농약과학회지 (Korean J. Pestic. Sci.)
Vol. 30, No. 3, pp. 212-222 (2026)
Open Access https://doi.org/10.7585/kjps.2026.30.3.212
ORIGINAL ARTICLES

Online ISSN 2287-2051
Print ISSN 1226-6183
Check for updates

한국 등록농약 피부투과 물성공간의 군집부재 진단과 space-filling 기반 대표 물질 선정

신지영[1] · 홍소혜[1] · 양시영[1] · 조유미[1] · 김솔아[1] · 정미혜[1] · 문지현[1] · 박수진[2*†] · 김태곤[3*†]

[1]농촌진흥청 국립농업과학원 농산물안전성부 독성위해평가과, [2]농촌진흥청 연구정책국 농자재산업과
[3]전북대학교 농업생명과학대학 스마트팜학과

Diagnosing Clusterability and Selecting Representative Compounds by Space-filling Design in the Skin Permeation Property Space of Korean Registered Pesticides

Ji-Young Shin[1], So-Hye Hong[1], Si-Young Yang[1], You-Mi Jo[1], Sol-A Kim[1], Mi-Hye Jeong[1], Ji-Hyun Moon[1], Soo-Jin Park[2*†], and Tae-Gon Kim[3*†]

[1]Toxicity and Risk Assessment Division, Department of Agro-Food Safety & Crop Protection, National Institute of Agricultural Sciences, Rural Development Administration (RDA), Wanju 55365, Republic of Korea
[2]Agro-materials Industry Division, Rural Development Administration (RDA), Jeonju 54875, Republic of Korea
[3]Department of Smart Farm, College of Agriculture & Life Sciences, Jeonbuk National University

(Received on August 19, 2026. Revised on September 8, 2026. Accepted on September 8, 2026)

Abstract Dermal absorption is a key input in pesticide risk assessment. However, its experimental determination is costly, and in silico models have emerged as alternative methods. Before such models are developed, the substances to be measured must be selected. This selection is a design problem that precedes model development. The selection strategy depends on the structure of the physicochemical space. The clusterability of 356 registered pesticide active ingredients was assessed using five diagnostics based on different principles. To verify their effectiveness, the same procedure was applied to synthetic positive and negative controls matched in size, dimensionality, and range. No cluster structure was detected (dip tests unimodal; 0/180 configurations above the silhouette threshold; ARI 0.21-0.52). The silhouette for the real data (0.231) approached that of the negative control (0.192), whereas clusters were clearly detected in the positive control (0.576). Nevertheless, the Hopkins statistic was 0.951, indicating that the compounds were concentrated in certain regions, and the void fraction in five dimensions was approximately 99%, precluding uniform coverage at high dimensionality. Accordingly, 25 compounds were selected by farthest-point selection within log Kp quintiles. The selection outperformed 1,000 random draws in both mean coverage distance (0.0087 vs. 0.0114) and maximum gap (0.4193 vs. 0.4959). The selected compounds can serve as a basis for constructing training data under limited experimental resources.

Key words: Clusterability, Dermal absorption, Pesticide, In silico, Space-filling design

서 론

농약 원제(active ingredient)는 인체에 여러 경로로 노출되며, 농작업자의 경우 경피 노출이 주된 경로로 알려져 있다(Krieger, 1995). 이에 따라 피부에 부착된 양 중 실제로 체내에 흡수되는 비율, 즉 피부흡수율은 농작업자 위해성 평가의 핵심 입력값으로 사용된다(EFSA, 2017; OECD, 2004).

피부흡수율의 표준 산출법은 OECD 시험지침 428(in vitro 피부흡수 시험)에 규정되어 있지만, 고비용 및 전문인력과 시설이 필요하다는 한계가 있다. 이러한 제약으로

†The authors contributed equally to this work
*Corresponding author
E-mail: taegon@jbnu.ac.kr, jiny0451@korea.kr

규제 현장에서는 실측값 대신 기본값이 널리 사용된다(EFSA, 2017; RDA, 2026). 국내에서는 단일 기본값을 일률적으로 적용하는 반면, 유럽식품안전청(European Food Safety Authority, EFSA)은 제형에 따라 기본값을 적용하고 있다(EFSA, 2017). 또한 옥탄올-물 분배계수(log P)와 분자량(molecular weight, MW)의 기준으로 기본값을 구분 적용하는 방식도 사용되어 왔으며, 국내에서도 이 기준에 따라 유효성분의 기본값을 분류하고 노출량을 산출한 사례가 보고되었다(You et al., 2017). 그러나 일률적 기본값과 이분법적 임계값은 물질 간 물성 차이를 충분히 반영하지 못하므로 화학물질별 예측하는 방향으로 전환이 요구된다.

이러한 요구에 대응하여 컴퓨팅 기술(in silico) 기반 대체 시험법이 그 대안으로 관심을 받아왔다. 다만 예측모델 개발에는 학습 및 검증용 실측 자료가 필연적으로 요구되는데, 국내에 등록된 수백 종의 농약 원제 전체를 실측하는 것은 현실적으로 불가능하다. 따라서 어떤 물질을 우선 측정할 것인가라는 선정 문제가 모델 개발에 선행한다. 선정이 편향되면 학습 및 검증 자료 자체가 물성 스펙트럼의 일부에만 치우치게 되므로, 이는 모델링에 앞서 해결되어야 할 설계 문제이다(Daszykowski et al., 2002). 한편 최근 농약 및 화학물질 연구에서는 실측값 없이 물리화학·구조 정보만으로 화학 공간을 특성화하고 이를 근거로 시험 대상을 선정하는 흐름이 나타나고 있다(Kar et al., 2025; Renai et al., 2025).

그런데 실측 자원의 배분 방식은 물성 공간의 구조에 따라 달라진다. 공간이 서로 구분되는 군집으로 이루어져 있다면 각 군집의 대표를 뽑는 분류 기반 선정이 타당하고(Golbraikh and Tropsha, 2002), 경계 없는 연속체라면 공간 전체를 고르게 덮는 space-filling 설계가 타당하다(Kennard and Stone, 1969). 이때 분류 기반 선정은 공간에 군집이 실재한다는 것을 전제로 하지만, 군집화 알고리즘은 구조가 없는 자료에서도 분할을 산출하므로 사전에 자료의 구조를 확인하지 않으면 얻어진 군집이 자료의 성질인지 알고리즘의 산물인지 구별할 수 없다(Adolfsson et al., 2019). 한편 space-filling 설계는 한정된 표본을 물성 공간에 고르게 배치함으로써 대표성을 확보하며, 이러한 배치가 예측 신뢰도를 높인다는 점은 QSAR 분야에서 입증되었다(Martin et al., 2012). 두 전략 중 어느 쪽이 적절한지가 자료의 구조에 의존한다는 점은 지적된 바 있다(Daszykowski et al., 2002).

그럼에도 앞서의 연구들은 화학공간을 고르게 탐색한다는 방침만을 채택하였을 뿐, 해당 공간이 그 방침에 적합한 구조인지는 검정하지 않았다. 또한 국내 등록 농약 원제의 물성공간이 어떤 구조를 갖는지는 아직 보고된 바가 없다.

본 연구는 국내 등록 농약 원제의 물성공간에 군집 구조가 실재하는지를 진단하고, 그 결과에 부합하는 대표물질 선정 전략을 제시하는 것을 목적으로 한다. 군집 구조의 실재 판정(clusterability)에서는 단일 지표에 의존한 판정의 위험성이 지적되어 왔으므로(Adolfsson et al., 2019; Ronan et al., 2016), 원리가 서로 다른 복수의 진단을 교차 확인하였다. 또한 진단 절차 자체의 검정력를 양성 및 음성 대조군으로 검증하였으며, 선정 결과의 타당성은 무작위 선정 대비 커버리지로 확인하였다. 흡수율 실측값은 사용하지 않으며, 따라서 구조 진단과 그에 근거한 선정까지를 범위로 한다.

