Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Development of machine learning models for the prediction of the skin sensitization potential of small organic compounds

Cumulative Dissertation
with the aim of achieving a doctoral degree
Doctor rerum naturalium (*Dr. rer. nat.*)

at the Faculty of Mathematics, Informatics and Natural Sciences
Department of Chemistry
Universität Hamburg

submitted by
Anke Wilm
Master of Science, Universität Hamburg
born in Frankfurt a.M.

April 03, 2022

The presented thesis was prepared from August 2017 till March 2022 under the supervision of Assoc.-Prof. Dr. Johannes Kirchmair at the Center of Bioinformatics, Universität Hamburg.

Reviewers:
Assoc.-Prof. Dr. Johannes Kirchmair
Prof. Dr. Ralph Holl

Date of submission:
April 03, 2022
Date of examination:
June 24, 2022

I. Acknowledgments

This work was funded by Beiersdorf AG, Hamburg, and implemented at the Universität Hamburg, Center for Bioinformatics (ZBH), in cooperation with HITeC, the Research and Technology Transfer Center of the Department of Informatics.

I would like to thank my supervisor, Assoc.-Prof. Dr. Johannes Kirchmair for encouraging me with so much advice, knowledge, trust, and vigor. I am thankful I had the opportunity to develop and learn so much during the last few years working together.

I would also like to express my gratitude to Beiersdorf AG for this wonderful opportunity to participate in a synergistic collaboration between academia and industry. I am especially grateful to Dr. Jochen Kühnl for enabling this project and his continues, strong support. Furthermore, I am thankful to Dr. Jochen Kühnl, Dr. Horst Wenck, Dr. Andreas Schepky, and the experimental toxicology group at Beiersdorf AG for fruitful discussions that enriched my research and fostered my personal development. Many thanks go also to Dr. Lothar Hotz, Wiebke Frauen, and Silke Peters from HITeC for their excellent project support.

A big thank you goes to Prof. Dr. Ulf Norinder, who supported me in developing and interpreting CP models. Many thanks also to all the people who contributed to my papers as co-authors. In addition and in particular, I would like to thank the entire former and present COMP3D team, at the Universität Hamburg, University of Bergen and University of Vienna. I especially want to thank Anya, Christina, Christoph, Conrad, Isabel, Marina, Neann, Nils, Steffen, and Wojciech for inspiring scientific discussions and pleasant private breaks. Conrad implemented the infrastructure for our web server and our git repository. I learned so much when working together with you. Anya, Conrad, Marina and Neann helped me by proofreading parts of this thesis. My friend Laura Thompson did the final proofreading of the whole manuscript. Many thanks also go to our admins Gerd and Jörn, who provided us with the perfect environment for high class computing and also helped when individual requirements arose. Many thanks to you all!

Last, but not least, I want to thank my family, friends, and neighbors for all the support during the last few years. I would not have and could not have finished this thesis without your support.

II. List of publications originating from this work

P1 **Wilm, A.**, Kühnl, J., Kirchmair, J., Computational approaches for skin sensitization prediction, *Critical Reviews in Toxicology*, **48**(9) (2018) 738–760.

P2 **Wilm, A.**, Stork, C., Bauer, C., Schepky, A., Kühnl, J., Kirchmair, J., Skin doctor: Machine learning models for skin sensitization prediction that provide estimates and indicators of prediction reliability, *International Journal of Molecular Sciences*, **20**(19) (2019) 4833.

P3 **Wilm, A.**, Norinder, U., Agea, M. I., de Bruyn Kops, C., Stork, C., Kühnl, J., Kirchmair, J., Skin doctor CP: Conformal prediction of the skin sensitization potential of small organic molecules, *Chemical Research in Toxicology*, **34**(2) (2020) 330–344.

P4 **Wilm, A.**, Garcia de Lomana, M., Stork, C., Mathai, N., Hirte, S., Norinder, U., Kühnl, J., Kirchmair, J., Predicting the skin sensitization potential of small molecules with machine learning models trained on biologically meaningful descriptors, *Pharmaceuticals*, **14**(8) (2021) 790.

III. Contents

IV. List of abbreviations

ACC accuracy

ACD allergic contact dermatitis

AD applicability domain

AI artificial intelligence

AOP adverse outcome pathway

CCR correct classification rate

CDK chemistry development kit

CP conformal prediction

CV cross-validation

DA defined approach

DPRA direct peptide reactivity assay

ECFP extended-connectivity fingerprints

FN false negative

FP false positive

GARD genomic allergen rapid detection

GPMT guinea pig maximization test

h-CLAT human cell line activation test

hTCPA human T-cell priming assay

IATA integrated approach to testing and assessment

ICD irritative contact dermatitis

InChI international chemical identifier

KNN k-nearest neighbor

LASSO least absolute shrinkage and selection operator

LLNA local lymph node assay

MACCS molecular access system

MCC Matthews correlation coefficient

MEST mouse ear-swelling test

MIE molecular initiating event

ML machine learning

MOE molecular operating environment

NERDD new e-resource for drug discovery

NPV negative predictive value

OCD occupational contact dermatitis

OECD organisation for economic co-operation and development

PaDEL pharmarceutical data exploration laboratory

PC principal component

PCA principal component analysis

PPV positive predictive value

(Q)SAR (quantitative) structure activity relationship

RAI relative alkylation index

RF random forest

SD structure data

SMILES simplified molecular-input line-entry system

SVM support vector machine

TN true negative

TP true postive

UMAP uniform manifold approximation and projection

U-Sens Myeloid U937 skin sensitisation test

1. Zusammenfassung

Das allergische Kontaktekzem (ACD) ist eine unangenehme und weit verbreitete Erkrankung, die durch wiederholten Hautkontakt (z.B. über Konsumgüter oder Arbeitsmittel) mit einer hautsensibilisierenden Substanz ausgelöst werden kann [5–7]. Um die Entstehung von ACD zu verhindern, ist eine sorgfältige Risikobewertung des Hautsensibilisierungspotenzials neu entwickelter Chemikalien und Substanzen erforderlich. In der Vergangenheit beruhte diese vor allem auf Ergebnissen von Tierversuchen [8]. Inzwischen ist erwünscht (und teilweise auch gesetzlich vorgeschrieben [9–13]), das Hautsensibilisierungspotenzial neuer Substanzen mit tierversuchsfreien Alternativen wie in-vitro und in-chemico Tests sowie computergestützten Methoden einzuschätzen [14, 15]. Im Vergleich zu experimentbasierten Testansätzen bieten hierbei computergestützte Methoden mehrere Vorteile, darunter eine kürzere Testdauer und geringere Kosten. Sie sind daher eine vielversprechende Säule für eine tierversuchsfreie Risikobewertung des Hautsensibilisierungspotenzials kleiner Moleküle.

