-92-0; 18.4%), 탄산수소나트륨(CAS 등록번호: 144-55-8; 17.3%), d-리모넨(CAS 등록번호: 5989-27-5; 16.4%), 수산화나트륨(CAS 등록번호: 1310-73-2; 13.5%), 리날룰(CAS 등록번호: 78-70-6; 13.3%) 순으로 확인되었다(Table 1).

2. 생식·수유독성 및 에스트로겐 수용체와 반응할 수 있는 물질 선정

유럽화학물질청의 분류, 표지 및 포장(EU CLP) 분류를 통해 생식·수유독성을 유발하는 물질은 53개(세정제 43개, 제거제 10개, 합성세제 26개, 표백제 15개, 섬유유연제 19개에서 중복물질 제외)로 확인되었다(Fig. 2C). 에스트로겐 수용체와 결합하거나 활성화하는 물질을 선별하기 위해 18개 ToxCast *in vitro* assay (AECA (1개), ATG (3개), NVS (3개), OT (8개), Tox21 (3개))를 확인하였으며 각 assay에는 시험한 화학물질의 수가 24~8,305개, 데이터세트 내에 양성 물질은 74.0~92.8%

Table 1. Top fifteen most frequently occurred chemicals in cleaning and laundry products, synonyms appearing in product ingredient lists, and ingredient usage

Chemical name	CAS number	No. of products containing this chemical (%)	Synonyms appearing in product ingredient lists	Usage
Water	7732-18-5	642 (31.8%)	Water | Purified water | Aqua	Ubiquitous
Ethanol	64-17-5	393 (19.4%)	Ethanol | Ethyl alcohol	Ubiquitous
Dodecanol-1-ol, ethoxylated	9002-92-0	372 (18.4%)	Dodecan-1-ol, ethoxylated | Dodecyl alcohol, ethoxylated | Ethoxylated lauryl alcohol	Surfactants
Sodium hydrogen carbonate	144-55-8	350 (17.3%)	Sodium hydrogen carbonate | Sodium bicarbonate | Baking soda	Assist in cleaning performance
(d)-Limonene	5989-27-5	331 (16.4%)	(d)-Limonene | (R)-p-mentha-1,8-diene	Fragrance
Sodium hydroxide	1310-73-2	272 (13.5%)	Sodium hydroxide | NaOH	pH stabilizer
Linalool	78-70-6	269 (13.3%)	Linalool | Linalol | 3,7-dimetil-1,6-octadien-3-ol	Fragrance
Propane-1,2-diol	57-55-6	241 (11.9%)	Propane-1,2-diol | Propylene glycol | 1,2-propanediol | Mono propylene glycol	Preservative
Sodium chloride	7647-14-5	227 (11.2%)	Sodium chloride | NaCl | Salt	Formulation stabilizer
Glycerin	56-81-5	222 (11.0%)	Glycerol | Glycerin	Moisturizer
1,2-benzisothiazol-3(2H)-one	2634-33-5	218 (10.8%)	1,2-benzisothiazol-3(2H)-one | 1,2-benzothiazolin-3-one | BIT	Preservative
Sodium carbonate	497-19-8	210 (10.4%)	Sodium carbonate | Carbonic acid sodium salt	Assist in cleaning performance
Butylphenyl methylpropional	80-54-6	210 (10.4%)	Butylphenyl methylpropional | 2-(4-tert-butylbenzyl) propionaldehyde | p-tert-butyl-alpha-methylhydrocinnamic aldehyde	Fragrance
Hexyl cinnamal	101-86-0	201 (9.9%)	Hexyl cinnamal | α-Hexylcinnamaldehyde | 2-benzylideneoctanal | 2-(phenylmethylene)octanal	Fragrance
Citric acid	77-92-9	193 (9.5%)	Citric acid | 1,2,3-Propanetricarboxylic acid, 2-hydroxy-	Stabilizer

