ARTICLE

Physicochemical properties determining drug detection in skin

Wout Bittremieux[1,2,3] | Rohit S. Advani[1,2,4] | Alan K. Jarmusch[1,2,5] |
Shaden Aguirre[1,2] | Aileen Lu[1,2] | Pieter C. Dorrestein[1,2] | Shirley M. Tsunoda[1]

[1]Skaggs School of Pharmacy and
Pharmaceutical Sciences, University
of California San Diego, La Jolla,
California, USA

[2]Collaborative Mass Spectrometry
Innovation Center, University of
California San Diego, La Jolla,
California, USA

[3]Department of Computer Science,
University of Antwerp, Antwerpen,
Belgium

[4]Department of Chemistry, University
of Massachusetts-Amherst, Amherst,
Massachusetts, USA

[5]Immunity, Inflammation, and Disease
Laboratory, Division of Intramural
Research, National Institute of
Environmental Health Sciences,
National Institutes of Health, Research
Triangle Park, North Carolina, USA

Correspondence
Shirley M. Tsunoda, Skaggs School
of Pharmacy and Pharmaceutical
Sciences, University of California San
Diego, 9500 Gilman Drive, MC 0657, La
Jolla, CA 92093-0657, USA.
Email: smtsunoda@health.ucsd.edu

Funding information
W.B. was supported by the Research
Foundation – Flanders (12W0418N).
This research was supported in
part by the Intramural Research
Program of National Institute of
Environmental Health Sciences of the
National Institutes of health (NIH;
ES103363-01).

Abstract

Chemicals, including some systemically administered xenobiotics and their bio-transformations, can be detected noninvasively using skin swabs and untargeted metabolomics analysis. We sought to understand the principal drivers that determine whether a drug taken orally or systemically is likely to be observed on the epidermis by using a random forest classifier to predict which drugs would be detected on the skin. A variety of molecular descriptors describing calculated properties of drugs, such as measures of volume, electronegativity, bond energy, and electrotopology, were used to train the classifier. The mean area under the receiver operating characteristic curve was 0.71 for predicting drug detection on the epidermis, and the SHapley Additive exPlanations (SHAP) model interpretation technique was used to determine the most relevant molecular descriptors. Based on the analysis of 2561 US Food and Drug Administration (FDA)-approved drugs, we predict that therapeutic drug classes, such as nervous system drugs, are more likely to be detected on the skin. Detecting drugs and other chemicals noninvasively on the skin using untargeted metabolomics could be a useful clinical advancement in therapeutic drug monitoring, adherence, and health status.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?
A multitude of chemicals that an individual encounters in daily life can be detected on the skin surface using untargeted metabolomic analysis, including topical and systemically administered xenobiotics.

WHAT QUESTION DID THIS STUDY ADDRESS?
Can machine learning be used to predict whether systemically administered drugs are observed on the epidermis and provide insights into the complex underlying biochemical processes?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?
Our machine-learning model found relevant molecular descriptors related to volume, electronegativity, bond energy, and electrotopology to be strong predictors

Wout Bittremieux and Rohit S. Advani contributed equally to this work.

of drug observance on the epidermis. Our model predicted that certain categories of drugs, such as those affecting the nervous system, are more likely to be observed.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

Our machine-learning model demonstrated several physicochemical properties of drugs that predict detection on noninvasively obtained skin swabs. Using skin swabs may be a paradigm shift in how we monitor drugs, measure drug adherence, and monitor health and disease noninvasively.

INTRODUCTION

The skin provides a physical and chemical barrier to environmental insults and supports immunological function and thermoregulation. Additionally, the bacteria, viruses, and fungi that comprise the skin microbiome provide an essential function in protection against microbial pathogens, educating the immune system, and breaking down products.[1] Traditionally, topical formulations of drugs are desired in certain medical conditions to either deliver drugs from the skin to the systemic circulation (e.g., transdermal scopolamine) or to deliver drugs locally to the skin and minimize the systemic toxicity of these drugs (e.g., topical corticosteroids). Interestingly, a recent study demonstrated systemic concentrations above the US Food and Drug Administration (FDA) safety threshold of the sunscreen compounds avobenzone, oxybenzone, and octocrylene up to 21 days postadministration, despite the widespread assumption that these commonly used topical products are considered "safe."[2]

Skin permeation by xenobiotics has been investigated for many years.[3] The majority of drugs permeate across the bulk of the epidermis, along a concentration gradient, in three pathways—intracellular, intercellular, and follicular—with the intercellular pathway believed to provide the principal route for drug permeation.[4] Several mechanistic studies have shown that extracellular lacunar domains comprise a pore pathway for penetration of polar and nonpolar molecules across the stratum corneum. Following penetration across the stratum corneum, drugs diffuse across the viable epidermis and dermis and are carried away into the bloodstream by the capillaries of the dermis. Ideal drug candidates for permeation through the skin include those with low molecular weight, solubility in water and oils to achieve an appropriate concentration gradient, an elevated but balanced partition coefficient, and a low melting point for solubility purposes.[5] Currently, there exist several models to predict skin permeability and data resources of experimental skin permeation values and their corresponding protocols.[6]