Ziel dieser Arbeit ist die Entwicklung und Bewertung zuverlässiger und anwendbarer computergestützter Methoden für die Vorhersage des Hautsensibilisierungspotenzials kleiner organischer Substanzen. Besonderes Augenmerk wird auf Aspekte gelegt, die die Nutzbarkeit und Akzeptanz der entwickelten Modelle für die Risikobewertung erhöhen. Dies beinhaltet insbesondere eine solide Datenbasis für die Modellentwicklung und -bewertung, solide Größen zur Bestimmung der Zuverlässigkeit der Vorhersagen und eine verbesserte Interpretierbarkeit der Modelle.

In einem ausführlichen Übersichtsartikel [P1] haben wir zunächst die öffentlich zugänglichen Daten zur Hautsensibilisierung sowie die vorhandenen Vorhersageprogramme und -ansätze zusammengefasst und diskutiert. Wir haben eine Vielzahl unterschiedlicher Ansätze mit teilweise zueinander komplementären Vor- und Nachteilen identifiziert, die eine vielversprechende Grundlage für die computergestützte Vorhersage von Hautsensibilisierungspotenzial bilden. In der Praxis wird jedoch keines der untersuchten Modelle als alleinstehender Ansatz für die Risikobewertung akzeptiert.

In unserem ersten Projekt haben wir den größten öffentlich zugänglichen Local-Lymph-Node-Assay (LLNA)-Datensatz (1416 Substanzen) zusammengestellt und seine Relevanz für den chemischen Raum von Kosmetika, Arzneimitteln und

Pestiziden nachgewiesen. Auf der Grundlage dieses Datensatzes haben wir, basierend auf maschinellen Lernverfahren (ML), 58 Modelle für die Vorhersage des binären Hautsensibilisierungspotenzials kleiner organischer Moleküle entwickelt, optimiert und bewertet. Mit wenigen Ausnahmen erreichten die optimierten Modelle eine vergleichbare Vorhersagekraft mit einer Genauigkeit (ACC) von bis zu 0,76 und einem Matthews-Korrelationskoeffizienten (MCC) von bis zu 0,55. Durch die Implementierung eines definierten Anwendungsbereiches (AD) und zweier Zuverlässigkeitsgrößen konnten wir die Vorhersagekraft der Modelle für eine Teilmenge unserer Daten bei gleichzeitiger Verringerung der Efficiency/Coverage erhöhen. Als Ergebnis unserer Analyse wurden zwei der leistungsfähigsten Modelle in der Skin Doctor Suite implementiert, einem über einen Webserver öffentlich zugänglichen Programm für die Vorhersage des Hautsensibilisierungspotenzials. Neben der eigentlichen Vorhersage gibt die Skin Doctor Suite auch Informationen über den AD und die beiden Zuverlässigkeitsgrößen einer jeden Substanz aus.

Im Anschluss an die Entwicklung der Skin Doctor Suite haben wir unseren LLNA-Datensatz weiter verbessert, indem wir alle Moleküle zusätzlich einer manuellen Qualitätsprüfung unterzogen. Damit ein ML-Modell für Risikobewertung und Zulassung eingesetzt werden kann, ist eine definierte und mathematisch belegte Zuverlässigkeit für jede einzelne Vorhersage vorteilhaft. Daher haben wir eine unserer leistungsstärksten Kombinationen aus Modellierungsalgorithmus, Hyperparametern und Deskriptoren aus der Skin Doctor Suite verbessert, indem wir das Modell in ein aggregiertes Mondrian conformal prediction (CP)-Framework eingebettet haben. Dies ermöglicht die mathematisch robuste Berechnung der Zuverlässigkeit jeder einzelnen Vorhersage und umgeht die Notwendigkeit zusätzlicher Grenzwerte für die Definition des AD oder der Zuverlässigkeitsgrößen. Wenn beispielsweise eine Error Significance von 0,20 zugelassen wird, können ACC, MCC und Coverage/Efficiency von 0,78, 0,56 bzw. 0,82 erreicht werden. Dieses binäre Modell, Skin Doctor CP, wurde ebenfalls auf unserem Webserver veröffentlicht. Um zwei verschiedene Klassen von Sensibilisierern (schwache bis mittlere und starke bis extreme Sensibilisierer) weiter zu unterscheiden, wurde ein zusätzliches aggregiertes Mondrian CP Modell trainiert und mit dem ursprünglichen Skin Doctor CP-Workflow zu einem ternären Modell kombiniert. Die ternäre Klassifizierung schien bei der Analyse der globalen Leistung erfolgreich zu sein, zeigte jedoch deutliche Schwächen bei der lokalen Leistung der unterrepräsentierten Klasse der starken bis extremen Sensibilisierern. Diese und weitere Schwächen konnten auch in einem von Di et al. [16] veröffentlichten ternären Modell festgestellt werden, indem dessen lokale Leistungen näherungsweise berechnet wurden.

Für die Akzeptanz eines Modells in der Risikobewertung ist eine gute Interpretierbarkeit von Vorteil. Darüber hinaus kann ein gut interpretierbares Modell Einblicke in dessen Mechanismus sowie in den zugrundeliegenden biologischen

Hintergrund geben. Um die Interpretierbarkeit unserer Modelle zu erhöhen, haben wir in einer dritten Studie einen strengen Selektionsprozess auf einen Satz von 750 berechneten Bioaktivitätsdeskriptoren angewendet. Auf diese Weise haben wir zehn biologisch sinnvolle Deskriptoren identifiziert, die als einzige Deskriptoren eines aggregierten Mondrian CP-Workflows Verwendung fanden. Das endgültige Modell erreichte eine Vorhersagekraft, die mit der unserer früheren, weniger interpretierbaren Modelle vergleichbar ist (ACC von 0,76, MCC von 0,53 und Coverage/Efficiency von 0,82, bei einer Error Significance von 0,20). Des Weiteren wurden der LLNA-Datensatz und drei Referenzdatensätze (bestehend aus Kosmetika, Pharmazeutika und Pestiziden) in dem von den diesen Deskriptoren aufgespannten chemischen Raum analysiert. Hierbei zeigte sich eine hohe Eignung der ausgewählten Deskriptoren zur Beschreibung des Sensibilisierungspotentials der untersuchten Moleküle.