포함되어 있다(Table 2). 783개 대상물질 중 ToxCast *in vitro* assay 자료가 존재하는 화학물질은 520개이었고, 이 중 에스트로겐 수용체와 반응하는 데 양성을 보인 물질은 총 310개(세정제 241개, 제거제 40개, 합성세제 173개, 표백제 87개, 섬유유연제 146개에서 중복물질 제외)이었다(Fig. 2D). ToxCast의 데이터 갭을 해결하기 위해 ToxCast의 17개 ER assay를 토대로 심층학습 모델을 구축하였다. ToxCast의 클래스 불균형을 조정하기 위해 SMOTEENN을 사용하였고, 층화된 5중 교차검증을 통해 모델의 정확도는 90.69~98.03%로 확인되었다(Table 3). 모델의 예측 정확도를 나타내는 또다른 지표인 민감도(true positive rate=sensitivity)는 0.93~1.00, 특이도(true negative rate=specificity)는 0.84~0.97이었으며, OT, Tox21 source의 assay를 비교적 잘 예측하는 것으로 나타났다(Table 3). 파이-상관계수를 통해 ToxCast의 14개 ER assay (ACEA, ATG, Tox21 source)가 *in vivo* 생식·수유독성 자료와 상관성이 높은 것으로 나타났으며, NVS source의 3가지 assay는 연관성이 낮은 것으로 나타났다(Table 4). Tox21 source의 3개 ER assay는 파이-상관계수가 0.15보다 커 *in vivo* 생식·수유독성 자료와 강한 상관관계가 있음이 확인되었다(Table 4). ToxCast *in vitro* assay 자료가 없는 473개의 화학물질 중 PubChem CID, SMILES, DTXSID가 모두 존재하는 245개 물질에 대해 심층학습 모델로 예측독성 자료를 확보하였고, 살리실산류(예: 살리실산 펜틸(CAS 등록번호: 2050-08-0), 살리실산 2-메틸뷰틸(CAS 등록번호: 51115-63-0)), 글리콜에테르류, 사이클로실록세인류, 에탄올아민류를 포함한 42개 물질이 양성 반응을 가진 것으로 확인되었다(Fig. 2E). 42개 물질 중에서 세정제, 제거제, 합성세제, 표백제, 섬유유연제에 함유된 물질은 각각 25개, 5개, 20개, 7개, 17개이었다(Fig. 2E).

3가지 방법(EU CLP 분류, ToxCast, 심층학습 모델)을 통해 생식·수유독성 또는 에스트로겐 수용체 반응물질로 판별된 것은 235종이며(Fig. 1), 선정된 물질은 대부분 향료의 구성성분(세정제 44.8%, 제거제 22.2%, 합성세제 50.4%, 표백제 79.3%, 섬유유연제 73.6%)으로 사용되고 있다. 두 가지 독성에 모두 양성을 보이는 물질은 13종으로(Table 5), 부틸페닐메틸프로피오날(CAS 등록번호: 80-54-6), 트라이에탄올아민(CAS 등록번호: 102-71-6), 부틸화하이드록시톨루엔(CAS 등

Table 2. Number of active and inactive chemicals based on 18 ToxCast *in vitro* assays