In contrast, less is known about "inverse penetration": drugs moving from the systemic circulation to the epidermis. Patzelt et al. suggested five possible inverse penetration pathways, consisting of: (i) the intracellular; (ii) intercellular; and (iii) follicular pathways, similar to topically administered substances; and (iv) inverse penetration via sweat, or (v) via the desquamation process.[7] Based on a literature study of 11 systemically administered substances and their recovery in the skin, they concluded that lipophilic substances predominantly reach the skin surface via the sebum, whereas hydrophilic substances utilize the sweat for delivery to the skin surface. Inverse cellular penetration and desquamation—which occurs on longer time scales—are less relevant, although no indication for inverse intracellular penetration was found in the literature. Concentrations on the epidermis of a few systemically administered compounds have been reported in the literature.[7,8] For example, the antifungal agent fluconazole was detected in significantly higher concentrations in the stratum corneum than in plasma and for a longer duration after cessation of therapy.[9] Similarly, other systemically administered antifungal agents were detected in high concentrations on the skin and exhibited slow clearance from both skin and nails.[8]

Comprehensive knowledge of the underlying mechanisms is relevant in the dermatological field, as a multiplicity of pharmaceutics are administered systemically to address skin disorders, and would enable important applications of xenobiotic skin detection using noninvasive methods to determine adherence of drugs, for therapeutic drug monitoring, the extent of metabolism, and to assess organ and health status. Skin swab samples have previously been used to determine individual skin chemistry profiles, including topically applied chemicals, such as avobenzone, octocrylene, as well as others found in soap, lotions, cosmetics, and anti-mosquito sprays and lotions.[10] Furthermore, we have recently demonstrated that systemically administered drugs, such as citalopram, diphenhydramine, and the N-acetyl metabolite of sulfamethoxazole, can be detected in skin swab samples of the hands, forearm, forehead, and axilla.[11] Utilizing

untargeted metabolomics and analysis of these data using the Global Natural Products Social Molecular Networking (GNPS)[12] infrastructure, we achieved the detection of these compounds on the epidermis of patients that were prescribed these drugs, thus concluding that systemically administered drugs can be detected on the skin surface. Additionally, our recent study in healthy humans demonstrated a delayed time course between plasma and skin concentrations of diphenhydramine and its metabolites ranging from 1.5 to 10 h (https://biorxiv.org/cgi/content/short/2021.11.22.469638v1).

The full mechanism and pathways of chemicals and drugs moving from the systemic circulation to the epidermis are unknown. Additionally, not all xenobiotics can be detected on the skin. A notable example is the immunosuppressive drug tacrolimus, which was not detected in our skin swab samples.[11] We sought to understand the physicochemical and pharmacokinetic properties that allow some systemically administered drugs to be detected on the epidermis and not others. Using existing skin swab data, we trained a random forest classifier that is able to accurately predict whether a compound will be observed on the epidermis.

METHODS

Data origin

No human subjects were recruited for this study and all data were assessed retrospectively. All data were anonymous and obtained from open mass spectrometry data contained in GNPS and ReDU. GNPS is a public data repository and analysis infrastructure for untargeted metabolomics data.[12] Analysis tools available on GNPS include spectral library searching, molecular networking to compute spectral similarities between tandem mass spectrometry (MS/MS) spectra and detect related compounds,[13] and MASST to query MS/MS spectra against all public metabolomics data and associated sample information to investigate their context.[14] ReDU is an associated system to capture metadata of public data in GNPS using validated controlled vocabularies (redu.ucsd.edu).[15]

To collect drugs that are observable on the epidermis, MS data and associated metadata were selected using ReDU (March 24, 2019) by filtering for files that were annotated as pertaining to human skin samples using the Uberon ontology of anatomy terms[16] (Table S1). This resulted in a list of 5629 files from a heterogeneous set of 20 previously performed studies with data deposited to GNPS (Table S2).

Additionally, prescription records available in conjunction with data from a previous kidney transplant study[11] were used to define drugs which were prescribed to

individuals but were not observed in skin samples in that study (GNPS/MassIVE dataset identifier MSV000081548). Skin samples were obtained from 15 individuals at two different clinic visits—without regard to timing with their medications—on 10 locations on the body (bilateral collection of the forehead, nasolabial area, axillary, backhand, and palm). The subjects of that study were prescribed many (>5) medications simultaneously. Of the 58 different medications in that study, 50 drugs were previously not detected in skin samples[11] and offer "negative" examples for which we have experimental data. Negative examples will include both the lack of transport to the epidermis, but also the lack of detection due to sample preparation (e.g., some drugs might not be detected due to the chosen extraction conditions). The eight drugs or drug metabolites that were detected in skin swabs in that study are part of presumed "positive" compounds that are observed on the epidermis. Further, these particular examples are supported with experimental data and matching prescription records (i.e., the drugs were detected in the subjects to whom they were prescribed).[11]

Data processing

All peak files filtered through ReDU were analyzed using MS/MS library searching on GNPS against all available public spectral libraries (version 2.0; GNPS task ID: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=53e265f8f6994f0196bf9bccd8d1b513). MS/MS library searching resulted in 175 drugs that were identified in the human skin files (level 2 annotation according to the Metabolomics Standards Initiative[17]), filtered using a list of curated drugs and drug metabolites as they are recorded in the GNPS MS/MS reference libraries. Duplicate annotations were removed and drugs available in topical formulations were excluded, resulting in a final list of 95 compounds. Based on the empirical measurement of these drugs or drug metabolites in publicly available MS data we presume that these 95 compounds are "positive" examples of drugs that appear on the epidermis.

Combined, 145 unique compounds were retained for the machine learning, 95 positive examples and 50 negative examples. The full list of compounds and information on whether they were observed on the epidermis or not is available in Table S3.