재료 및 방법

국내 등록 농약 원제의 물리화학 정보를 수집하여 분석

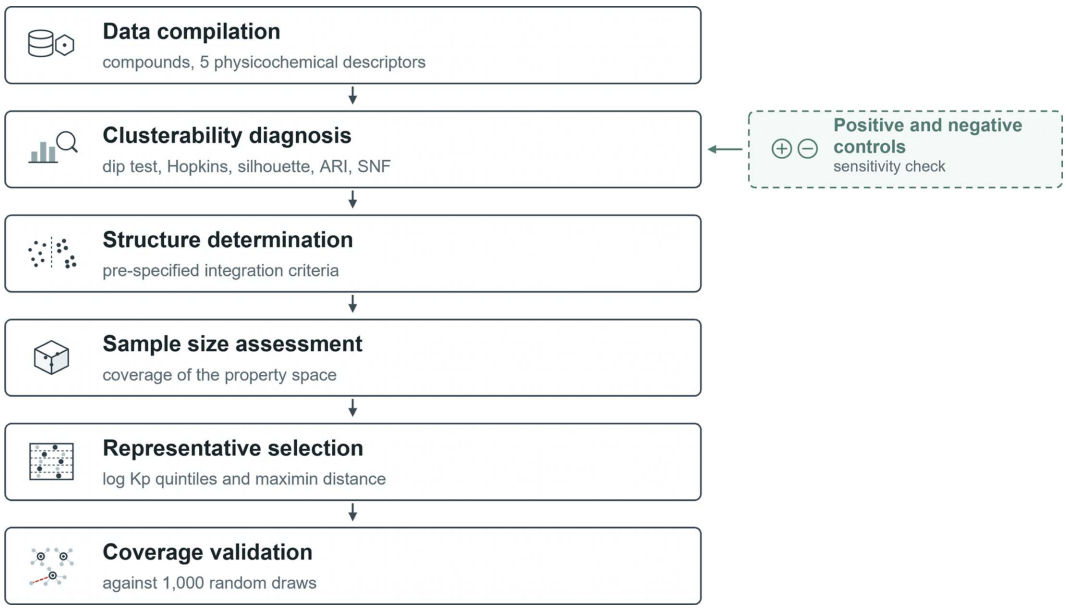


Fig. 1. Study workflow for diagnosing the structure of the physicochemical property space and selecting representative compounds.

Table 1. Definition of the input views used for the clusterability diagnostics

View	Input variables	Preprocessing	Distance	Diagnostics applied	N
Physicochemical view	log P, MW, water solubility, vapour pressure, melting point (5 variables)	$\log_{10}$ of water solubility and vapor pressure, followed by z-score standardization across all variables	Euclidean	dip test, Hopkins, PCA, UMAP, SNF (physicochemical side)	356 (354 for SNF)[a]
Structural view	Morgan/ECFP4 fingerprint (2048-bit, radius 2)	Binary fingerprint	Tanimoto (1-similarity)	SNF (structural side)	354

[a]The SNF cohort comprises 354 of the 356 substances; five organometallic/inorganic or novel active ingredients lacking molecular structure information (SMILES) were excluded.

대상 집합을 구성하고, 물질을 물성 공간에 배열할 좌표축으로 피부투과계수를 산정하였다. 이어서 원리가 서로 다른 복수의 지표로 물성 공간의 군집 구조 실재 여부를 진단하였으며, 진단 절차 자체의 타당성은 인공 양성 및 음성 대조군으로 검증하였다. 판정된 구조 아래에서 보유 표본이 물성 공간을 충분히 덮는지를 정량화한 뒤, 대표 농약을 선정하고 무작위 선정과의 커버리지 대조로 그 타당성을 확인하였다 (Fig. 1).

대상 물질 및 물리화학적 변수

데이터세트는 한국에 등록된 농약 중 원제에 대한 CAS번호를 기준으로 Pesticide Properties Data Base(PPDB)를 통해 물리화학적 정보를 수집하였으며, 염과 금속이 존재하는 데이터를 제외하였다.

물성을 반영한 변수 선택의 경우 피부 투과 예측에 사용되는 물리모델 중 주로 사용되어 온 다섯 가지를 채택하였다. 2차원은 log P와 MW으로 구성하였고(Potts and Guy, 1992), 여기에 융점을 추가하여 3차원으로 확장하였다(OECD, 2022; Waters and Quah, 2022). 그리고 수용해도와 증기압을 추가로 5차원을 적용하여 분석하였다(Eleftheriadou et al., 2019; OECD, 2022).

log P 및 MW를 보유한 원제(400종)와 5개 변수를 모두 보유한 원제(356종)를 분석 목적에 따라 사용하였다. log P와 MW로 산출한 log Kp(Potts and Guy, 1992)를 정렬 축으로 하는 space-filling 및 대표 선정에는 400종을 선정하였고, 다차원 및 커버리지 분석은 다섯 변수 모두 보유한 356종을 사용하였다.

수용해도와 증기압은 상용로그(log Kp) 변환 후 사용하였으며, 거리 기반 분석인군집, PCA(principal component analysis), UMAP(uniform manifold approximation and projection), 커버리지 시 변수 간 척도 차이가 거리를 지배하지 않도록 z-정규화(z-score)로 표준화하였다.

피부투과계수(Kp) 산정

각 원제의 물성 공간 상에 배치하는 좌표 및 그룹화 축으로 활용하기 위해 각 원제의 피부투과계수(Kp)는 Potts and Guy (1992) 식으로 산정하였다.

$$\log Kp = -2.72 + 0.71 \log P - 0.0061 MW \quad (1)$$

여기서 Kp는 정상상태 피부투과계수(cm/h)이고, log P는 옥탄올-물 분배계수의 상용로그이며, MW는 분자량(g/mol)이다.

군집가능성 진단

물성 공간에 자연적인 군집 구조가 존재하는지를 원리가 서로 다른 다섯 가지 진단으로 평가하였다. 물성 공간은 다섯 개의 물성 변수로 정의하되, dip 검정은 1변량 검정이므로 다변량 적용 시 데이터를 1차원으로 축소한 뒤(제1주성분 및 두 점 사이 거리) 검정하였으며, 개별 물성 변수와 정렬 축 log Kp의 주변분포에 대한 dip은 보조 및 탐색 지표로만 산출하였다.

각 진단의 입력 뷰와 적용 대상은 Table 1과 같다. 단변량 및 저차원 진단(dip test, Hopkins, PCA, UMAP)은 5 물성 변수로 정의된 물성 뷰에 적용하였고(N = 356, 5변수 완전 보유 데이터), 다중뷰 융합(similarity network fusion, SNF)은 물성 뷰와 구조 뷰를 함께 사용하였다(N = 354).

Dip test of unimodality

각 변수의 주변분포가 단봉(unimodal)인지 다봉(multimodal)인지를 Hartigan and Hartigan (1985)의 dip 통계량으로 검정하였다(diptest 패키지). (i) 제1주성분에 투영한 값(PCA-dip)과 (ii) 모든 두 점 사이의 유클리드 거리 분포에 각각 dip을 적용하였다. 다변량 자료는 dip을 직접 적용할 수 없어 1차원으로 축소해야 하며, 점 간 거리 분포를 이용하는 방안은 Kalogeratos and Likas (2012)에서 착안하였으며, 모든 두 점 사이 거리를 하나의 분포로 모아 검정하는 방법을 따랐다(Adolfsson et al., 2019). 추가로 각 물성변수 및 log Kp의 주변분포에 대한 개별 dip도 산출하였다.