2. Abstract

Allergic contact dermatitis (ACD) is a common and distressful condition among workers and consumers which is induced by the repeated contact of the skin to a skin sensitizing substance [5–7]. To prevent the induction of ACD, a careful risk assessment according the skin sensitization potential or potency of newly developed chemicals and substances is required. Historically, skin sensitization risk assessment was mainly conducted by animal experiments [8]. Currently, it is desired (and partly legally required [9–13]) to assess skin sensitization potential with non-animal alternatives such as in vitro and in chemico assays and computational methods [14, 15]. Compared to testing approaches, computational methods tout several advantages, including reduced testing time, and lower costs. Thus, computational methods are a promising pillar for a non-animal risk assessment of the skin sensitization potential and potency of small molecules.

In this thesis, we aim to support the development of reliable and applicable computational tools for the prediction of skin sensitization potential and potency of small molecules. Special emphasis is placed on aspects to increase the models' usability and acceptance for risk assessment by providing a solid data basis for model development and evaluation, solid measures of reliability and increased interpretability linked to the biological processes of the induction of skin sensitization. We summarized and critically reviewed the publicly available data for skin sensitization as well as the existing computational tools and approaches in an extensive review article [P1]. We identified a variety of different approaches with partially complementary advantages and disadvantages, comprising a promising base for the computational prediction of skin sensitization potential and potency. However, none of the models reviewed are presently accepted as a standalone approach for risk assessment.

In our first project, we compiled the largest publicly available local lymph node assay (LLNA) data set (1416 compounds) and demonstrated its relevance for the chemical space covered by cosmetics, pharmaceuticals, and pesticides. Based on this data set, we developed, optimized and evaluated 58 machine learning (ML) models for the prediction of binary skin sensitization potential of small molecules. With few exceptions, the optimized models reached comparable performance and did not exceed an accuracy (ACC) of 0.76, and Matthews correlation coefficient (MCC) of 0.55. By implementing an applicability domain (AD) and two measures of reliability we could increase models'

predictivity for a subset of our data accompanied by a simultaneous decrease of models' coverage. As a result of our analysis, two of the best performing models have been implemented as the Skin Doctor Suite, a publicly available web server for the prediction of skin sensitization potential. The Skin Doctor Suite also comprises AD information on every single molecule of interest as well as the two reliability measures. All three of them are visualized with a simple and intuitive color coding.

Following the development of the Skin Doctor Suite, we further refined our LLNA data set with an additional manual data curation step. For a ML model to be applicable for risk assessment and regulator purposes, a defined and mathematically proven reliability for every single prediction is advantageous. Thus, we enhanced one of our best performing combinations of modeling algorithm, hyperparameters, and set of descriptors from the Skin Doctor Suite by enveloping the final model into an aggregated Mondrian conformal prediction (CP) framework. This allows for the enumeration of the reliability of every single prediction in a mathematically proven way and circumvents the need of any additional cutoff values for the definition of the AD or the reliability measures. For example, when allowing an error significance of 0.20, ACCs, MCCs, and coverage/efficiency of 0.78, 0.56, and 0.82 could be realized, respectively. This final binary classifier, Skin Doctor CP, was published on our web server. To further differentiate two different classes of sensitizers (weak to moderate and strong to extreme sensitizers), an additional classifier was trained and combined with the original Skin Doctor CP workflow into a ternary classifier. The ternary classification was successful when analyzed by the global performance, but revealed non-negligible weaknesses in the local performance of the underrepresented class of strong to extreme sensitizers. Even more weaknesses could be detected in a ternary model published by Di et al. [16] when estimating the local performances of this model.

For a model to be accepted for risk assessment, a high interpretability is preferable. Additionally, an interpretable model can contribute insights into the mechanism of the model as well as the underlying biological background. To promote the interpretability of our models, we applied a strict feature selection process to a set of 750 calculated bioactivity descriptors in our third study. By doing this, we identified ten biologically meaningful descriptors which are capable of serving as the only descriptors of an aggregated Mondrian CP workflow. The final model resulted in a performance comparable to the one of our former, less interpretable models (ACC of 0.76, MCC of 0.53, and coverage/efficiency of 0.82, at an error significance of 0.20). An analysis of the LLNA data set and three reference data sets (comprising molecules labeled as cosmetics, pharmaceuticals, and pesticides) in the chemical space spanned by these descriptors demonstrated high discriminative capacities of the descriptor set selected.

3. Introduction

Repeated exposure of the skin to a sensitizing substance can induce allergic contact dermatitis (ACD). ACD manifests as often uncomfortable rashes or skin lesions at the site of exposure and can result in itching, burning, and pain [17]. In a large meta study conducted in 2007, about 20% of the general population of North America and Western Europe were found to be allergic to at least one contact allergen [5]. Depending on the study and the examined profession, ACD is responsible for approximately 20% to 50% of reported cases of occupational contact dermatitis (OCD) which is estimated to cause up to 30% of all reported occupational diseases (0.5–1.9 cases of OCD per 1000 full-time workers) [18]. For those affected, ACD can be a distressing problem resulting in sick leaves and health expenses [6, 7]. To prevent the induction of this condition, skin sensitization is an important endpoint for the safety assessment of new chemicals and consumer products.

3.1 Background

The development of skin sensitization can be divided into two phases [19]: First, during the induction phase, a cutaneous immune response is triggered by the contact of the skin with the sensitizing substance and results in an immunological priming. Secondly, during the elicitation phase, a new contact of the skin with this substance results in a symptomatic immune response associated with an inflammatory reaction at the site of the contact. The mechanisms behind the induction of skin sensitization are described in the currently accepted skin sensitization adverse outcome pathway (AOP) [20]. The pathway consists of eleven steps that comprise four key events. The first key event or molecular initiating event (MIE) is called haptenization and describes the molecular interaction of the substance with skin peptides and proteins. Molecules that need activation through autoxidation or enzymatic reactions prior to protein binding are called pre- and pro-haptens, respectively [21]. Within the second and third key event, activation of kertinocytes and dendritic cells are described, respectively. Finally, the fourth key event deals with the proliferation of hapten-specific T cells. The AOP does not only provide mechanistic insight in the processes behind skin sensitization induction, but also helps to structure existing knowledge about individual compounds and mixtures [22, 23].