Machine learning epidermis prediction

Feature generation and preprocessing

A random forest classifier was used to predict whether drugs are expected to be observed on the epidermis. First,

Mordred,[18] a cheminformatics software tool to efficiently compute a large variety of molecular descriptors, was used to generate molecular descriptors for all 145 compounds, such as calculated measures of volume, electronegativity, bond energy, and electrotopology (see Table S4 for relevant descriptor examples). Molecular descriptors that were missing for one or more drugs were omitted, resulting in a feature table consisting of 929 unique descriptors per compound. Next, a classification pipeline was built to predict the probability of observing a drug on the epidermis. The classification pipeline consisted of preprocessing steps to remove irrelevant features and a random forest classifier. Preprocessing steps included removing all features whose variance was below 0.05 and removing one of the features for which their pairwise Pearson correlation exceeded 0.95.

Random forest classifier training and evaluation

A random forest classifier[19] using 1000 trees was trained to predict the epidermis probability. Evaluation and hyperparameter tuning of the classification pipeline were done using nested cross-validation. Two levels of stratified shuffle splitting consisting of 100 iterations of random splitting in 80% training data and 20% test data were performed. In the inner cross-validation loop, 10 iterations of randomized searching were used for hyperparameter optimization of the random forest. The following different hyperparameters were evaluated: tree depth (i.e., the longest path from the root node to the leaf nodes) between 5 and 9 (inclusive), minimum number of samples per leaf node between 1 and 9 (inclusive), and minimum number of samples to split an internal node between 2 and 9 (inclusive). The random forest classifier with optimal hyperparameters was subsequently evaluated in the outer cross-validation loop. The number of features retained in the final classification pipeline after removing uninformative features was 287, and trees with depth eight, minimum two samples per leaf node, and minimum two samples to split a node were most frequently found to be optimal. For each split, the balanced accuracy, true positive rate, false positive rate, and precision were computed for both the training data and test data. Model performance was assessed based on the receiver operating characteristic (ROC) curve and precision–recall curve.

SHapley Additive exPlanations model interpretability

Important features for epidermis prediction were determined using SHapley Additive exPlanations (SHAP),[20] a model interpretability method founded in game theory.

Briefly, SHAP explains machine-learning predictions by using interpretable local models to approximate a complex black box model. Kernel SHAP was used to explore the trained classification pipeline. To determine the important features, 50 training samples determined by K-means clustering, with the cluster centroids weighted by the number of samples assigned to them, were used as the background dataset. To investigate the features of importance of individual compounds, if they were part of the training dataset, the random forest classifier with optimal hyperparameters was retrained using a leave-one-out strategy prior to SHAP analysis.

FDA-approved drugs and biotransformations

Drug names, SMILES representations, and Anatomic Therapeutic Chemical (ATC) codes for 2561 FDA-approved drugs were retrieved from DrugBank (version 5.1.7)[21] on December 23, 2020. Mordred was used to generate the same features for these drugs as used during model training, and the probability of observing these drugs on the epidermis was determined using the trained classification pipeline. Additionally, potential biotransformation products of the drugs were generated using the BioTransformer tool.[22] The human super transformer mode, which combines an Enzyme Commission-based transformer, a CYP450 (phase I) transformer, a phase II transformer, and a human gut microbial transformer were used to predict potential biotransformation products after a single transformation step. This resulted in 23,693 putative biotransformation metabolites derived from the FDA-approved drugs, for which similarly the probability of observing them on the epidermis was predicted using the trained classification pipeline.

Code availability

All analyses were performed in Python 3.8. RDKit (version 2020.09.3)[23] and Mordred (version 1.2.0)[18] were used to generate molecular descriptors. A GPU-accelerated version of the random forest algorithm, available as part of the cuML library (version 0.18.0)[24] was used in combination with Scikit-Learn (version 0.24.1)[25] for data preprocessing and model evaluation. SHAP (version 0.39.0)[20] was used to compute the features of importance. BioTransformer (version 2.0.1)[22] was used to generate biotransformation products. Additionally, NumPy (version 1.20.1),[26] SciPy (version 1.6.0),[27] and Pandas (version 1.1.5)[28] were used for scientific computing, and matplotlib (version 3.3.4)[29] and Seaborn (version 0.11.1)[30] were used for visualization purposes.

All code is available at https://github.com/bittremieux/drugs_epidermis as open source under the permissive BSD license.

RESULTS

Occurrence of drugs on the epidermis

Based on the rich metadata associated with the MS/MS data, we were able to select 5629 MS/MS peak files that contain samples collected from human body sites from 20 publicly available datasets (Table S2). These data originate from a variety of studies, including, for example, a 3D molecular cartography of the human skin study in which paired MS and sequencing data was collected to investigate the spatial relationships of human skin with hygiene, the microbiota, and the environment (MSV000078556)[31]; a study in which skin samples were obtained from healthy human volunteers that were given single doses of caffeine, omeprazole, midazolam, and dextromethorphan on 2 separate days (8 days apart) and a 7-day course of cefprozil (MSV000082493)[32]; and the kidney transplant study described in more detail above (MSV000081548).[11]

For our secondary analysis, we extracted 145 curated drugs from these data to build a machine-learning model to predict whether drugs occur on the epidermis. Additionally, based on the Uberon anatomy ontology,[16] these drugs were mapped to the body site on which they were detected (Figure 1). The different rates of drug occurrence throughout the body suggest that there will be distinct detection of chemicals and xenobiotics in skin. As an example, our previous study showed that the N-acetyl metabolite of sulfamethoxazole was detected in armpit skin samples but not in other skin sites sampled, such as forehead, palms, and forearm.[11] More polar compounds may be more likely detected in more aqueous areas of the skin where sweat is more concentrated, such as the armpit.