Hopkins

통계량 데이터의 군집화 경향(clustering tendency)을 Hopkins 통계량으로 평가하였다. 0.5는 균등 분포, 1에 가까울수록

비균등(non-uniform)을 의미한다. 무작위 표본 추출에 따른 변동을 줄이기 위해 시드 0-19의 20회 반복 평균을 사용하였다. 통계량 산출 시 최근접이웃거리는 데이터 차원수 d만큼 거듭제곱하여 합산하였다(Hopkins and Skellam, 1954; Wright, 2022).

차원 축소(PCA 및 UMAP)

표준화된 다차원 공간을 PCA와 UMAP(McInnes et al., 2018)으로 2차원에 투영하여 군집의 시각적 분리 여부를 확인하였다. UMAP이 만들어내는 인위적 군집 착시를 배제하기 위해 PCA 결과와 교차 확인하였다(Kobak and Linderman, 2021).

다중뷰 융합(SNF) 게이트

물성 정보만으로 드러나지 않는 군집이 분자 구조 정보를 함께 고려하면 나타나는지 확인하기 위해, 물성 뷰와 구조 뷰를 SNF(Wang et al., 2014)로 융합하였다. 물성 뷰는 5 물성 변수이며, 구조 뷰는 각 화합물의 Morgan/ECFP4 지문(2048-bit, radius 2)에서 산출한 Tanimoto 거리(1-유사도)를 입력으로 한다(뷰 정의는 Table 1). 이웃수 K, 커널 스케일 μ, 군집수 C의 모든 조합(총 180개 설정)을 클러스터링 하였다(Table 2). 탐색 범위는 Wang et al. (2014)의 권장($\mu \in [0.3, 0.8]$) 및 표본수 기반 이웃수 규칙($K \approx N/20 \sim N/40$)에 따라 설정하였고, 군집수 C는 과소 및 과대분할을 배제하기 위해 4–8로 사전 지정하였다. 각 설정의 군집 유효성은 군집이 형성된 본래 특성 공간에서 실루엣 계수(silhouette coefficient; Rousseeuw, 1987)로 평가하였다. 실루엣 임계값 0.25는 분석 전에 사전 고정(pre-specified)하였으며, 결과를 관찰한 뒤 조정하지 않았다. 다만 군집 존재 여부의 최종 판정의 임계값이 아니며, 대조군과의 상대 비교에 근거하여 판정하였다.

분석은 Python에서 수행하였으며, 분자 지문은 RDKit, SNF 융합은 snfpy, 실루엣 및 spectral clustering은 scikit-learn(Pedregosa et al., 2011)을 사용하였다.

Table 2. Hyperparameter search grid for the SNF fusion-network clustering

Hyperparameter	Search values	Count[a)]
K (number of neighbors)	15, 20, 25, 30	4
μ (kernel scale)	0.3, 0.5, 0.7	3
C (number of clusters)	4, 5, 6, 7, 8	5
View	Physicochemical / structural / fused	3

[a)]A total of 180 configurations were evaluated, corresponding to the $4 \times 3 \times 5 = 60$ combinations of K, μ, and C applied to each of the three views (physicochemical, structural, fused).

그룹화 경계의 안전성(ARI)

물성 공간에 자연적 경계가 있다면 그룹화 방법을 바꾸어도 유사한 경계가 재현되어야 한다. log Kp를 기준으로 400종을 네 가지 방법(등간격 구간화, 분위 구간화, log Kp 1차원 k-means, log Kp·log P·MW의 3변수 k-means; 모두 k = 5)으로 5개 그룹으로 나눈 뒤, 방법 간 그룹 라벨의 일치도를 ARI(adjusted Rand index; Hubert and Arabie, 1985)로 산출하였다.

군집 존재 여부의 종합 판정은 단일 지표에 의존하지 않고, dip과 PCA/UMAP를 핵심 판별자로, Hopkins와 SNF를 보조 지표로 삼는 규칙을 분석 전에 정하였다.

양성 및 음성 대조군에 의한 검정력 검증

군집가능성 진단 절차 자체의 검정력를 확인하기 위해 양성 및 음성 대조군을 구성하였다. 세 데이터셋(양성, 음성, 실데이터)은 표본수(N = 356), 차원(5), 변수별 값 범위를 동일하게 맞추고 분포만 달리하였다. 양성대조군은 make_blobs (scikit-learn)로 4개 중심(cluster_std = 0.8)의 군집 구조를 생성한 뒤 실데이터의 변수 범위로 척도를 맞추었으며, 음성대조군은 동일 범위의 균등분포로 생성하여 군집이 존재하지 않도록 하였다. 실제 데이터는 356종 다섯 변수에 앞서 기술한 것과 동일한 전처리(log 변환 및 z-점수 표준화)를 적용하였다. 세 데이터셋에 dip, Hopkins, PCA, UMAP, 실루엣(k = 4)을 동일하게 적용하여, 실데이터의 진단 결과가 양성·음성 대조군 중 어느 쪽과 일치하는지 비교하였다.

물성 공간의 표본 커버리지 분석

한정된 수의 표본으로 다차원 물성 공간을 대표하려면, 표본이 공간을 어느 정도 밀도로 점유하는지가 전제되어야 한다. 이를 확인하기 위해 보유 표본의 기하학적 커버리지를 차원수 증가에 따라 정량화하였다.

차원 조합은 앞서 제시한 변수 근거에 따라 중첩된 세 가지로 정의하였다.

Min-max 정규화 물성 공간에서 각 표본을 반경 r(0.05, 0.1, 0.2) 안(hypersphere)을 점유하는지 판정하였으며, void fraction을 산출하고, 최근접이웃거리(NN distance)와 동일 밀도 커버리지에 필요한 표본수(N_req)를 평가하였다. 3가지 진단이 소수 이상치에 좌우되는지 확인하기 위해 이상치를 포함한 전체 데이터와 극단값을 제거하지 않은 상하위 1% 경계값으로 대체하는 윈저화(winsorization) 방법으로 각각 분석하였다. 아울러 커버리지 분석에 투입한 5개 물성 변수가 서로 독립적인 축을 구성하는지 확인하기 위해 변수 간 Pearson 상관계수와 분산팽창계수(variance inflation factor, VIF)를 산출하여 다중공선성 여부를 점검하였다.

대표 화합물 선정

한정된 실측 자원을 고르게 배분하기 위해, 대표 물질 선

정을 space-filling 표본으로 하였다. log Kp를 오분위로 구간화한 뒤(qcut, q = 5), 각 구간 내에서 log P·MW 평면상 최원점(farthest-point) 탐욕 알고리즘(greedy algorithm; Kennard and Stone, 1969)으로 구간당 5종씩, 총 25종을 선정하였다.

최원점 선정은 기존 선택점과의 최소거리가 최대인 후보를 추가하는 절차이며, 아직 선정되지 않은 용도의 후보에 1.3배 가중치를 부여하였다. 선정 후, 실험적으로 측정 불가능한 물질(거대분자 및 금속 착물 등 초저투과 영역)을 제외하였다. 대표 25종이 무작위 선정보다 물성 스펙트럼을 고르게 포괄하는지를 흡수율(y) 없이 공간 커버리지만으로 평가하였다. log Kp를 [0, 1]로 min-max 정규화한 뒤, 모집단(극단 꼬리를 절단하지 않은 400종) 각 점에서 가장 가까운 대표까지의 거리를 구하여 그 평균과 최댓값을 커버리지 지표로 삼았다. 동일 모집단에서 25종을 무작위 추출하는 절차를 1,000회 반복하고, 두 지표를 대표 선정과 비교하였다.