Assay		AEID	No. of total chem.	No. of active chem. (%)	No. of target chem. (%) -active-	No. of target chem. (%) -inactive-
Source	Name					
ACEA*	ER_80hr	2	3,031	2,575 (85.0%)	215 (27.5%)	28 (3.6%)
ATG*	ERa_TRANS_up	117	3,799	2,972 (78.2%)	189 (24.1%)	62 (7.9%)
	ERb_TRANS2_up	1,367	24	18 (75.0%)	0 (0.0%)	0 (0.0%)
	ERE_CIS_up	75	3,800	2,813 (74.0%)	198 (25.3%)	53 (6.8%)
NVS*	NR_bER	708	1,088	957 (88.0%)	44 (5.6%)	6 (0.8%)
	NR_hER	714	1,177	933 (79.2%)	47 (6.0%)	12 (1.5%)
	NR_mERa	725	958	759 (79.2%)	36 (4.6%)	11 (1.4%)
OT*	ER_ERaERa_0480	742	1,857	1,707 (91.9%)	177 (22.6%)	7 (0.9%)
	ER_ERaERa_1440	743	1,857	1,686 (90.8%)	170 (21.7%)	14 (1.8%)
	ER_ERaERb_0480	744	1,857	1,610 (86.7%)	169 (21.6%)	15 (1.9%)
	ER_ERaERb_1440	745	1,857	1,527 (82.2%)	162 (20.7%)	22 (2.8%)
	ER_ERbERb_0480	746	1,857	1,622 (87.3%)	171 (21.8%)	13 (1.7%)
	ER_ERbERb_1440	747	1,857	1,602 (86.3%)	167 (21.3%)	17 (2.2%)
	ERa_EREGFP_0120	750	1,857	1,654 (89.1%)	171 (21.8%)	13 (1.7%)
	ERa_EREGFP_0480	751	1,857	1,656 (89.2%)	171 (21.8%)	13 (1.7%)
Tox21*	ERa_BLA_Agonist_ratio	785	8,305	7,707 (92.8%)	306 (39.0%)	4 (0.5%)
	ERb_BLA_Agonist_ratio	2,115	7,871	6,204 (78.8%)	276 (35.2%)	34 (4.3%)
	ERa_LUC_VM7_Agonist	788	8,305	7,707 (92.8%)	303 (38.7%)	7 (0.9%)

*ACEA: arachidonyl-2'-chloroethylamide assay; ATG: Attagene Factorial™ assay, which provides high-content assessment of over 90 different gene regulatory pathways and all 48 human nuclear receptors; NVS: Novascreen® assay, which provides information on binding to estrogen receptor; OT: Odyssey Thera; Tox21: Assays run by the National Institutes of Health's National Center for Advancing Translational Sciences (NCATS) as part of the Federal Tox21 program.

Table 3. Description of deep learning models based on the ToxCast assays

Assay name	AEID	Intended target	Model accuracy (%)	True positive rate	True negative rate	False positive rate	False negative rate
ACEA_ER_80hr	2	ESR	98.03	0.97	0.87	0.13	0.03
ATG_ERa_TRANS_up	117	ESR1	97.17	0.95	0.85	0.15	0.05
ATG_ERb_TRANS2_up*	1,367	ESR2	-	-	-	-	-
ATG_ERE_CIS_up	75	ESR1	97.58	0.93	0.84	0.16	0.07
NVS_NR_bER	708	ESR1	90.69	0.95	0.88	0.12	0.05
NVS_NR_hER	714	ESR1	91.31	0.94	0.89	0.11	0.06
NVS_NR_mERa	725	ESR1	94.97	0.96	0.87	0.13	0.04
OT_ER_ERaERa_0480	742	ESR1	95.07	0.99	0.97	0.03	0.01
OT_ER_ERaERa_1440	743	ESR1	97.54	0.98	0.94	0.06	0.02
OT_ER_ERaERb_0480	744	ESR1,2	95.58	0.99	0.97	0.03	0.01
OT_ER_ERaERb_1440	745	ESR1,2	97.17	0.97	0.93	0.07	0.03
OT_ER_ERbERb_0480	746	ESR2	97.34	1.00	0.96	0.04	0.00
OT_ER_ERbERb_1440	747	ESR2	96.37	0.97	0.92	0.08	0.03
OT_ERa_EREGFP_0120	750	ESR1	95.62	0.98	0.91	0.09	0.02
OT_ERa_EREGFP_0480	751	ESR1	96.20	0.97	0.94	0.06	0.03
Tox21_ERa_BLA_Agonist_ratio	785	ESR1	95.22	0.98	0.93	0.07	0.02
Tox21_ERb_BLA_Agonist_ratio	2,115	ESR2	94.87	0.96	0.88	0.12	0.04
Tox21_ERa_LUC_VM7_Agonist	788	ESR1	97.19	0.95	0.92	0.08	0.05

*Since this assay does not contain many test chemicals, it was excluded from the construction of the deep learning model.