Machine learning to predict whether drugs occur on the epidermis

Using a random forest classifier, we were able to predict whether drugs will be observed on the epidermis with an area under the ROC curve (AUC) obtained during cross-validation of 0.71 ± 0.10 (Figure 2) and an area under the precision–recall curve of 0.82 ± 0.07 (Figure S1). As an example, two drugs that were present in our data and which were previously reported to be present on the skin, the antimycotics fluconazole,[9] and ketoconazole,[33] are strongly predicted to be observed on the skin. This performance indicates that machine learning can be used to successfully
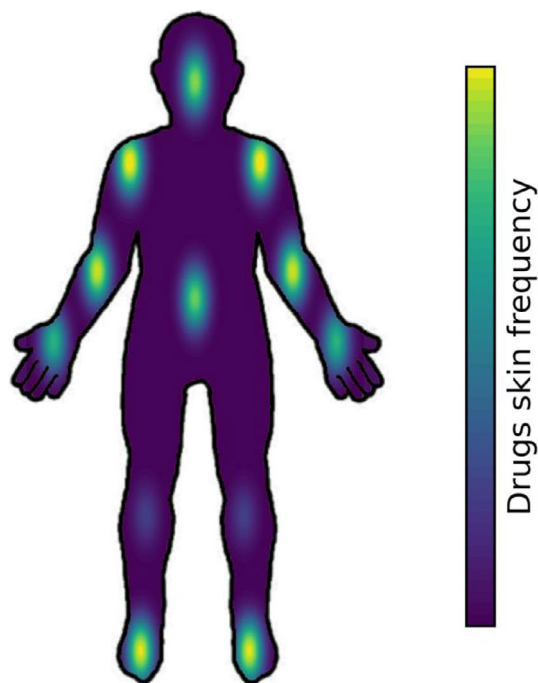

FIGURE 1 Body sites of the drugs found through spectral library searching. Body sites for the identified drugs were retrieved from the Uberon annotations specified in ReDU, and drug counts per body site were normalized by the total number of ReDU entries for each body site

approximate the complex underlying biochemical processes leading to drugs being observed on the epidermis. As we were constrained by the limited availability of ground truth data in this study, we hypothesize that as more training data becomes available it will be possible to produce even more accurate machine learning models (Figure S2).

We tried to gain insight into the molecules' physical properties that result in drugs being present on the epidermis. The SHAP model interpretation technique was used to determine the most relevant features, consisting of molecular descriptors generated by Mordred, for the classifier performance (Figure 3, Table S4). The top-ranked features are computed measures of volume (ATSC7v), electronegativity (PEOE_VSA1, PEOE_VSA9), bond energy (ATSC6d), and electrotopology (EState_VSA1). By investigating the SHAP values for individual features, we can derive that in general smaller compounds (Van der Waals volume) with a smaller bonding potential (electronegativity) are more likely to be observed on the epidermis. We can hypothesize that through heterogeneous biochemical processes such molecules diffuse faster and thus will be secreted to the epidermis.

Additionally, SHAP can be used to interpret predictions for individual drugs. The antihistamine drug diphenhydramine was experimentally observed on the epidermis in a previous healthy human clinical study (https://biorx
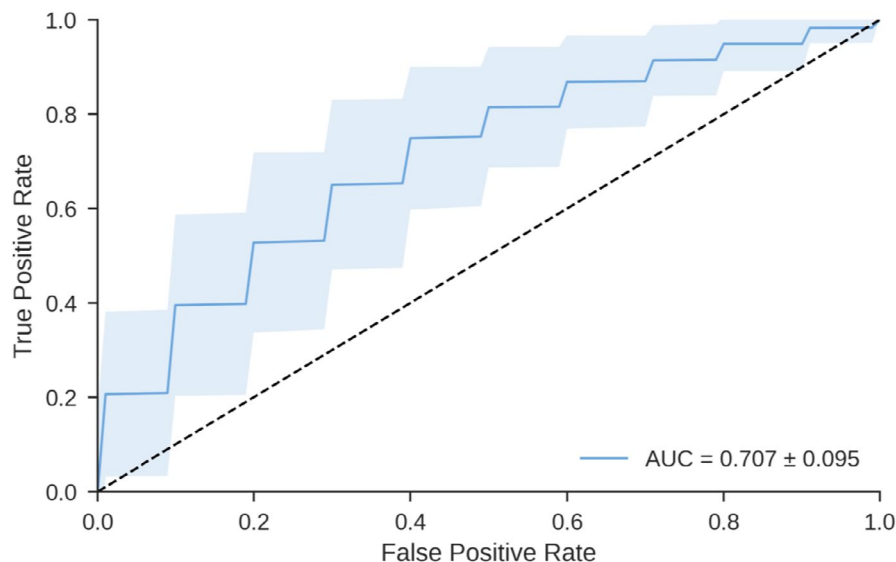
FIGURE 2 ROC curve indicating the performance of the random forest classifier to predict whether drugs can be observed on the epidermis. The curve is the mean ROC curve over 100 random stratified training (80% of the data) and test (20% of the data) splits. The standard deviation over the splits is indicated by the shaded area. The mean AUC is 0.707, with a standard deviation of 0.095. AUC, area under the curve; ROC, receiver operating characteristic
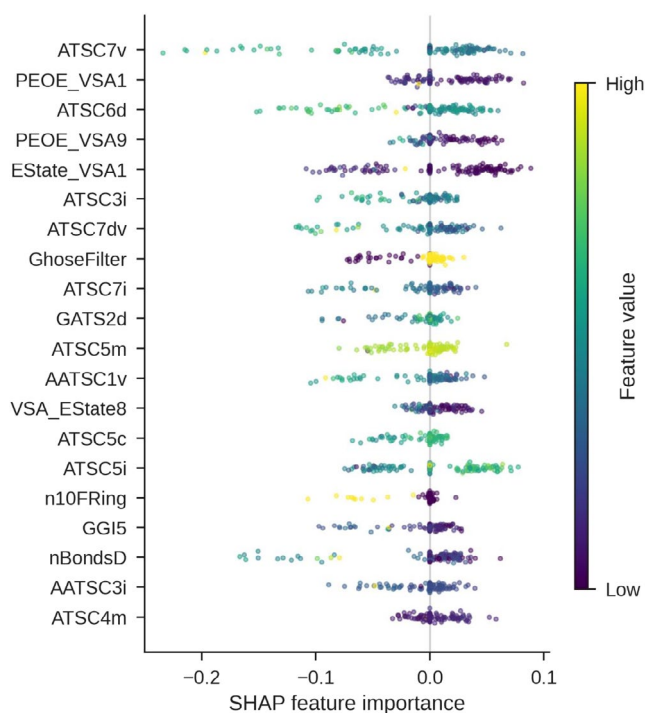
FIGURE 3 SHAP features of importance for the top 20 most important Mordred features from the random forest classifier for the 145 training compounds. A positive SHAP feature importance contributes to drugs predicted to appear on the epidermis, whereas a negative SHAP feature importance contributes to drugs predicted to not appear on the epidermis. The top-ranked features capture information about the volume, electronegativity, bond energy, and electrotopology of the molecules. See Table S4 for a full description of the Mordred features. SHAP, SHapley Additive exPlanations