Traditionally, the skin sensitization potential of a substance is assessed by animal or (very rarely) human in vivo studies. Since the early 1940s, animal experiments on skin sensitization potential have been conducted on guinea pigs [8]. Defined experiments on this species, such as the Buehler Test and the guinea pig maximization test (GPMT), have been available since the 1960s. After the induction phase (occluded exposure in the case of the Buehler Test and a combination of occluded exposure and intradermal injections in the case of GPMT) both protocols assess skin sensitization potential by the visible investigation of a patch test conducted on the flanks of the animals. Alternative experiments conducted on mice, such as the mouse ear-swelling test (MEST), came into focus several decades later. The MEST combines epidermal exposure with intradermal injections within the induction phase. In the elicitation phase, one of the mouse's ears is exposed to the test substance and the swelling of the exposed ear compared to the control is measured in the assays readout. All three assays only provide an uncertain quantification of skin sensitization potential (skin sensitization potency) and are thus better suited for binary classification of skin sensitization potential.

Currently, the local lymph node assay (LLNA) conducted in mice has become the method of choice for in vivo assessment of skin sensitization potential and potency. After repeated contact of the mice with the substance of interest, animals are sacrificed and the cell proliferation in draining lymph nodes is measured [24]. The concentration percentage of test compound that produces a 3-fold increase in cell proliferation is called the EC3 value. Compared to the qualitative readouts from other animal experiments, the EC3 value presents the advantage of providing a relatively stable and reproducible measure of skin sensitization potency [25]. The quantitative assessment of potency compared to a qualitative potential is desired since it reduces the need to completely exclude skin sensitizing substances and could allow for their application in safe concentrations [26]. The LLNA is also considered advantageous with respect to animal welfare since the objective of the experiment is the induction phase only [8].

Very rarely, measured data on skin sensitization potential from human clinical trials are available [27, 28]. Since they are usually conducted on less controlled studies, they are considered less reliable than the corresponding animal experiments [14]. Nevertheless, a comparison of human data and LLNA has been conducted several times: The LLNA is reported to predict binary skin sensitization potential in humans with a balanced accuracy between 0.58 and 0.88 depending on the study and the data analyzed [29]. Based on a carefully curated data set provided by Cosmetics Europe [30] (in the original version comprising 128 substances), a balanced accuracy of 0.68 can be expected [29, 30].

While the LLNA is considered the gold standard for skin sensitization prediction in terms of reliability, it is desired to replace in vivo experiments with non-animal alternatives due to ethical reasons [14, 15]. This shift is also promoted by regulatory authorities, for example in the EU (by e.g. EU Directive 2010/63/EU, REACH and the Cosmetic Products Regulation [9–11]) and the US (by e.g. Tox21 and ToxCast program [12, 13]) direct towards non-animal alternatives. To address the single key events from the skin sensitization AOP without animal testing, several experimental in vitro and in chemico methods are currently at hand. [8, 31, 32] Five of them are fully validated and covered by an organisation for economic co-operation and development (OECD) guideline and two more are under review: The first key event can be addressed by the direct peptide reactivity assay (DPRA) [33], while the second key event can be addressed by KeratinoSens, LuSens, and SENS-IS [34]. The third key event of the AOP can be addressed by IL8-Luc, Myeloid U937 skin sensitisation test (U-Sens), human cell line activation test (h-CLAT) and genomic allergen rapid detection (GARD) assay [35]. The fourth key event could in theory be addressed by the human T-cell priming assay (hTCPA) [36]. In practice, this key event is usually not assessed by non-animal alternatives due to experimental difficulties and a lack of verification data. Multiple key events of the AOP can be covered by a combination of non-animal tests within so called defined approaches (DAs) or integrated approaches to testing and assessment (IATAs) [37–39]. This are promising routes to increase the reliability and coverage of non-animal approaches.

In 2018, when our review article [P1] was published, the largest publicly available skin sensitization data set was collected and published by Alves et al. [40] and comprised 1000 LLNA, 138 human, 194 DPRA, 190 KeratinoSens, 160 h-CLAT data points. At the same time, the best curated (but thus smaller) data set was published by Cosmetics Europe [30]. This data set comprises an almost complete matrix of LLNA, human, DPRA, KeratinoSens, h-CLAT, U-Sens, SENS-IS data for 128 well-curated substances. Due to its relatively small size, the data set is better suited for model evaluation than for model building. Both data sets are further described and discussed in our review article [P1]. A more recent LLNA data set published by Di et al. [16] is introduced in chapter 3.5 of this thesis.

An advantage in terms of testing time and economic costs compared to any testing method can be gained by the application of computational models [41]. A variety of different approaches to model skin sensitization potential and potency in silico have been developed and refined in recent years. They can be divided into three approaches: i) expert knowledge or rule-based approaches, which encode expert knowledge on potential reactivity and reaction pathways of a molecule of interest, ii) similarity based approaches, which are based on the assumption that similar molecules will have similar biological properties, and

iii) (quantitative) structure activity relationship ((Q)SAR) approaches, which develop and apply mathematical functions to calculate the bioactivity of compounds based on their molecular features. Depending on the mathematical function describing the correlation between descriptors and skin sensitization potential or potency, two types of (Q)SAR approaches can be further differentiated into linear and non-linear models. While rule-based, similarity based and linear (Q)SAR approaches can intuitively be interpreted by human investigators, non-linear machine learning (ML) based (Q)SAR models often are more akin to a "black box". Nevertheless, they have proven to be a powerful tool for the prediction of skin sensitization potential and potency, since they are able to capture complex correlations between molecular features and toxic activity. Moreover, several computational prediction tools may be combined in hybrid models or be integrated in IATA or DAs together with non-animal assays to improve the predictivity and applicability of stand-alone approaches. However, a strong dependency of the models' performance and applicability on the quality and quantity of the underlying data can be observed for all theoretical approaches predicting skin sensitization potential and potency [42, 43].