재현성

분석 결과가 특정 실행 환경이나 우연에 의해 달라지는 것을 방지하기 위해 모든 무작위 절차에는 고정 시드(random_state = 42)를 적용하였다. void fraction의 참조점 추출, 대조군 생성, UMAP 투영, 실루엣 및 ARI 산출의 k-means, 무작위 선정 대조에는 고정 시드를 사용하였다. Hopkins 통계량은 참조점 추출에 따른 변동이 크므로 시드 0–19의 20회 반복 평균으로 산출하였다. 대표 선정의 최원점 알고리즘은 결정적이므로 시드 의존성이 없다. 분석은 Python(3.9.23)의 scikit-learn, diptest, umap-learn 등 표준 라이브러리로 수행하였다.

결과 및 고찰

피부투과 물성 공간의 군집 부재

5개의 물성 변수로 정의한 국내 등록 농약 원제의 피부투과 물성 공간에서는 자연적인 군집 구조가 관찰되지 않았다.

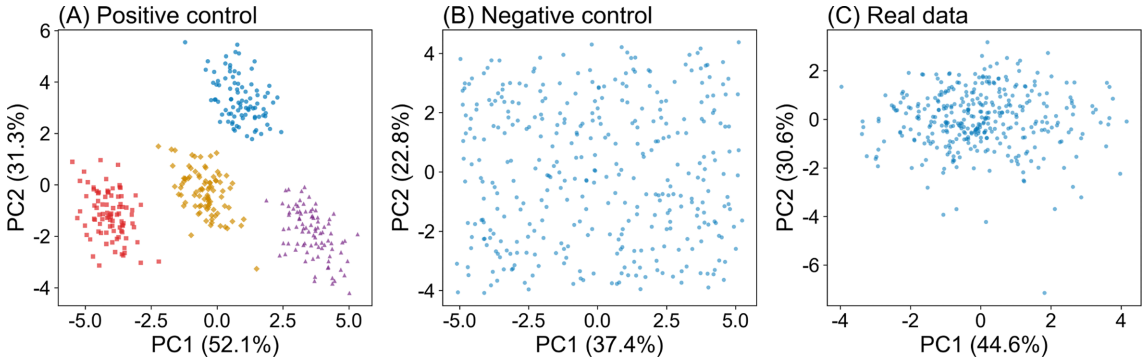


Fig. 2. Two-dimensional principal component analysis (PCA) projections of the three datasets. (A) Positive control, four Gaussian clusters generated by make_blobs; (B) negative control, a uniform random distribution; (C) real data, 356 registered pesticides described by five physicochemical properties (log P, MW, melting point, water solubility, vapour pressure) after log transformation and z-score standardization. Axis labels indicate the percentage of variance explained by each principal component. In (A), colors and marker shapes denote the four generative clusters. All three datasets share the same sample size (N = 356), dimensionality (5), and were constructed over the same per-variable value range.

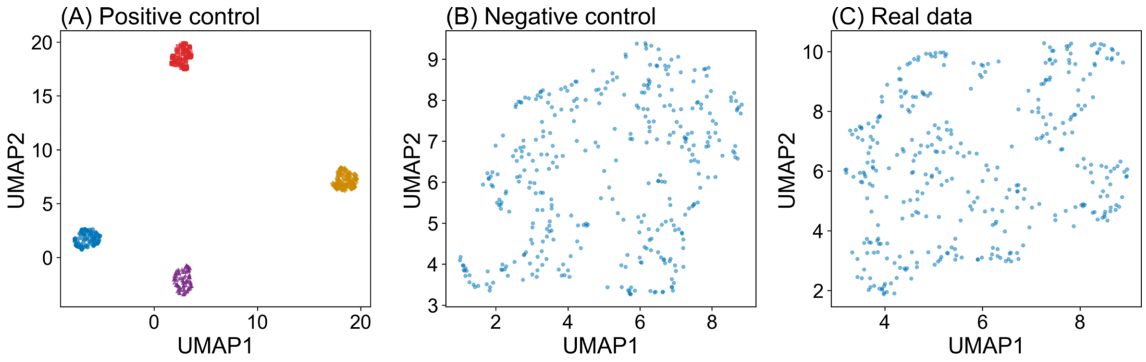


Fig. 3. Two-dimensional UMAP projections of the three datasets. (A) Positive control, four Gaussian clusters generated by make_blobs; (B) negative control, a uniform random distribution; (C) real data, 356 registered pesticides described by five physicochemical properties after log transformation and z-score standardization. In (A), colors and marker shapes denote the four generative clusters. UMAP parameters: n_neighbors = 15, min_dist = 0.1. All three datasets share the same sample size (N = 356), dimensionality (5), and per-variable value range. UMAP axis coordinates are dimensionless and their absolute scale carries no direct interpretation.

다봉성 검정, 차원축소 시각화, 다중뷰 군집성 평가 및 그룹 일치도 분석의 네 가지 독립적인 진단은 모두 일관되게 군집 부재(clusterability absence)를 지지하였다.

먼저, 다변량 자료를 1차원으로 축소하여 수행한 다봉성 검정에서도 군집의 증거는 확인되지 않았다. 제1주성분에 대한 dip 검정(PCA-dip)과 두 점 사이 거리 기반 dip 검정(Dip-dist)은 모두 단봉 분포를 나타냈으며(각각 $p = 0.965$, $p = 0.991$), 개별 물성 변수의 주변분포 역시 모두 단봉으로 나타났다($p = 0.96–0.99$). 특히 dip-dist는 좌표가 아닌 두 점 사이 거리를 이용하므로 투영 방향에 영향을 받지 않으며, 방향성과 무관한 군집성 검정으로 알려져 있다(Adolfsson et al., 2019; Kalogeratos and Likas, 2012). 따라서 축 선택이나 투영 방향에 관계없이 자료는 하나의 연속적인 분포를 형성하는 것으로 판단되었다.

차원축소 결과에서도 동일한 양상이 확인되었다(Figs. 2 and 3). PCA에서는 분리된 집단 없이 하나의 연속적인 점구름(point cloud)이 형성되었으며, 비선형 구조를 보존하는 UMAP에서도 명확한 군집이나 경계는 관찰되지 않았다. 즉, 선형 및 비선형 차원축소 모두에서 자료는 독립적인 하위 집단보다 연속적인 물성 스펙트럼의 형태를 나타냈다.

반면 Hopkins 통계량은 0.951로 계산되어 자료가 물성 공간 전체에 균등하게 분포하지 않고 특정 영역에 집중되어 있음을 시사하였다. 그러나 Hopkins 통계량은 공간적 비균등성(spatial inhomogeneity)을 평가하는 지표로, 높은 값은 데이터가 균등하게 분포되어 있지 않음을 의미할 뿐 자연적인 군집의 존재를 직접 의미하지는 않는다. 따라서 본 결과는 물질이 일부 물성 영역에 밀집되어 있음을 보여주지만, 앞선 군집성 진단 결과와 모순되지는 않는다.

군집 부재가 선택한 5개의 물성 변수에 의존한 결과인지를 확인하기 위해 구조 지문(fingerprint)과 물성 정보를 결합한 SNF 기반 다중뷰 분석을 수행하였다. 사전에 설정한 실루엣 임계값(0.25)을 기준으로 180개의 매개변수 조합을 평가한 결과, 임계값을 만족한 경우는 없었다(0/180). 구조 뷰 단독의 최대 실루엣 계수는 0.042, 융합망에서는 0.035로 모두 기준값의 5분의 1 이하 수준에 머물렀다. 즉, 화학구조 정보를 추가한 경우에도 자연적인 군집은 형성되지 않았으며, 이러한 결과는 군집 부재가 특정 물성 변수의 선택에 따른 산물이 아님을 시사한다.