iv.org/cgi/content/short/2021.11.22.469638v1). Using a leave-one-out training strategy to not bias the classifier, it was also strongly predicted to be present on the epidermis (Figure 4a). The most relevant features contributing to this prediction are its lack of accessible Van der Waals surface area with a low electrotopological state (EState_VSA1), a high atomic mass autocorrelation (ATSC4m), a small Van der Waals surface area with low partial charge (PEOE_VSA1), and a low autocorrelation weighted by sigma electrons (ATSC6d, ATSC7d). In contrast, although the related compound diphenhydramine N-hexose is structurally similar, it is predicted to not appear on the epidermis (Figure 4b), in part because of its increased accessible Van der Waals surface area with a low electrotopological state (EState_VSA1), as well as its higher autocorrelation weighted by Van der Waals volume (ATSC7v), its high autocorrelation weighted by ionization potential (ATSC7i), and its high topological radius (Radius). This is consistent with our experimental results (https://biorxiv.org/cgi/content/short/2021.11.22.469638v1). In a previous study,[11] citalopram was detected in the skin samples of the only subject to which it was prescribed. This empirical observation is confirmed by the machine-learning model (Figure 4c), as citalopram is very strongly predicted to be observed on the epidermis due to its lack of accessible Van der Waals surface area with a low electrotopological state (EState_VSA1), its low topological charge (GGI10), and its low autocorrelation weighted by valence electrons (ATSC7dv). Conversely, tacrolimus is very strongly predicted to not appear on the epidermis (Figure 4d), primarily due to its high number of double bonds (nBondsD), its high number of Kier–Hall dssC atom types (motif "C(=[*])([*])[*]"),[34] and its high autocorrelation weighted by ionization potential (ATSC5i, ATSC7i, and ATSC8i). This prediction matches its absence in the skin samples of 14 subjects who were prescribed tacrolimus.[11] This analysis demonstrates how machine-learning techniques can be used to obtain insights into the complex internal biochemical mechanisms that lead systemically administered drugs to be observed on the epidermis.

Exploring presence of drugs and related biotransformations on the epidermis

To expand our knowledge of the variety of drugs that are likely to be observed on the epidermis beyond the training data consisting of 145 drugs, we retrieved 2561 FDA-approved drugs from DrugBank.[21] Furthermore, we utilized BioTransformer[22] to predict potential biotransformation products of the FDA-approved drugs, resulting in 23,693 putative biotransformation metabolites. These biotransformations include phase I metabolism products (e.g., Cytochrome P450), enzyme commission-based metabolism products, phase II metabolism products (e.g., Uridine 5′-diphospho-glucuronosyltransferase), and gut microbial transformation products, and they cover a number of different reaction types, including hydrolysis, oxidation and reduction, and conjugation.

The probability of observing the FDA-approved drugs and their potential biotransformation products was predicted using the trained random forest model. To investigate whether specific types of drugs were more likely to occur on the epidermis, we grouped the drugs and the corresponding biotransformation products using the ATC Classification System (Figure 5). This indicates, for example, that hormonal preparations, such as corticosteroids, are least likely to be observed on the epidermis, whereas nervous system drugs, such as analgesics, antiepileptics, antidepressants, and antipsychotics are more likely to be detected on skin.

DISCUSSION

So far, little is known about which chemicals and drugs move from the systemic circulation to the epidermis. Here, we have demonstrated that machine learning can be used to gain insights into these complex processes for the first time. Using publicly available MS data, we have trained a random forest model to predict whether drugs will occur on the epidermis. Notably, the classifier correctly predicted the presence on the skin of the antimycotics fluconazole[9] and ketoconazole,[33] matching previous experimental results.

To obtain insights into the complex processes that underlie reverse penetration of drugs to the epidermis, the SHAP model interpretability method was used to investigate which molecular descriptors are most relevant for prediction using the random forest. In general, we observe that smaller compounds with a smaller bonding potential are more likely to be observed on the epidermis. Although further studies are needed to fully understand the underlying biochemical processes, we hypothesize that through heterogeneous mechanisms such molecules diffuse faster and thus will be secreted to the epidermis. Additionally, we used SHAP to investigate predictions for drugs with a known experimentally derived ground truth. This demonstrates how detailed and individualized insights for specific drugs can be obtained to explore whether they will appear on the epidermis or not.