Computational approaches to predict skin sensitization potential and potency (including stand-alone models, hybrid models and models incorporating experimental data) published before 2018 are extensively reviewed and discussed within our review article [P1].

P1 **Wilm, A.**, Kühnl, J., Kirchmair, J., Computational approaches for skin sensitization prediction, *Critical Reviews in Toxicology*, **48**(9) (2018) 738–760

Available at:
https://www.tandfonline.com/doi/full/10.1080/10408444.2018.1528207

A. Wilm, J. Kühnl and J. Kirchmair conceptualized the work. A. Wilm analyzed the literature and wrote the manuscript, with contributions from J. Kühnl and J. Kirchmair. J. Kühnl and J. Kirchmair supervised the work. All authors have read and agreed to the published version of the manuscript.

Computational tools for the prediction of skin sensitization potential and potency of chemical substances published after the publication of the review article are described in section 3.5 of this thesis. The most common schemes for evaluation of the models as well as methodological background for ML and (Q)SAR modeling approaches are provided in sections 3.4 and 3.2 of this thesis, respectively.

CRITICAL REVIEWS IN TOXICOLOGY
https://doi.org/10.1080/10408444.2018.1528207

Taylor & Francis
Taylor & Francis Group

REVIEW ARTICLE

OPEN ACCESS Check for updates

Computational approaches for skin sensitization prediction

Anke Wilm[a,b], Jochen Kühnl[c] and Johannes Kirchmair[a,d,e]

[a]Center for Bioinformatics, Universität Hamburg, Hamburg, Germany; [b]HITeC e.V, Hamburg, Germany; [c]Front End Innovation, Beiersdorf AG, Hamburg, Germany; [d]Department of Chemistry, University of Bergen, Bergen, Norway; [e]Computational Biology Unit (CBU), University of Bergen, Bergen, Norway

ABSTRACT

Drugs, cosmetics, preservatives, fragrances, pesticides, metals, and other chemicals can cause skin sensitization. The ability to predict the skin sensitization potential and potency of substances is therefore of enormous importance to a host of different industries, to customers' and workers' safety. Animal experiments have been the preferred testing method for most risk assessment and regulatory purposes but considerable efforts to replace them with non-animal models and *in silico* models are ongoing. This review provides a comprehensive overview of the computational approaches and models that have been developed for skin sensitization prediction over the last 10 years. The scope and limitations of rule-based approaches, read-across, linear and nonlinear (quantitative) structure–activity relationship ((Q)SAR) modeling, hybrid or combined approaches, and models integrating computational methods with experimental results are discussed followed by examples of relevant models. Emphasis is placed on models that are accessible to the scientific community, and on model validation. A dedicated section reports on comparative performance assessments of various approaches and models. The review also provides a concise overview of relevant data sources on skin sensitization.

ARTICLE HISTORY
Received 17 April 2018
Revised 3 September 2018
Accepted 21 September 2018

KEYWORDS
Allergic contact dermatitis (ACD); skin sensitization; *in silico* prediction; rule-based approaches; read-across; quantitative structure–activity relationship (QSAR) modeling; machine learning; defined approaches (DAs); integrated approaches to testing and assessment (IATAs); model validation

Table of contents

Introduction

Skin sensitizers are substances able to induce T cell-mediated type IV hypersensitivity immunoreactions in susceptible individuals after topical exposure. Repeated exposure eventually results in clinical manifestations such as skin reddening and itchy rashes, commonly termed allergic contact dermatitis (ACD) (Kimber et al. 2011). ACD is a commonly observed symptom among the general population. A large meta-study reported a weighted average prevalence of 19.5% of ACD involving at least one allergen, most commonly nickel, preservatives, and fragrances, in the general population (Thyssen et al. 2007). ACD is also a major cause of occupational illness (Lushniak 2004; Winkler et al. 2015), and its pervasiveness among hairdressers and dental technicians, for example, is well described (Goebel et al. 2018; Heratizadeh et al. 2018). The mechanisms involved in the induction of skin sensitization have been subject to extensive research and are relatively well understood. The currently accepted adverse outcome pathway (AOP) for skin sensitization comprises a total of 11 steps. Of these, four steps are considered to be key events: (i) molecular interaction of the substance with skin peptides and proteins ("haptenization"; this is the molecular initiating event (MIE)), (ii) activation and inflammatory responses of keratinocytes, (iii) activation of the skin's

CONTACT Johannes Kirchmair johannes.kirchmair@uib.no Department of Chemistry, University of Bergen, Allégaten 41, Bergen, N-5020, Norway

dendritic cells, and (iv) proliferation of hapten-specific T cells (OECD 2012).

Most skin sensitizers are low-molecular-weight xenobiotic chemicals that bind covalently to skin proteins through a Michael addition, Schiff base formation, bimolecular nucleophilic substitution (S_N2), nucleophilic aromatic substitution (S_NAr), or acyl transfer (Aptula and Roberts 2006; Roberts, Aptula, et al. 2007; Chipinda et al. 2011; Enoch et al. 2011; Roberts 2013). Another type of common sensitizers is metals that form coordination complexes with skin proteins and, in the case of nickel, can directly interact with receptors (e.g. human Toll-like receptor 4; TCR) on immune cells (Garner 2004; Schmidt et al. 2010; Martin et al. 2011). Although some substances bind to skin proteins directly, others need to undergo activation through autoxidation (pre-haptens) or enzymatic reactions (pro-haptens) (Karlberg et al. 2008). The relevance of skin permeability in skin sensitization is a field of active research and is not yet fully understood (Alves et al. 2015a; Fitzpatrick et al. 2017a, 2017b).

Human data on skin sensitization remain sparse and of varying quality and reproducibility. From a regulatory point of view, animal experiments for skin sensitization currently constitute the most authoritative testing method for most risk assessment and regulatory purposes. Three animal experiments are accepted for regulatory purposes by the Organization for Economic Co-operation and Development (OECD): the guinea pig maximization test (GPMT), the Buehler guinea pig test (BGPT), and the rodent local lymph node assay (LLNA). Historically, the GPMT and BGPT have been the methods of choice (Ezendam et al. 2016). They have largely been succeeded by the LLNA, which is currently considered the most advanced animal testing system and serves as the primary reference method for the validation of alternatives to animal testing (AATs) (Anderson et al. 2011). Recent studies found that the LLNA correctly discriminates human skin sensitizers from non-sensitizers in approximately two out of three (Alves, Capuzzi, et al. 2018) to three out of four cases (Hoffmann et al. 2018). Besides the identification of a sensitization hazard, the LLNA also allows the determination of the skin sensitization potency of a substance as an EC3 value. The EC3 value is the concentration at which a substance evokes a three-fold stimulation of cell proliferation (measured in draining lymph nodes) in the treated groups compared with the control group. Knowing the potency of a substance is of high interest for risk assessment as this knowledge may allow the application of substances at a safe level of exposure (Adler et al. 2011; Goebel et al. 2017; Kimber et al. 2017).