또한 네 가지 그룹화 방법(등간격, 분위수, 1차원 k-means, 3차원 k-means; $k = 5$)으로 생성한 그룹 라벨 간, ARI는 0.21–0.52 범위로 나타났다. 서로 다른 방법 간 그룹 일치도가 낮았다는 것은 자료에 안정적이고 재현 가능한 군집 구조가 존재하지 않음을 의미하며, 앞선 군집성 진단 결과를 뒷받침하였다.

종합하면, 국내 등록 농약 원제의 피부투과 물성 공간은 서로 분리된 자연 군집이 아니라 연속적인 물성 스펙트럼으로 구성된 것으로 판단되었다. 따라서 대표 물질 선정은 군집별 대표를 추출하는 방식보다 물성 공간을 균등하게 덮는 space-filling 전략이 적합하며, 피부흡수율 예측 역시 분류 기반 접근보다 연속형 회귀 모델을 적용하는 것이 타당함을 시사한다.

양성 및 음성 대조군에 의한 진단 검정력 검증

군집 부재 결론이 진단 절차의 검정력 부족에서 비롯된 것이 아님을 확인하기 위해, 양성 및 음성 대조군과 실데이터를 동일 조건($N = 356$, 5차원, 변수 범위 동일)에서 비교하였다(Table 3, Figs. 2 and 3).

핵심 판별자 3 가지(dip, PCA/UMAP, 실루엣)에서 실데이터는 음성 대조군과 동일한 양상이었지만, 양성 대조군은 상반된 양상을 보였다. 특히 실루엣 계수는 군집이 정의상 존재하지 않는 음성 대조군에서도 군집수를 $k = 4$로 강제하였을 때 0.192으로 산출되었으며, 실데이터의 0.231은 이 값에 근접하였다. 동일한 검증 절차에서 양성 대조군에서는 군집을 명확히 검출하였으므로(다봉 2/5, 4개 섬 분리, 실루엣 0.576), 실데이터의 "군집 부재"에 대한 결론은 검정의 둔감이 아니라 실제 구조의 부재에 기인함을 확인하였다.

한편 Hopkins 통계량은 실데이터(0.951)가 군집이 명확한 양성 대조군(0.998)과 거의 같은 수준으로 높았으나 군집이 없는 음성 대조군(0.472)과는 크게 달랐다. 즉 실데이터는 군집이 없음에도(dip 다봉 0/5, 실루엣 0.231) 양성대조군 군집 데이터에 필적하는 높은 Hopkins 값을 보임으로써 비균등성을 확인할 수 있었으나, 비균등의 원인이 군집인지

Table 3. Comparison of clusterability diagnostics for the positive control, negative control, and real data

Dataset[a)]	Dip multimodal variables[b)]	Hopkins	PCA 2D explained variance[c)]	Silhouette (k = 4)	PCA/UMAP	Verdict
Positive control (clustered)	2/5	0.998	0.834	0.576	Four separated islands	Clusters detected
Negative control (uniform)	0/5	0.472	0.602	0.192	Single cloud	No clusters
Real data	0/5	0.951	0.752	0.231	Single cloud	No clusters

[a)]All three datasets share the same sample size ($N = 356$), dimensionality (5), and per-variable value range; they differ only in their distribution. [b)]Number of variables (of five) with a dip-test p-value < 0.05. [c)]Proportion of total variance explained by the first two principal components.

밀집인지는 구분되지 않았다. 이는 Hopkins가 균등성 여부를 나타내는 지표이기 때문이며, 단독 판별 지표로 사용해서는 안 됨을 대조군 실험으로 확인하였다. 다만 이는 Hopkins가 판정에 기여하지 못한다는 뜻이 아니라, 군집이 부재한 상태에서 관찰된 이러한 치우침은 무작위 추출이 밀집 영역에 쏠릴 수 있음을 뜻하며, 대표 물질 선정 방식과 직접 관련된다.

커버리지 분석

군집을 이루지 않으면서 특정 영역에 밀집한 구조에서는 집단별 층화도 무작위 추출도 적절하지 않았다. 군집이 없는 경우에는 나눌 경계가 없고, 밀집한 구조의 경우 무작위 추출이 밀집 영역을 과표집한다. 따라서 공간 전체를 고르게 덮는 설계를 채택하였다. 이에 보유 표본(356종)이 물성 공간을 차원이 증가함에 따라 충분히 포괄하는지를 3가지 진단으로 평가하였다.

Void fraction

표본으로부터 반경 r안에 들지 않는 빈 영역의 비율은 차원이 증가할수록 급격히 상승하여, 5차원에서 약 99%였다(전체 99.1%, 윈저화 98.9%; 반경 0.1 기준). 2차원과 3차원의 void fraction은 각각 전체 39.1% 및 74.6%로, 차원 증가에 따라 급격히 상승하였다. 윈저화 적용 시의 5차원에서도 유사한 값으로 수렴하였다(Fig. 4).

최근접이웃거리

점 간 최근접이웃거리는 차원 증가에 따라 증가하였으며, 차원수에 따른 거리 증가를 √(차원수)로 보정한 후에도 증가 경향이 유지되었다(Fig. 5).

필요 표본수

동일 밀도의 충진에 필요한 표본수는 차원 증가에 따라

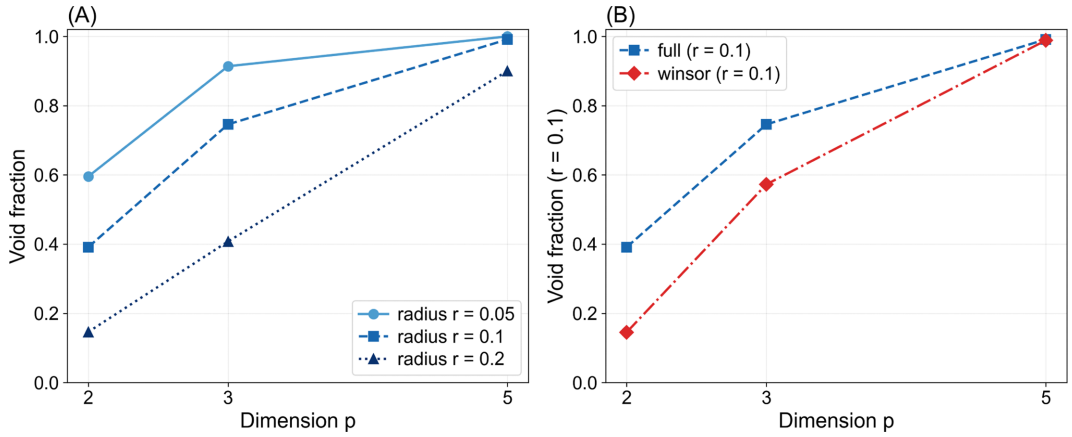


Fig. 4. Void fraction of the physicochemical space as a function of dimensionality (p = 2, 3, 5). The void fraction is the proportion of random reference points in the min-max normalized space that lie farther than radius r from every sample. (A) Full data at three radii (r = 0.05, 0.1, 0.2). (B) Full data versus the winsorized version (top/bottom 1% clipped) at r = 0.1. N = 356.