Applying our random forest model to over 2500 FDA-approved drugs and their biotransformations gives insight into additional drugs and their metabolites that may be detected on the skin surface. For those drugs with a low probability of skin detection, we hypothesize that either these drugs are fully processed within the body, rather than secreted to the epidermis, or their physicochemical properties prevent access to the skin surface. For example, a high degree of lipophilicity might prevent access to the skin surface, as the hydrophilic viable epidermis is a barrier to lipophilic substances. Alternatively, highly hydrophilic compounds are unlikely to be detected on the skin either, as the hydrophobic stratum corneum is a barrier for hydrophilic substances. The variety of important physicochemical properties underlying the prediction of drug detection on the skin indicates that there is no single process for moving compounds from the systemic circulation to the epidermis, but rather that there are unique interplays between the specific drugs and their relevant transport pathways.

Notably, median epidermis prediction values for the FDA approved drugs and their biotransformations in different ATC drug classes range from ~35% to ~60% of drugs in each category. There is substantial variation in predicted probability; we speculate that this observation reflects that specific physicochemical properties of the drugs are the driver of this phenomenon rather than the ATC class. Nevertheless, broad generalizations can be made; for example, steroid hormones were predicted to not be detected on the epidermis, which is consistent with our experimental data for budesonide, fludrocortisone, prednisone, and prednisolone; whereas amitriptyline, citalopram, cyclobenzaprine, escitalopram, gabapentin, ketamine, nortriptyline, and venlafaxine were detected in our data, consistent with our model prediction for nervous system drugs (Table S4).

An important caveat of the current work is that there was only a limited amount of heterogeneous data available for secondary analysis. Because the data were derived from various publicly available untargeted metabolomics studies for secondary analysis, no consistent experimental protocol was used. Additionally, prescribed drugs might not be detected on the skin due to lack of transport from the systemic circulation to the epidermis, but also due to incompatibilities with sample preparation. Although we achieved encouraging predictive performance, the machine-learning performance can be further improved upon by obtaining and incorporating more