Table 4. Potential assays related to reproductive toxicity analyzed by phi-coefficient correlation

Assay name	AEID	ToxCast *in vitro* data		Deep learning prediction		Active average (%)	Phi-correlation with reproductive toxicity	
		Active/ total chemical	Active (%)	Active/ total chemical	Active (%)		Phi-coefficient	p-value
ACEA_ER_80hr	2	215/243	88.5	49/68	72.0	84.9	0.108	0.003
ATG_ERa_TRANS_up	117	189/251	75.3	28/153	18.3	53.7	0.109	0.002
ATG_ERb_TRANS2_up*	1,367	0/0	0.0	-	-	-	-	-
ATG_ERE_CIS_up	75	198/251	78.9	32/153	20.9	56.9	0.101	0.005
NVS_NR_bER	708	44/50	88.0	13/41	31.7	62.6	0.000	0.989
NVS_NR_hER	714	47/59	79.7	31/75	41.3	58.2	0.004	0.914
NVS_NR_mERa	725	36/47	76.6	11/28	39.3	62.6	0.038	0.288
OT_ER_ERaERa_0480	742	177/184	96.2	22/169	13.0	56.4	0.122	0.001
OT_ER_ERaERa_1440	743	170/184	92.4	34/199	17.0	53.3	0.092	0.010
OT_ER_ERaERb_0480	744	169/184	91.8	83/169	49.1	71.3	0.106	0.003
OT_ER_ERaERb_1440	745	162/184	88.0	80/199	40.2	63.2	0.088	0.013
OT_ER_ERbERb_0480	746	171/184	92.9	76/169	44.9	70.0	0.116	0.001
OT_ER_ERbERb_1440	747	167/184	90.8	55/199	27.6	58.0	0.108	0.003
OT_ERa_EREGFP_0120	750	171/184	92.9	57/133	42.8	71.9	0.116	0.001
OT_ERa_EREGFP_0480	751	171/184	92.9	57/156	36.5	67.0	0.104	0.004
Tox21_ERa_BLA_Agonist_ratio	785	306/310	98.7	24/87	27.6	83.1	0.180	0.000
Tox21_ERb_BLA_Agonist_ratio	2,115	276/310	89.0	52/148	35.1	71.6	0.216	0.000
Tox21_ERa_LUC_VM7_Agonist	788	303/310	97.7	40/87	46.0	86.4	0.172	0.000

*Since this assay does not contain many test chemicals, it was excluded from the construction of the deep learning model.

Table 5. Substances that are positive for both reproductive toxicity and ER-mediated toxicity among target chemicals

No.	Chemical name	CAS no.	Contained product*				
			CL	RE	SD	BA	FS
1	2-(4-tert-butylbenzyl)propionaldehyde	80-54-6	O	×	O	O	O
2	2,2',2''-nitrilotriethanol	102-71-6	O	O	O	O	×
3	2,6-di-tert-butyl-p-cresol	128-37-0	O	×	O	O	O
4	2-butoxyethanol	111-76-2	O	O	×	×	×
5	octamethylcyclotetrasiloxane	556-67-2	O	×	O	O	O
6	2,2'-iminodiethanol	111-42-2	O	×	O	×	×
7	trisodium 5-hydroxy-1-(4-sulphophenyl)-4-(4-sulphophenylazo)pyrazole-3-carboxylate	1934-21-0	O	×	O	×	×
8	oxalic acid	144-62-7	O	×	O	×	×
9	m-xylene	108-38-3	O	×	×	×	×
10	methyl salicylate	119-36-8	×	×	O	×	×
11	3-(4-tert-butylphenyl)propionaldehyde	18127-01-0	×	×	×	O	×
12	2-pyrrolidone	616-45-5	×	O	×	×	×
13	p-mentha-1,3-diene	99-86-5	O	×	×	×	×

*CL: cleaners, RE: removers, SD: synthetic detergents, BA: bleaching agents, FS: fabric softeners.

록번호: 128-37-0) 등이 선정되었다.