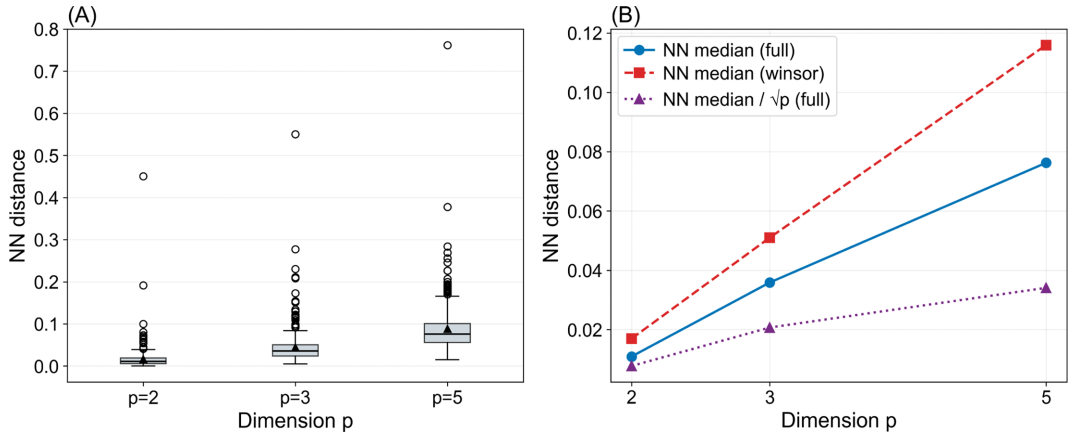


Fig. 5. Nearest-neighbor (NN) distance in the min-max normalized space as a function of dimensionality (p = 2, 3, 5). (A) Distribution of per-point NN distances for the full data (boxplots; black triangles indicate the mean, open circles indicate outliers). (B) Median NN distance for the full and winsorized (top/bottom 1% clipped) data, together with the axis-corrected median (median / √p) for the full data. N = 356.

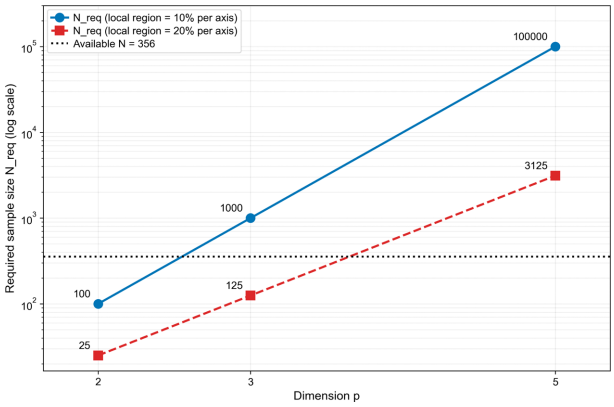


Fig. 6. Required sample size (N_req) to place one point per local region as a function of dimensionality (p = 2, 3, 5), on a logarithmic scale, compared with the available sample (N = 356; dotted line). N_req is shown for local-region sizes of 10% and 20% per axis.

증가하여, 5차원에서는 추정 모집단 규모(국내 등록농약 농촌진흥청 농약안전정보시스템 2025년 6월 24일 기준)를 넘었다(Fig. 6).

Void fraction과 필요한 표본수가 차원에 따라 급격히 증가한 양상은 차원의 저주로 알려진 현상과 부합하며, 차원이 늘수록 공간의 부피가 기하급수적으로 커지기 때문이다(Ronan et al., 2016). Space-filing 설계의 목표가 공간의 어느 지점에서도 가까운 설계점이 존재하도록 배치하는 데 있음을 고려하면(Joseph, 2016), 차원이 높아질수록 많은 표본이 필요해진다. 따라서 피부 투과성을 정의하는 5변수 공간을 균등하게 채우는데 필요한 표본수는 국내 등록 농약 원제 전체를 동원하여도 도달할 수 없는 규모이기 때문에 대표 선정은 이보다 낮은 차원의 축에서 수행되어야 한다. 표본을 늘려도 목표 수준에 도달하지 못하는 경우가 존재한다는 점은 다른 분야의 표집 설계 연구에서도 보고되어 있다(He et al., 2022).

아울러 커버리지 분석의 5 물성 변수가 실질적으로 서로 다른 축을 구성하는지 확인한 결과, log P와 수용해도만이 강한 음의 상관을 보였고(r = -0.81) 나머지 아홉 변수쌍은 모두 |r| < 0.56였다(Table 4). 모든 변수의 VIF가 5 미만(최대 4.40, 수용해도)으로 다중공선성은 관찰되지 않았다.

대표 물질 선정

log Kp 오분위 구간별로 log P·MW 평면상 최원점 선정을 수행하여 총 25종을 선정하였다(Fig. 7). 선정된 대표 물질의

Table 4. Pearson correlation matrix and variance inflation factors (VIF) among the five physicochemical variables

Variable	log P	MW	Melting point	Water solubility[a)]	Vapour pressure[a)]	VIF[b)]
log P	1.00	0.39	-0.24	-0.81	0.12	4.10
MW	0.39	1.00	0.07	-0.56	-0.42	1.80
Melting point	-0.24	0.07	1.00	0.05	-0.34	1.38
Water solubility[a)]	-0.81	-0.56	0.04	1.00	0.12	4.40
Vapour pressure[a)]	0.12	-0.42	-0.31	0.12	1.00	1.50

Notes: N = 356. Diagonal entries are self-correlations (1.00). [a)]Water solubility and vapour pressure were $\log_{10}$-transformed prior to correlation, consistent with the coverage analysis. [b)]Variance inflation factor, obtained by regressing each variable on the remaining four; VIF > 5 is conventionally taken as an indication of multicollinearity.

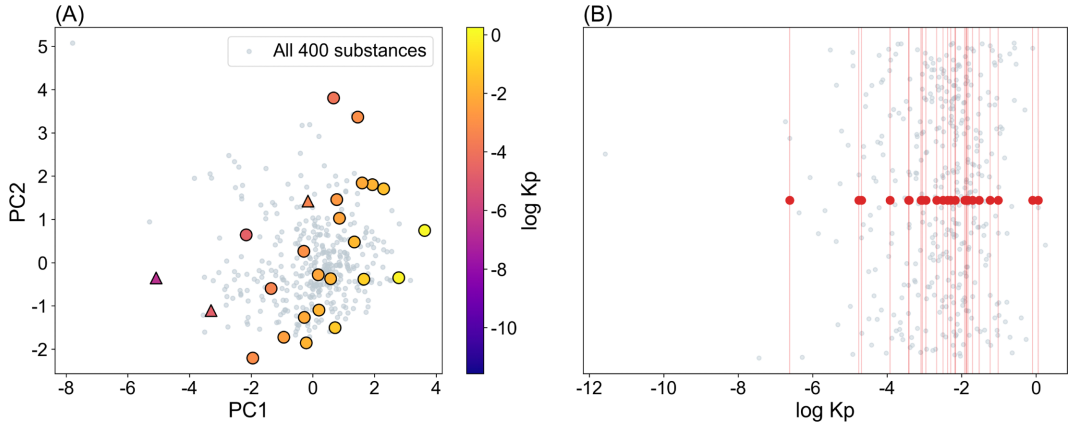


Fig. 7. Coverage of the 25 selected representative compounds. (A) Principal component analysis (PCA) of the population in the log P-MW-log Kp space; gray points denote all 400 substances and the enlarged markers denote the 25 representatives, colored by log Kp. Circles indicate compounds retained from the initial farthest-point selection; triangles indicate compounds replaced by experimentally feasible substitutes within the same log Kp quintile. (B) Distribution of the 25 representatives (red points and vertical lines) along the log Kp axis, overlaid on the population (gray points).

log Kp 분포는 -6.62 ~ +0.06으로 5구간 전체를 포괄하였며, 물성값은 Table 5에 제시하였다.

Table 5. Physicochemical properties of the 25 representative compounds selected by space-filling sampling across the five log Kp quintiles.