The LLNA has significant error rates and outcomes can vary considerably. For example, an analysis of 87 substances for which binary LLNA results have been recorded from more than one study (using the same vehicle) identified contradictory outcomes for 19 (22%) of these substances (Dumont et al. 2016). Discordance is higher when different solvents or more than two potency classes are considered, as also reported by Hoffmann (2015).

The LLNA and animal experiments, in general, evoke ethical concerns, and their value for human risk assessment is the subject of ongoing debate (Hartung 2013; Hartung 2017; Alves, Capuzzi, et al. 2018; Hoffmann et al. 2018). Effective since 2013, the European Union's 7th amendment of the cosmetics directive (EUR-Lex 2009) prohibits the sale of cosmetics tested on animals. This leaves a challenging environment for the European cosmetics industry as the risk assessment for the qualification of cosmetic ingredients through alternative testing means such as *in vitro*, *in chemico,* and *in silico* methods is a paradigm shift (Goebel et al. 2012; Ezendam et al. 2016; Goebel et al. 2017; EUR-Lex 2009). Substantial efforts have been made by academic researchers, individual companies and associations from cosmetics, pharmaceutical and fragrance industries as well as institutional laboratories to replace animal experiments with a combination of alternative methods and assessment strategies in compliance with the 3 R (refinement, reduction, and replacement of animal usage in laboratory procedures) concept (Russell and Burch 1959; Basketter et al. 2012; Nendza et al. 2013; Johansson and Lindstedt 2014; Reisinger et al. 2015; Bergers et al. 2016; Ezendam et al. 2016).

Various non-animal testing methods for skin sensitization are available today (Mehling et al. 2012; Thyssen et al. 2012; Reisinger et al. 2015; Ezendam et al. 2016). Six testing methods, addressing the first three key events of the AOP, have been accepted by the OECD for regulatory purposes so far: The direct peptide reactivity assay (DPRA) addresses the MIE of the AOP by measuring the reactivity of a compound toward lysine or cysteine-containing peptides (OECD 2015a). KeratinoSens™ (EURL ECVAM; KeratinoSens assay for the testing of skin sensitizers.) and LuSens (EURL ECVAM; LuSens Assay) address the second key event of the AOP by measuring the activation of the transcription factor Nrf2 in keratinocytes (OECD 2015b). The U937 cell line activation test (U-SENS™), human cell line activation test (h-CLAT), and interleukin-8 reporter gene assay (IL-8 Luc assay) address the third key event of the AOP (OECD 2017a). U-SENS™ and h-CLAT assess the induction of cell surface marker (CD54/CD86) expression in dendritic-like cells (U937 and THP-1, respectively) as a measure for immunogenic cell activation, and the IL-8 Luc assay measures dendritic cell activation through changes in IL-8 cytokine secretion. In recent studies, these non-animal testing methods obtained accuracies (or correct classification rates, CCRs) in the range of 65% to 80% when measured against LLNA and human data (Hirota et al. 2017; Alves, Capuzzi, et al. 2018; Hoffmann et al. 2018).

Besides the OECD-accepted assays, several other promising approaches are in development and/or in the process of validation. Some assays, such as the SENS-IS assay (Cottrez et al. 2015) and the genomic allergen rapid detection (GARD) assay (Johansson et al. 2011, 2013) utilize genomic biomarker signatures to discriminate sensitizing from non-sensitizing substances. Genes relevant to the SENS-IS prediction model were identified by a combination of data mining, literature review, and experimental determination and include (i) a selection of 17 genes that contain a Keap1-Nrf2 signaling pathway-activated antioxidant response element in their promotor and (ii) 21 genes associated with several biological processes (inflammation, danger signals, cell migration) relevant to the activation of dendritic cells. Because of the use of skin models, the SENS-IS assay integrates skin penetration

and metabolism properties of substances although the epidermis model may not completely reflect the in vivo situation. The prediction model of the GARD assay builds on a gene panel selected by an unbiased, genome-wide profiling of the transcriptional response of MUTZ-3 cells to a training set of 20 sensitizing and 20 non-sensitizing substances. The most descriptive genes were identified by principal component analysis (PCA) of differentially expressed genes and subsequent algorithm-based backward elimination (Johansson et al. 2011, Johansson, Rydnert et al. 2014). The SENS-IS assay, which is based on the EpiSkin skin model, addresses keratinocyte activation as the second key event of the AOP, whereas the GARD assay assesses the third key event by analyzing gene expression changes in a human myeloid leukemia cell (MUTZ-3)-derived cell line. Both assays were reported to show high accuracy for hazard identification (SENS-IS: 93% and 91% compared with LLNA or human data, respectively (Cottrez et al. 2016); GARD: 86% accumulated accuracy compared with LLNA data (Johansson, Rydnert et al. 2014; Johansson 2017)). Moreover, both assays were reported to indicate *in vivo* potency. Other approaches to improve predictions are based on the integration of additional parameters to existing testing concepts. For example, the peroxidase peptide reactivity assay (PPRA) adds a peroxidase-dependent oxidation of chemicals with the purpose to improve the detection of pro-haptens with *in chemico* assays. Potential differences of cell lines and primary cells regarding their metabolic capacity and biological responses to external stimuli motivated the development of an optimized protocol for the use of human peripheral blood monocyte-derived dendritic cells (Reuter et al. 2011). Hennen et al. (2011) reported that co-culture of HaCaT cells and THP-1 cells increases the response of THP-1 cells to skin sensitizers compared with that of a monoculture of THP-1 cells. This pertains to the induction of CD54 and CD86, which are readouts essential for the h-CLAT prediction model. The added metabolic capacity of HaCaT cells and the release of keratinocyte danger signals are potential explanations (Hennen et al. 2011). Although the metabolic capacities of cell-based in vitro assays are limited, recent findings indicate that non-animal testing methods are also able to identify sensitizers that require activation through autoxidation or metabolism (Patlewicz et al. 2016; Urbisch, Becker, et al. 2016).