IV. 고 찰

동물을 이용한 전통적인 독성시험으로는 증가하는 화학물질의 속도를 감당하기 어려우므로 화학물질의 독성을 신속·정확하게 평가할 수 있는 대체시험법의 개발과 활용이 필요하다. 사람에 대한 역학연구 또는 동물을 이용한 독성시험(*in vivo*)을 토대로 설정되는 독성 분류등급에 시험관 내(*in vitro*) 독성 빅데이터 자료와 심층학습 모델의 예측독성 자료를 추가하면 분자수준의 독성기전부터 최종 독성영향까지 고려한 화학물질 선별 시스템 구축이 가능하다. 본 연구에서는 이러한 시스템 구동의 일례로 세정제품과 세탁제품에 함유된 성분들 중에서 에스트로겐 수용체와 반응하여 생식·수유장애를 일으킬 수 있는 화학물질들을 선별하기 위해 EU CLP 분류, ToxCast 빅데이터 자료, 심층학습 모델의 예측독성 자료를 활용하였으며, *in vivo-in vitro-in silico*를 아우르는 화학물질 선별 시스템의 구축 가능성을 시연하였다.

생활화학제품에 포함된 성분의 독성을 확인하려면 우선 다양한 방식으로 기재된 성분명을 하나의 동의어로 통일하고, 화학물질 식별자와 연결하는 작업이 필요하다. 본 연구의 대상제품에서 88.3%의 물질에 대해 CAS 등록번호와 연결할 수 있었고, 이 중에서 34.6%는 2개 이상의 동의어가 제품에 기재됨을 확인하였다. 생활화학제품의 성분을 표시하는 데 다양한 동의어가 사용되면, 소비자가 제품을 구매할 때 의사 결정에 방해가 될 수 있으며,[2] 제품 위해성 평가 시 정확한 정보전달이 이루어지기 어렵다. 소비자의 알 권리와 위해성 소통을 증진시키고, 정확한 정보가 전달될 수 있도록 함유된 개별성분을 일원화하여 기재하고 확인할 수 있는 시스템을 구축하는 것이 필요하다.

생활화학제품의 위해성을 관리하기 위해서는 제품 안에 포함된 화학물질들의 노출 및 독성 자료가 확보되어야 한다. 그러나 연구대상 물질 중 439개(56.1%) 물질은 ToxCast 자료를 확보할 수 없었다. 광범위한 독성 데이터베이스에 시험자료가 수록된 물질들은 대부분 환경 잔류성이 크고 대량 생산되는 물질이며 소비자가 자주 노출될 수 있는 물질은 상대적으로 자료를 확보하기가 어렵다.[2,10] 이러한 자료의 공백을 채우기 위해 인공지능 기술을 활용한 예측독성이 제안되고 있으며, 심층학습을 토대로 내분비계 장애물질을 확인하는 방법도 소개되고 있다.[32,33] 본 연구에서는 ToxCast의 17개 ER assay를 토대로 심층학습 모델을 구축하였고, 안전확인대상 생활화학제품에 포함된 화학물질들 중 ToxCast ER assay 자료가 없는 물질들을 토대로 에스트로겐 수용체 양성 물질을 선별하였다. 심층학습 모델을 통해 살리실산 펜틸, 살리실산 2-메틸뷰틸 등이 에스트로겐 수용체 양성 물질로 예측되었고, 이 물질들과 구조가 유사한 살리실산류(예: 살리실산 페닐, 살리실산 벤질)의 ER 활성이 보고된 바 있다.[34] 컴퓨터 기반의 화학물질 독성 예측 프로그램이 널리 활용되면 미리 그 물질의 독성을 예측하거나 신규 환경오염물질의 독성을 신속하게 파악하는 데 활용될 수 있다. 본 연구에서는 심층학습 모델의 교차검증을 통해 모델의 정확도, 민감도, 특이도가 높음을 확인했으나, 모델의 신뢰도를 높이기 위해 추후에 구축된 모델의 외부 데이터를 이용한 검증 작업이 필요하다.