Quintile	Active ingredient	log P	MW	log Kp
Q1	glufosinate-P	-3.96	181.13	-6.62
Q1	mesotrione	0.11	339.32	-4.69
Q1	daminozide	-1.53	160.17	-4.76
Q1	flubendiamide	4.14	682.39	-3.92
Q1	propyrisulfuron	2.90	455.88	-3.42
Q2	hexazinone	1.17	252.31	-3.41
Q2	hymexazol	0.30	99.15	-3.09
Q2	broflanilide	5.20	663.30	-3.05
Q2	hexythiazox	2.67	352.88	-2.96
Q2	cycloprothrin	4.19	482.36	-2.67
Q3	metolcarb	1.70	165.21	-2.50
Q3	cyflumetofen	4.30	447.45	-2.38
Q3	penflufen	3.30	317.41	-2.29
Q3	methabenzthiazuron	2.64	221.27	-2.18
Q3	pyflubumide	5.34	535.52	-2.18
Q4	metconazole	3.85	319.83	-1.92
Q4	chlorfluazuron	5.80	540.65	-1.88
Q4	forchlorfenuron	3.30	247.68	-1.87
Q4	dichlobenil	2.70	172.01	-1.83
Q4	imibenconazole	4.94	411.70	-1.70
Q5	acrinathrin	6.30	541.44	-1.53
Q5	cyprodinil	4.00	225.29	-1.23
Q5	oxadiazon	5.33	345.20	-1.02
Q5	etofenprox	6.90	376.49	-0.10
Q5	pyridalyl	8.10	491.12	0.06

실험적으로 측정 불가능한 저투과 물질(거대분자 및 금속 착물 등)을 제외함에 따라 log Kp 하한이 -11.6에서 -6.6으로 상승하였으며, 이 영역은 모델의 외삽 한계로 명시하였다. 최원점 알고리즘은 결정적이므로 선정 결과는 시드에 무관하며, 실험 부적합 물질 교체 후에도 5구간 전체의 대표성과 물성 공간 분포가 유지되었다.

선정된 대표 25종의 커버리지 우수성은 무작위 선정과의 비교로 확인되었다(Fig. 8). log Kp 일차원 축에서 평균 커버 거리는 대표 선정이 0.0087로 무작위 추출 1,000회 평균(0.0114) 보다 낮았다. 최대 빈틈 또한 대표 선정이 0.4193로 무작위 평균(0.4959)보다 낮았다. 두 지표 모두 임의의 범위 절단 없이 극단 꼬리를 포함한 400종 전체에서 산출되었다.

종합하면, 국내 등록 농약 원제의 피부투과 관련 물성공간에 뚜렷한 군집 구조가 존재하지 않음을 다섯 가지 진단과 양성 및 음성 대조군을 통해 확인하였다. 또한, 군집이 부재한 연속 물성 공간에서 한정된 실측 자원을 배분하는 문제를 space-filling 표본 설계로 정식화하여, log Kp 오분위 구간을 포괄하는 25종의 대표 물질을 도출하였다. 무작위 선정과의 커버리지 비교를 통해 이 선정이 임의 추출보다 물성 스펙트럼을 더 균등하게 대표함을 확인하였다. 다만 선정이 log Kp 축에서 수행되었으므로 5변수 공간 전체에서의 균등성은 보장되지 않으며, 이는 표본이 확충된 이후의 과제로 남는다. 이 결과는 후속 동물대체 시험법의 실측 대상 선정 및 예측 모델 구축의 학습자료로 활용된다. 또한 분석에 사용된 농약의 물성은 다른 위해성 평가 분야에서도 공통적으로 이용되는 항목이므로, 군집 부재라는 진단 결과는 피부흡수 이외의 평가 영역에도 활용될 수 있다.

감사의 글

본 연구는 2026년 농촌진흥청 국립농업과학원 농업과학

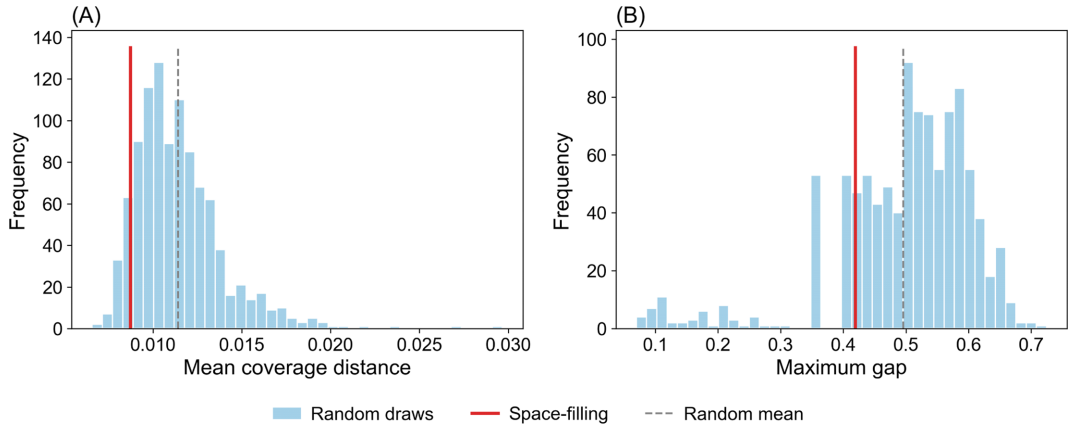


Fig. 8. Coverage of the space-filling selection versus random selection along the log Kp axis, evaluated on the full population of 400 substances (including the extreme tails) in the min-max normalized [0,1] space. (A) Mean coverage distance and (B) maximum gap; for both metrics, lower values indicate better coverage. Histograms show the distribution over 1,000 random draws of 25 compounds (seed 42); the red solid line marks the space-filling selection and the gray dashed line marks the mean of the random draws.

기술 연구개발사업 ‘피부흡수율 동물대체시험법 개발을 통한 농약노출 위해성평가기준 개선(PJ017781)’에서 지원받아 수행되었으며, 2026년도 농촌진흥청 국립농업과학원 전문연구원 과정 지원사업에 의해 이루어진 것입니다.

Author Information and Contributions

Ji-Young Shin, Toxicity and Risk Assessment Division, Department of Agro-food Safety, National Institute of Agricultural Sciences, Wanju 55365, Korea, Researcher, https://orcid.org/0000-0003-3177-161X, Conceptualization, methodology, formal analysis, visualization, writing-original draft and project administration.

So-Hye Hong, Toxicity and Risk Assessment Division, Department of Agro-food Safety, National Institute of Agricultural Sciences, Wanju 55365, Korea, Researcher, https://orcid.org/0000-0001-9958-5552, Conceptualization, methodology, investigation.

Si-Young Yang, Toxicity and Risk Assessment Division, Department of Agro-food Safety, National Institute of Agricultural Sciences, Wanju 55365, Korea, Senior Researcher, https://orcid.org/0009-0000-7308-969X, Conceptualization, methodology and editing.

Sol-A Kim, Toxicity and Risk Assessment Division, Department of Agro-food Safety, National Institute of Agricultural Sciences, Wanju 55365, Korea, Researcher, https://orcid.org/0000-0002-4055-7511, Data curation, validation and editing.

You-Mi Jo, Toxicity and Risk Assessment Division, Department of Agro-food Safety, National Institute of Agricultural Sciences, Wanju 55365, Korea, Senior Researcher, https://orcid.org/0009-0008 6733-7181, Investigation, software, editing.

Mi-Hye Jeong, Toxicity and Risk Assessment Division, Department of Agro-food Safety, National Institute of Agricultural Sciences, Wanju 55365, Korea, Senior Researcher, https://orcid.org/0009-0007-8724-1059, Validation, writing-review and editing.

Ji-Hyun Moon, Toxicity and Risk Assessment Division, Department of Agro-food Safety, National Institute of Agricultural Sciences, Wanju 55365, Korea, Researcher, https://orcid.org/0009-0002-9245-3113, Data curation and editing.