(a) Diphenhydramine

(b) Diphenhydramine N-hexose

(c) Citalopram

(d) Tacrolimus
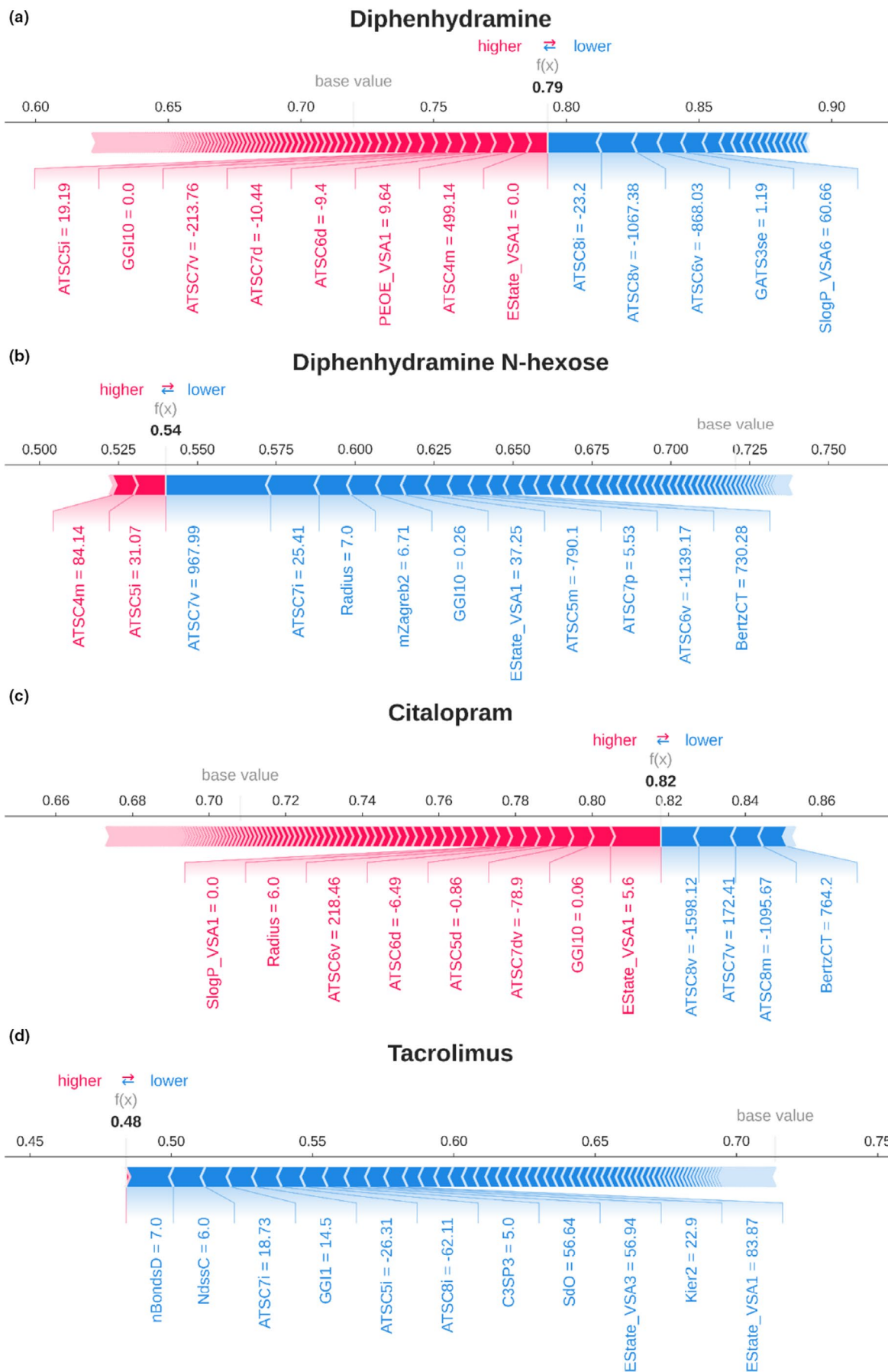
FIGURE 4 Force plots[35] of the SHAP values to interpret predictions of individual drugs. The most important features, their values, and the direction in which they contribute to the predictions (higher/red: observed, lower/blue: not observed) are displayed. The horizontal axis shows the model probability, with the prediction score indicated by "f(x)." Scores above the expected value based on the training data ("base value") constitute positive predictions, and scores below the expected value constitute negative predictions. The size of the bars for individual features indicates their magnitude contributing to a positive or negative prediction. (a) Diphenhydramine is predicted to be observed on the epidermis. (b) Diphenhydramine N-hexose is predicted to not be observed on the epidermis. (c) Citalopram is predicted to be observed on the epidermis. (d) Tacrolimus is predicted to not be observed on the epidermis. SHAP, SHapley Additive exPlanations
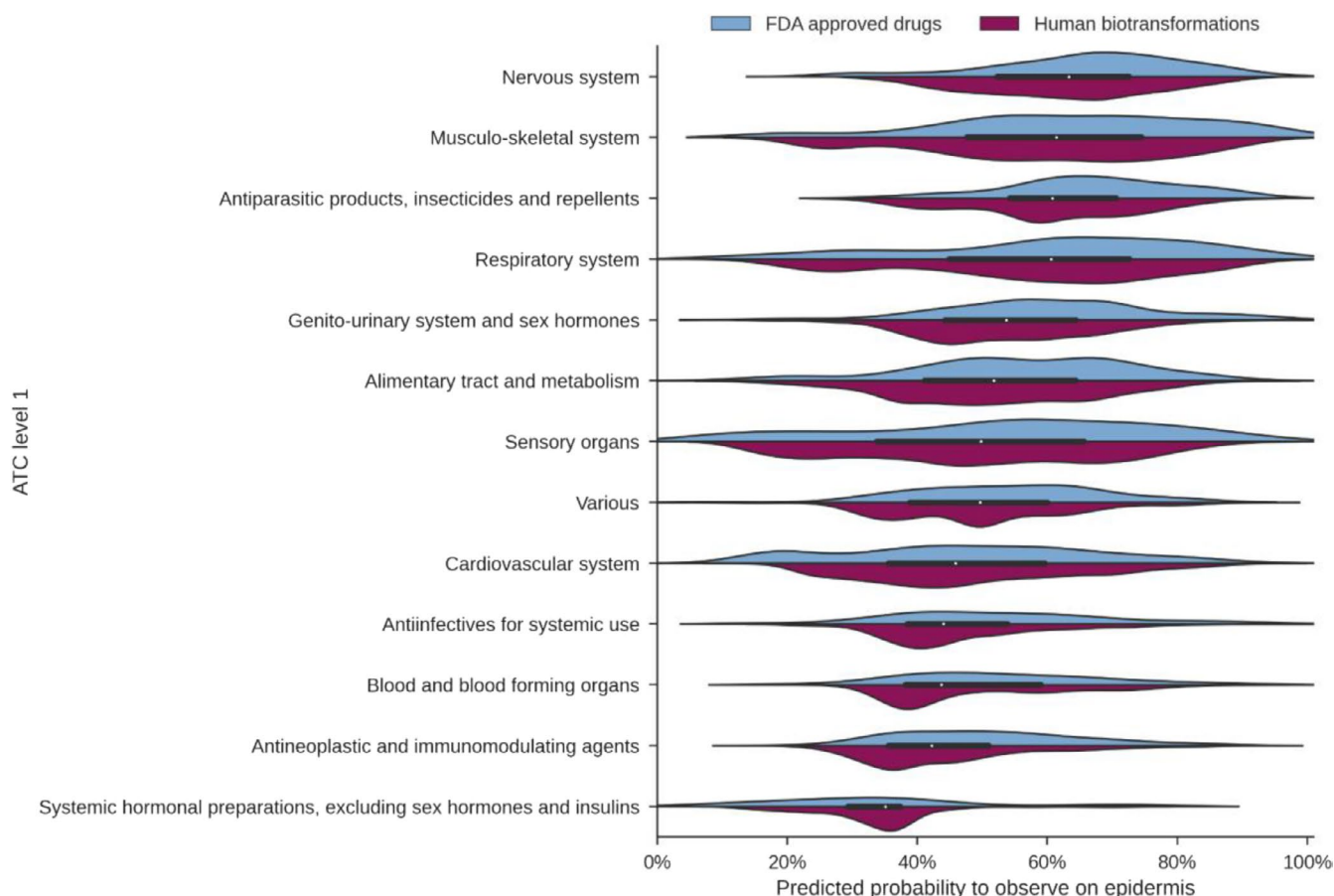
FIGURE 5 Prediction scores for 2561 FDA approved drugs and their 23,693 biotransformations, subdivided by their drug class in the ATC classification system. ATC, Anatomic Therapeutic Chemical; FDA, US Food and Drug Administration

relevant experimental data, as indicated by the learning curve. Ideally, this should include a controlled measurement of both positive and negative examples of drugs and other xenobiotics that are commonly consumed and their status of being observed on the epidermis.

Our machine-learning model is the first attempt to predict xenobiotic skin detection using physicochemical properties. There will likely be future iterations of this model as we advance our understanding of the complex processes governing molecular transport from the systemic circulation to the surface of the skin. The use of noninvasive skin swabs in clinical medicine could be a paradigm shift in how health and disease are monitored. Contemporary methods of blood draws and tissue biopsies are invasive and inconvenient for patients. In the future, we envision the use of noninvasive skin sampling to determine adherence of drugs, for therapeutic drug monitoring, the extent of metabolism, and to assess organ and health status.