세정제품과 세탁제품은 일반 소비자들이 가장 빈번하게 노출될 수 있는 안전확인대상 생활화학제품 중의 하나이다. 본 연구에서 3가지 방법(EU CLP 분류, ToxCast, 심층학습 모델)으로 에스트로겐 수용체 반응과 생식·수유독성에 모두 양성을 보이는 13종의 물질은 주로 세정제, 합성세제, 섬유유연제에 향료로 포함된다(Table 5). 이는 대상제품 내 향료의 안전관리가 중요함을 나타낸다. 향료는 원하는 향을 얻거나 제품의 다른 향을 가리기 위해 사용된다. 하나의 향료에는 일반적으로 50~300개의 화학물질이 포함되며[35] 단일물질로 구분하기 어려운 경우도 있다. 세정제품과 세탁제품(특히 섬유유연제)에 다빈도로 함유된 리모넨과 리날룰은 테르펜(terpene)의 일종으로서, 피부에 접촉할 경우 자극 및 알레르기를 유발하는 것으로 알려졌다.[36] 헥실신남알,[37] 제라니올[38] 등의 향료 성분은 에스트로겐성을 지녔다는 보고도 있다.

최근 화학물질의 유해성 평가나 규제 측면에서 특정한 독성기전과 연관된 물질을 선별하기 위해 독성발현경로와 심층학습 모델을 활용하는 사례가 많아지고 있다. 한 연구팀에서는 폐섬유화 유발 독성발현경로와 관련된 ToxCast 시험관 내(*in vitro*) 독성시험 자료를 토대로 심층학습 예측 모델을 만든 후 흡입독성 유발물질을 선별하였다.[19] 플라스틱에 함유된 첨가제 50종을 토대로 ToxCast 데이터베이스에서 자료를 확인할 수 없는 물질에 대해 첨가제의 독성발현 기전과 관련된 339개의 심층학습 모델을 구축하여 독성을 예측한 연구도 있다.[20] 본 연구에서는 분자 수준의 초기 현상(에스트로겐 수용체 반응 활성)부터 최종 독성영향(생식·수유독성)까지 고려한 화학물질 선별 시스템의 구동 가능성을 살펴보고자 했다. 파이-상관계수를 통해 ToxCast의 14개 ER assay가 *in vivo* 생식·수유독성 자료와 상관성이 높고 13종의 물질이 에스트로겐 수용체 반응을 통해 생식·수유독성을 유발할 수 있음을 확인하였다. 이는 생식·수유독성 유발물질의 약 25%가 에스트로겐 수용체 활성 기전과 연관됨을 의미한다. 이 결과는 상관분석으로 얻어진 것이기에 생물학적 분자 지표와 생식·수유독성 간의 직접적인 관계를 해석하는 데 일부 제한적일 수 있으며, 추후 생식·수유독성에 관련이 있는 다른 지표(예: 스테로이드호르몬 생합성) 들과의 상관성을 확인할 필요가 있다. 본 연구에서 시연한 *in vivo-in vitro-in silico* 화학물질 선별 시스템은 향후 독성자료가 부족한 다른 소비자제품 내 유해화학물질을 선별하는 데 활

용될 수 있을 것이다.

V. 결 론

본 연구에서는 동물 독성시험(*in vivo*)을 토대로 설정되는 독성 분류 시스템에 ToxCast *in vitro* 독성 빅데이터 자료와 심층학습을 통한 예측독성 자료를 활용하여 화학물질 선별 시스템의 구동 가능성을 살펴보았다. 위해성 평가가 필요한 안전확인 대상 생활화학제품 중 일반 소비자들에게 빈번하게 노출되는 세정제품(n=1,135)과 세탁제품(n=886)에 함유된 성분들을 대상으로 하였고, 783종 화학물질 중 에스트로겐 수용체와 반응하여 생식·수유독성을 유발할 수 있는 13종 화학물질을 선별하였다. 화학물질 독성평가 인프라를 갖추는 데 많은 비용과 시간이 드는 만큼 *in vivo-in vitro-in silico*를 통합한 화학물질 선별 시스템은 화학물질에 관한 빅데이터와 인공지능 기술을 활용해 동물실험을 최소화하고 효율적으로 독성평가 인프라를 갖추기 위한 방안이 될 것이다.