Soo-Jin Park, Agro-Materials Industry Division, Research Policy Bureau, Rural Development Administration, Jeonju 54875, Korea, Researcher, https://orcid.org/0000-0002 2522-3185, Conceptualization, methodology, writing-review and editing.

Tae-Gon Kim, Department of Smart Farm, College of Agricultural and Life Sciences, Jeonbuk National University, Jeonju 54896, Korea, Professor, https://orcid.org/0000-0002-7931-6627, Conceptualization, methodology, validation, writing-review and editing and supervision.

이해상충관계

저자는 이해상충관계가 없음을 선언합니다.

Literature cited

Adolfsson A, Ackerman M, Brownstein NC, 2019. To cluster, or not to cluster: an analysis of clusterability methods. Pattern Recognit. 88:13-26.

Daszykowski M, Walczak B, Massart DL, 2002. Representative subset selection. Anal. Chim. Acta 468(1):91-103.

Eleftheriadou D, Luette S, Kneuer C, 2019. In silico prediction of dermal absorption of pesticides - an evaluation of selected models against results from in vitro testing. SAR QSAR Environ. Res. 30(8):561-585.

European Food Safety Authority (EFSA), 2017. Guidance on dermal absorption. EFSA J. 15(6):4873.

Golbraikh A, Tropsha A, 2002. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J. Comput.-Aided Mol. Des. 16(5-6):357-369.

Hartigan JA, Hartigan PM, 1985. The dip test of unimodality. Ann. Stat. 13(1):70-84.

He J, Zhou X, Matsui T, Li F, Tanaka TST, 2022. Critical reevaluation of an efficient sampling design for assessing soil properties using bootstrap sampling and geostatistical analysis in Japanese large-scale paddy fields. Soil Sci. Plant Nutr. 68(5-6):536-546.

Hopkins B, Skellam JG, 1954. A new method for determining the type of distribution of plant individuals. Ann. Bot. 18(2):213-227.

Hubert L, Arabie P, 1985. Comparing partitions. J. Classif. 2(1):193-218.

Joseph VR, 2016. Space-filling designs for computer experiments: a review. Qual. Eng. 28(1):28-35.

Kalogeratos A, Likas A, 2012. Dip-means: an incremental clustering method for estimating the number of clusters. Adv. Neural Inf. Process. Syst. 25:2393-2401.

Kar S, Amin SA, Li L, 2025. Unveiling chemical space, scaffold diversity, critical structural features of pesticides: a comprehensive QSAR, qRASAR, machine learning studies

to predict pesticides toxicity. Sci. Total Environ. 1001:180489.
Kennard RW, Stone LA, 1969. Computer aided design of experiments. Technometrics 11(1):137-148.
Kobak D, Linderman GC, 2021. Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nat. Biotechnol. 39(2):156-157.
Krieger RI, 1995. Pesticide exposure assessment. Toxicol. Lett. 82/83:65-72.
Martin TM, Harten P, Young DM, Muratov EN, Golbraikh A, et al., 2012. Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model. 52(10):2570-2578.
McInnes L, Healy J, Melville J, 2018. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426.
Organisation for Economic Co-operation and Development (OECD), 2004. Guidance document for the conduct of skin absorption studies. Series on Testing and Assessment No. 28. Paris, France.
Organisation for Economic Co-operation and Development (OECD), 2022. Guidance notes on dermal absorption. Series on Testing and Assessment No. 156, 2nd ed. Paris, France.
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12:2825-2830.
Potts RO, Guy RH, 1992. Predicting skin permeability. Pharm. Res. 9(5):663-669.
Renai L, Turkina V, Hulleman T, Nikolopoulos A, Gargano AFG, et al., 2025. A novel chemical-space-dependent strategy for compound selection in non-target LC-HRMS method development using physicochemical and structural data. Environ. Sci. Technol. Lett. 12(9):1162-1168.
Ronan T, Qi Z, Naegle KM, 2016. Avoiding common pitfalls when clustering biological data. Sci. Signal. 9(432):re6.
Rousseeuw PJ, 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20:53-65.
Rural Development Administration (RDA), 2026. Registration criteria for pesticides and active ingredients. RDA Notification No. 2026-15, Annex 12. Jeonju, Korea. (In Korean)
Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, et al., 2014. Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods 11(3):333-337.
Waters LJ, Quah XL, 2022. Predicting skin permeability using HuskinDB. Sci. Data 9:584.
Wright K, 2022. Will the real Hopkins statistic please stand up? R J. 14(3):282-292.
You AS, Lee N, Park S, Jo Y, Oh JA, et al., 2017. Estimation of exposure dose and occupational risk assessment with various dermal absorption default in spraying of plant protection products. Korean J. Pestic. Sci. 21(4):434-452. (In Korean)

한국 등록농약 피부투과 물성공간의 군집부재 진단과 space-filling 기반 대표 물질 선정

신지영[1] · 홍소혜[1] · 양시영[1] · 조유미[1] · 김솔아[1] · 정미혜[1] · 문지현[1] · 박수진[2*†] · 김태곤[3*†]

[1]농촌진흥청 국립농업과학원 농산물안전성부 독성위해평가과, [2]농촌진흥청 연구정책국 농자재산업과
[3]전북대학교 농업생명과학대학 스마트팜학과

요 약 농약의 위해성 평가의 핵심 입력값인 피부흡수율은 시험 비용이 높아 컴퓨터 기반 예측 모델의 필요성이 증가하고 있다. 모델 학습에 사용될 실측 대상을 어떻게 고를 것인가는 모델 개발에 선행하는 설계 문제이다. 선정 전략은 물성 공간의 구조에 따라 달라지므로, 본 연구는 5개 물성 정보를 모두 확보할 수 있는 국내 등록 농약 원제 356종이 대상이며, 피부투과 물성 공간의 군집 구조 존재 여부를 원리가 다른 다섯 가지 진단으로 교차 평가하였다. 진단 절차의 타당성을 검증하기 위해 표본수, 차원, 변수 범위를 실제 데이터와 동일하게 맞춘 인공 양성 및 음성 대조군에 같은 절차를 적용하였다. 적용한 변수와 진단 범위에서 뚜렷하고 안정적인 군집 구조를 지지하는 증거는 확인되지 않았으며(다봉성 검정 단봉, 다중뷰 융합 실루엣 임계값 통과 0/180, 그룹화 방법간 일치도 0.21–0.52), 실데이터의 실루엣(0.231)이 음성 대조군(0.192)에 근접한 반면, 양성 대조군은 군집을 명확히 검출하여(0.576) 실데이터의 결과가 음성 대조군에 가까웠으며, 적용한 진단 절차가 명확히 분리된 군집을 검출할 수 있음을 양성 대조군에서 확인하였다. 다만 Hopkins 통계량은 0.951로 물질이 특정 영역에 치우쳐 있음을 나타내었고, 변수별 min-max 범위로 정의한 5차원 공간의 void fraction은 약 99%로 나타나, 현재 표본으로 해당 공간을 균등하게 충진하기 어려운 것으로 평가되었다. 이에 예측 log Kp를 오분위로 층화한 후 각 구간의 logP-MW 공간에서 최원점 알고리즘을 적용하여 25종을 선정하였으며, 두 커버리지 지표 모두 단순 무작위 추출 1,000회의 평균보다 낮았다. 본 연구에서 제안한 선정 25종 및 전략은 제한된 실측 자원을 효율적으로 활용하여 후속 피부흡수시험 대상의 우선순위 설정과 예측모델 개발을 위한 초기 학습자료 구축에 활용될 수 있다.

색인어: Clusterability, 피부흡수율, 농약, In silico, Space-filling design