CONFLICT OF INTEREST

P.C.D. is on the scientific advisory board of Sirenas, Cybele Microbiome, Galileo, and founder and scientific advisor of Ometa Labs LLC and Enveda. All other authors declare no competing interests for this work.

AUTHOR CONTRIBUTIONS

W.B., R.S.A., A.K.J., P.C.D., and S.M.T. wrote the manuscript. R.S.A., A.K.J., and S.M.T. designed the research. W.B., R.S.A., S.A., A.L., A.K.J., and S.M.T. performed the

research. W.B., R.S.A., A.K.J., and S.M.T. analyzed the data. W.B. and R.S.A. contributed new reagents/analytical tools.

ORCID

Wout Bittremieux https://orcid.org/0000-0002-3105-1359
Shirley M. Tsunoda https://orcid.org/0000-0002-3974-8038

REFERENCES

1. Belkaid Y, Segre JA. Dialogue between skin microbiota and immunity. *Science*. 2014;346:954-959.
2. Matta MK, Florian J, Zusterzeel R, et al. Effect of sunscreen application on plasma concentration of sunscreen active ingredients: a randomized clinical trial. *JAMA*. 2020;323:256-267.
3. Hadgraft J, Lane ME. Skin permeation: the years of enlightenment. *Int J Pharm*. 2005;305:2-12.
4. Yamashita F, Hashida M. Mechanistic and empirical modeling of skin permeation of drugs. *Adv Drug Deliv Rev*. 2003;55:1185-1199.
5. Kalia YN, Guy RH. Modeling transdermal drug release. *Adv Drug Deliv Rev*. 2001;48:159-172.
6. Stepanov D, Canipa S, Wolber G. HuskinDB, a database for skin permeation of xenobiotics. *Sci Data*. 2020;7:426.
7. Patzelt A, Antoniou C, Sterry W, Lademann J. Skin penetration from the inside to the outside: a review. *Drug Discov Today Dis Mech*. 2008;5:e229-e235.
8. Felton T, Troke PF, Hope WW. Tissue penetration of antifungal agents. *Clin Microbiol Rev*. 2014;27:68-88.
9. Wildfeuer A, Faergemann J, Laufen H, et al. Bioavailability of fluconazole in the skin after oral medication. *Mycoses*. 1994;37:127-130.
10. Bouslimani A, Melnik AV, Xu Z, et al. Lifestyle chemistries from phones for individual profiling. *Proc Natl Acad Sci USA*. 2016;113:E7645-E7654.
11. Jarmusch AK, Elijah EO, Vargas F, et al. Initial development toward non-invasive drug monitoring via untargeted mass spectrometric analysis of human skin. *Anal Chem*. 2019;91:8062-8069.
12. Wang M, Carver JJ, Phelan VV, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol*. 2016;34:828-837.
13. Aron AT, Gentry EC, McPhail KL, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat Protoc*. 2020;15:1954-1991.
14. Wang M, Jarmusch AK, Vargas F, et al. Mass spectrometry searches using MASST. *Nat Biotechnol*. 2020;38:23-26.
15. Jarmusch AK, Wang M, Aceves CM, et al. ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat Methods*. 2020;17:901-904.
16. Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. *Genome Biol*. 2012;13:R5.
17. Sumner LW, Amberg A, Barrett D, et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*. 2007;3:211-221.
18. Moriwaki H, Tian Y-S, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. *J Cheminform*. 2018;10:4.
19. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
20. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. arXiv [cs.AI], 2017. Available from http://arxiv.org/abs/1705.07874.
21. Wishart DS. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*. 2006;34:D668-D672.
22. Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, et al. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminform*. 2019;11:2.
23. Landrum G, et al. rdkit/rdkit: 2020_09_3 (Q3 2020) Release, 2020. 10.5281/zenodo.4318717.
24. RAPIDS Citations. Available from https://rapids.ai/citations.html.
25. Pedregosa F, et al. Scikit-learn: machine learning in Python. *J Machine Learn Res*. 2011;12:2825-2830.
26. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585:357-362.
27. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17:261-272.
28. McKinney W. & Others Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*. 445, 51-56 (Austin, TX, 2010)
29. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90-95.
30. Waskom M, et al. mwaskom/seaborn: v0.11.1 (December 2020). (Zenodo, 2020). 10.5281/ZENODO.592845.
31. Bouslimani A, et al. Molecular cartography of the human skin surface in 3D. *Proc Natl Acad Sci USA*. 2015;112:E2120-E2129.
32. Jarmusch AK, Vrbanac A, Momper JD, et al. Enhanced characterization of drug metabolism and the influence of the intestinal microbiome: a pharmacokinetic, microbiome, and untargeted metabolomics study. *Clin Transl Sci*. 2020;13:972-984.
33. Harris R, Jones HE, Artis WM. Orally administered ketoconazole: route of delivery to the human stratum corneum. *Antimicrob Agents Chemother*. 1983;24:876-882.
34. Pearlman RS. Molecular structure description. The Electrotopological State By Lemont B. Kier (Virginia Commonwealth University) and Lowell H. Hall (Eastern Nazarene College). Academic Press: San Diego. 1999. xx +245 pp. $99.95. ISBN 0-12-406555-4. *J Am Chem Soc* 2000;122:6340.
35. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. 2018;2:749-760.

SUPPORTING INFORMATION
Additional supporting information may be found in the online version of the article at the publisher's website.