감사의 글

본 연구는 한국연구재단(과제번호 2019R1A2C1002712)의 지원을 받아 수행되었습니다.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. Lee D, Lim H, Kim JH, Kim T, Hwang M, Seok K, et al. An investigation of consumer product co-use patterns - focusing on air-fresheners and deodorizer. *J Environ Health Sci.* 2018; 44(3): 275-282.
2. Gabb HA, Blake C. An informatics approach to evaluating combined chemical exposures from consumer products: a case study of asthma-associated chemicals and potential endocrine disruptors. *Environ Health Perspect.* 2016; 124(8): 1155-1165.
3. Sanderson H, Greggs W, Cowan-Ellsberry C, DeLeo P, Sedlak R. Collection and dissemination of exposure data throughout the chemical value chain: a case study from a global consumer product industry. *Hum Ecol Risk Assess Int J.* 2013; 19(4): 999-1013.
4. Park JW. The role of environmental law in assessing and managing risk of chemical products and biocides. *Environ Law Policy.* 2018; 20: 55-85.
5. Heo DA, Huh EH, Park JY, Moon KW, Lee K. An investigation of ingredients and hazardous substances in some consumer products - focusing on cleaners and disinfectants. *J Environ Health Sci.* 2015; 41(5): 314-326.
6. Heo JJ, Kim UJ, Oh JE. Simultaneous quantitative analysis of four isothiazolinones and 3-iodo-2-propynyl butyl carbamate in hygienic consumer products. *Environ Eng Res.* 2019; 24(1): 137-143.
7. Pyun DH, Kim YW, Hwang YW, Lee SY, Park SH. A hazard assessment of chemicals in liquid synthetic detergent using GreenScreen. *Korean J Hazard Mater.* 2021; 9(1): 30-40.
8. Park JY, Lim M, Lee K, Ji K, Yang W, Shin HS, et al. Consumer exposure and risk assessment to selected chemicals of mold stain remover use in Korea. *J Expo Sci Environ Epidemiol.* 2020; 30(5): 888-897.
9. Gore AC, Chappell VA, Fenton SE, Flaws JA, Nadal A, Prins GS, et al. EDC-2: the endocrine society's second scientific statement on endocrine-disrupting chemicals. *Endocr Rev.* 2015; 36(6): E1-E150.
10. Dodson RE, Nishioka M, Standley LJ, Perovich LJ, Brody JG, Rudel RA. Endocrine disruptors and asthma-associated chemicals in consumer products. *Environ Health Perspect.* 2012; 120(7): 935-943.
11. Rotroff DM, Dix DJ, Houck KA, Knudsen TB, Martin MT, McLaurin KW, et al. Using in vitro high throughput screening assays to identify potential endocrine-disrupting chemicals. *Environ Health Perspect.* 2013; 121(1): 7-14.
12. United States Environmental Protection Agency. ToxCast Dashboard. Available: https://comptox.epa.gov/dashboard/chemical_lists/toxcast [accessed 20 August 2021].
13. National Institute of Environmental Health Sciences. National Toxicology Program. Tox21: Toxicology in the 21st Century. Available: https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html [accessed 20 August 2021].
14. Judson RS, Magpantay FM, Chickarmane V, Haskell C, Tania N, Taylor J, et al. Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. *Toxicol Sci.* 2015; 148(1): 137-154.
15. Andersson N, Arena M, Auteri D, Barmaz S, Grignard E, Kienzler A, et al. Guidance for the identification of endocrine disruptors in the context of Regulations (EU) No 528/2012 and (EC) No 1107/2009. *EFSA J.* 2018; 16(6): e05311.
16. To KT, Fry RC, Reif DM. Characterizing the effects of missing data and evaluating imputation methods for chemical prioritization applications using ToxPi. *BioData Min.* 2018; 11: 10.
17. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif Intell Med.* 2010; 50(2): 105-115.
18. Tang W, Chen J, Wang Z, Xie H, Hong H. Deep learning for predicting toxicity of chemicals: a mini review. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev.* 2018; 36(4): 252-271.
19. Jeong J, Garcia-Reyero N, Burgoon L, Perkins E, Park T, Kim C, et al. 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.* 2019; 32(6): 1212-1222.
20. Jeong J, Choi J. Development of AOP relevant to microplastics based on toxicity mechanisms of chemical additives using ToxCast™ and