Animal experiments by nature cover the whole process of skin sensitization described in the AOP, including enzymatic or physiological activation of the sensitizer. In contrast, non-animal testing methods focus on single key events of the AOP (Ezendam et al. 2016; Casati et al. 2018). Therefore, the combination of different non-animal testing methods and integration with *in silico* methods is recommended, in particular for the task of potency prediction (Raunio 2011; Mehling et al. 2012; Johansson and Lindstedt 2014; Ezendam et al. 2016; Goebel et al. 2017; OECD 2017a; Casati et al. 2018). Recent studies indicate that such strategies can yield higher prediction accuracies in human hazard estimation than animal experiments (van der Veen et al. 2014; Urbisch et al. 2015; Alves, Capuzzi, et al. 2016; Benigni et al. 2016; Ezendam et al. 2016). Addressing each key event of the AOP individually can also be advantageous for the investigation of the underlying mechanisms (Steiling 2016).

Computational methods promise the ability to predict the skin sensitization potential of substances based solely on their molecular structures. Compared with experimental methods, computational approaches offer the advantage of producing predictions quickly, thus enabling the interactive optimization of compounds. In addition, these methods are cost-effective (Leontaridou et al. 2016), do not require materials for testing, and are not affected by difficulties common to experimental approaches, such as limited solubility, aggregate formation, and evaporation (Hartung 2013). Some recent studies suggest that *in silico* tools could eventually outperform *in vitro* and *in chemico* tools, provided that sufficient data will become available for model development (Asturiol et al. 2016). In contrast to experimental methods, *in silico* tools require defined molecular structures, which are not always accessible, such as in the case of some natural products (Kleinstreuer et al. 2018). In addition, computational methods are generally not applicable to mixtures and metals. Their predictivity and applicability are limited by the quality and quantity of available human, animal, and non-animal data. Luechtefeld et al. (Luechtefeld, Rowlands, et al. 2018) pointed out that future experimental testing efforts should, therefore, focus on the generation of data that can improve model development rather than individual compounds of interest.

In 2008, Patlewicz and Worth produced two reviews that provide a comprehensive overview of computational methods for skin sensitization prediction (Patlewicz and Worth 2008; Patlewicz et al. 2008). Recently, Alves et al. (Alves, Capuzzi, et al. 2018) published a perspective on skin sensitization prediction in which they discuss some of the most relevant computational approaches and data sources.

This work is a comprehensive review of relevant computational approaches for skin sensitization prediction, with a focus on methods and models that have been published after the reviews of Patlewicz et al. and are accessible to the public.

Data sets

Human data on skin sensitization should by nature be most suitable for the development of predictive methods for this endpoint in humans. However, human data remain scarce, vary in quality, and are often difficult to interpret because most of the available human data are no-observed-adverse-effect-levels (NOAELs), which are difficult to interpret and exploit in the context of model development (Politano and Api 2008). Deriving potency information from human epidemiological data is more complex than deriving it from animal testing experiments as it is based on the weighted analysis of (aggregated) exposure and the corresponding number of sensitization incidences. In consequence, non-animal testing approaches and *in silico* models have primarily been developed and validated based on animal data, in particular, LLNA data.

In recent years, a large number of data sets of human, animal, and non-animal data on skin sensitization have been used for model building. However, a closer look reveals that few of these data sets contain significant amounts of new measured data. Most of them are compiled from a few existing sources and thus have substantial overlaps with one another. The most important differences between these data sets is how the data were curated, conflicting information was handled and class labels were assigned.

Here, we will focus on two of the most relevant data sets on skin sensitization: the most comprehensive curated data set on the skin sensitization potential of substances (compiled by Alves et al.; Alves, Capuzzi, et al. 2018) and a high-density data set of compounds relevant to cosmetic application (compiled by Cosmetics Europe; Hoffmann et al. 2018).

Data set compiled by Alves et al.

The Alves data set includes binary LLNA data for 1000 compounds, DPRA data for 194 compounds, KeratinoSens™ data for 190 compounds, h-CLAT data for 160 compounds, and human data for 138 compounds. The data set was prepared following an elaborate data curation protocol that includes, among many other steps, the removal of entries with discordant biological outcomes for the same data type. The provenance of the data is documented. Data concordance and chemical space analyses provide additional information on the consistency and coverage of the individual subsets. For example, the authors found that 65% to 79% of the binary data recorded for any of the three non-animal testing methods are in agreement with the LLNA outcomes. They also reported that for 801 of the 1000 substances measured in the LLNA no other types of data were available.

The LLNA data that is included in the Alves data set was compiled from the work of Luechtefeld et al. (2016), Jaworska et al. (2013), and from the NICEATM LLNA database (ICCVAM 2013). LLNA data on 566 unique compounds (197 sensitizers and 369 non-sensitizers; after curation by Alves et al.) originate from the Luechtefeld data set, representing a collection of publically available *in vivo* and non-animal data on the skin sensitization potential of (primarily) high-production volume chemicals, all of which have been submitted for the REACH registration process. The REACH data set (as evaluated by Luechtefeld et al.) contains information on close to 20 000 studies conducted on the skin sensitization potential and potency of substances but requires further curation prior to use for model development. The current version of the REACH data set is available on the website of the European Chemical Agency (ECHA; ECHA. Homepage) and can be filtered through the OECD eChemPortal (OECD; eChemPortal). Most recently, Fitzpatrick et al. (2018) extracted GPMT and LLNA data on the skin sensitization potential of 1295 substances mainly from this database.

LLNA data on 145 substances included in the Alves data set originate from the work of Jaworska et al. (2013). The aim of Jaworska et al. was the compilation of a diverse, high-quality data set on the skin sensitization potency of substances for which LLNA, *in chemico* and *in vitro* data (i.e. DPRA, KeratinoSens™, and a CD86 activation assay based on the U937 cell line) are available. As such, the substances included in this data set cover different potency classes (from non-sensitizers to extreme sensitizers) and a wide range of physicochemical properties and usage classes (e.g. fragrances, preservatives, dyes, dye precursors, and solvents). Jaworska et al. applied strict quality filters. For example, they only included data derived in agreement with the corresponding OECD protocols and for which either a negative result or a clear dose-response curve is reported.