deep learning models combined approach. *Environ Int.* 2020; 137: 105557.
21. Ministry of Environment. Living Environment Safety Information System. Available: https://ecolife.me.go.kr/ecolife [accessed 20 August 2021].
22. National Institute of Environmental Research. National Chemicals Information System. Available: https://ncis.nier.go.kr [accessed 20 August 2021].
23. European Chemicals Agency. Information on Chemicals. Available: https://echa.europa.eu/information-on-chemicals [accessed 20 August 2021].
24. United States Environmental Protection Agency. CompTox Chemicals Dashboard. Available: https://comptox.epa.gov/dashboard [accessed 20 August 2021].
25. National Library of Medicine. PubChem. Available: https://pubchem.ncbi.nlm.nih.gov/ [accessed 20 August 2021].
26. Liu J, Mansouri K, Judson RS, Martin MT, Hong H, Chen M, et al. Predicting hepatotoxicity using ToxCast in vitro bioactivity and chemical structure. *Chem Res Toxicol.* 2015; 28(4): 738-751.
27. Jeong J, Bae SY, Choi J. Identification of toxicity pathway of diesel particulate matter using AOP of PPARγ inactivation leading to pulmonary fibrosis. *Environ Int.* 2021; 147: 106339.
28. Idakwo G, Thangapandian S, Luttrell J, Li Y, Wang N, Zhou Z, et al. Structure-activity relationship-based chemical classification of highly imbalanced Tox21 datasets. *J Cheminform.* 2020; 12(1): 66.
29. RDKit. RDKit: Open-Source Cheminformatics Software. Available: http://www.rdkit.org [accessed 20 August 2021].
30. Pérez-Enciso M, Zingaretti LM. A guide for using deep learning for complex trait genomic prediction. *Genes (Basel).* 2019; 10(7): 553.
31. Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med.* 2018; 18(3): 91-93.
32. Mukherjee A, Su A, Rajan K. Deep learning model for identifying critical structural motifs in potential endocrine disruptors. *J Chem Inf Model.* 2021; 61(5): 2187-2197.
33. Burgoon LD. Autoencoder Predicting Estrogenic Chemical Substances (APECS): an improved approach for screening potentially estrogenic chemicals using in vitro assays and deep learning. *Comput Toxicol.* 2017; 2: 45-49.
34. Zhang Z, Jia C, Hu Y, Sun L, Jiao J, Zhao L, et al. The estrogenic potential of salicylate esters and their possible risks in foods and cosmetics. *Toxicol Lett.* 2012; 209(2): 146-153.
35. Bickers DR, Calow P, Greim HA, Hanifin JM, Rogers AE, Saurat JH, et al. The safety assessment of fragrance materials. *Regul Toxicol Pharmacol.* 2003; 37(2): 218-273.
36. Audrain H, Kenward C, Lovell CR, Green C, Ormerod AD, Sansom J, et al. Allergy to oxidized limonene and linalool is frequent in the U.K. *Br J Dermatol.* 2014; 171(2): 292-297.
37. Gunia-Krzyżak A, Słoczyńska K, Popiół J, Koczurkiewicz P, Marona H, Pękala E. Cinnamic acid derivatives in cosmetics: current use and future prospects. *Int J Cosmet Sci.* 2018; 40(4): 356-366.
38. Howes MJ, Houghton PJ, Barlow DJ, Pocock VJ, Milligan SR. Assessment of estrogenic activity in some common essential oil constituents. *J Pharm Pharmacol.* 2002; 54(11): 1521-1528.

〈저자정보〉
이인혜(연구원), 이수진(대학원생), 지경희(교수)