The LLNA data for 516 substances (332 sensitizers and 184 non-sensitizers) included in the Alves data set originate from the NICEATM LLNA database (ICCVAM 2013). The NICEATM LLNA database is one of the most comprehensive collections of EC3 data on a diverse range of chemicals. The database has been compiled from, among other sources, the work of Gerberick et al. (2005), the work of Kern et al. (2010) and donated company data. The documentation of the data provenance allows the lookup of the original sources of individual data points. Importantly, the vehicle of each study is also recorded, which can support the interpretation of discordant data.

The Alves data set also contains human data on the skin sensitization potential of 138 substances. These data originate from the ICCVAM human database (ICCVAM 2011) and the Strickland data set (Strickland et al. 2017). The ICCVAM human database consists of 302 substances that have been compiled for the evaluation of LLNA potency prediction. The Strickland data set consists of 96 substances covering a wide area of product usages (e.g. manufacturing chemicals, food additives, pharmaceuticals, fragrances, personal care products, pesticides, and cosmetics). In addition to human data, the Strickland data set also contains LLNA data, outcomes from non-animal testing approaches (DPRA, KeratinoSens™, and h-CLAT), as well as (primarily measured) data on six relevant physicochemical properties for the same substances for which human data are provided.

The non-animal testing results collected by Alves et al. originate from the data set of Urbisch et al. (2015). The Urbisch data set includes LLNA-derived EC3 values for 213 substances for which information from at least two non-animal testing methods was available. The data set covers substances from diverse use contexts such as pharmaceuticals, pesticides, fragrances, and preservatives. The LLNA data are accompanied – when available – by human data and results from non-animal testing methods (DPRA, KeratinoSens™, LuSens assay, h-CLAT, myeloid U937 skin sensitization test (MUSST), and modified MUSST (mMUSST)).

Overall, the comprehensiveness and quality of the Alves data set make it one of the most valuable resources for the development of computational models. However, some information that may be of value for the interpretation and modeling of the data has not been transferred from the original source into the Alves data set, such as potency information (e.g. EC values or potency classes), data measured for mixtures, and information on repeated measurements. The documentation of data provenance allows the retrieval of such information from the original sources.

Cosmetics Europe data set

Cosmetics Europe, the trade association of the cosmetics industry in Europe, compiled an almost complete matrix of 128 substances measured with the DPRA, KeratinoSens™, h-CLAT, U-SENSTM, and SENS-IS assays (Hoffmann et al. 2018). The data are accompanied by LLNA-derived potency information, human potency categorization according to Basketter et al. (six classes; Basketter et al. 2014), and information on six (primarily measured) physicochemical properties associated with skin penetration and protein binding.

The substances included in this data set are of high relevance to the cosmetic application and include 58 fragrances, 16 preservatives, 9 actives, 7 surfactants, 7 dyes, 6 pharmaceuticals, and 25 substances assigned to other categories. Thirty-eight of these substances are Michael acceptors, 21 are Schiff base electrophiles, 11 are S_N2 electrophiles, 9 are acyl transfer agents, and 2 are S_NAr electrophiles (41 were not assigned to a reaction domain). The substances have a molecular weight between 30 and 605 Da and a water solubility (log*S*) between −7 and 2.

The LLNA data included in the Cosmetics Europe data set were retrieved from several sources, including the NICEATM LLNA database and a proprietary database from the Research Institute for Fragrance Materials (RIFM). For 57 substances, EC3 values were collected from more than one LLNA study and merged using a newly developed, median-like location parameter. Human data were collected from Basketter et al. (2014) and Api et al. (2017), whereas non-animal testing data were retrieved from, among others, Urbisch et al. (2015) and Natsch et al. (2013). The non-animal testing data not only comprise binary testing results but also several (partly quantitative) outcomes for each testing method.

All data included in the Cosmetics Europe data set are based on experiments following standard protocols, thereby facilitating the comparability and reliability of the data. Approximately one-third of the non-animal testing data were newly generated by Cosmetics Europe. Any external data were individually reviewed and cross-checked by a second reviewer to ensure the high quality of the data.

The number of substances covered by the Cosmetics Europe data set is comparable with that of other data sets that include different data types (e.g. the data sets of Strickland et al. (2017), Jaworska et al. (2013), and Urbisch et al. (2015)). However, the Cosmetics Europe data set includes significantly more results from non-animal testing methods. Its high quality and consistency make it a valuable resource, in particular for the benchmarking of new theoretical and experimental methods for the prediction of the skin sensitization potency of substances. In comparison with the Alves data set, the Cosmetics Europe data set has a much higher information density for individual substances and a clear focus on substances relevant to the cosmetic application. In contrast, the Alves data set is designed to cover the largest possible chemical space. As such, the Alves data set is particularly valuable for the development of machine learning models for the classification of substances into skin sensitizers and non-sensitizers.

Computational methods for skin sensitization prediction

In this section, we discuss important AATs to predict skin sensitization that either are pure theoretical methods (e.g. rule-based approaches, statistical models, machine learning, models and hybrid models) or (can) include a computational component. A schematic overview of these approaches is provided in Figure 1. Additional information on the most relevant *in silico* models is reported in Table 1.

(Q)SAR modeling approaches

(Q)SAR approaches aim to describe the correlation between the structure and activity of compounds (Gleeson et al. 2012; Cherkasov et al. 2014). In the current context, activity is the skin sensitization potential or potency. Classification models are primarily utilized for the categorization of compounds according to their predicted skin sensitization potential whereas regression models are most commonly used for the quantitative prediction of potency. For risk assessment, quantitative predictions are clearly preferred over categorical models because the latter generally require to adopt the assumption of maximum activity within the predicted category (Adler et al. 2011; Goebel et al. 2017; Kimber et al. 2017). Where possible, a combination of regression and classification models can be of advantage with respect to the accuracy, applicability domain (AD), and interpretability of models.

Depending on the scale of the chemical space covered, (Q)SAR models can either be local or global (Helgee et al. 2010). Local models cover a well-defined, narrow chemical space within which they generally obtain high prediction accuracy. In contrast, global models aim to cover the broadest possible chemical space and thus have a large AD, often at the cost of prediction accuracy. As is true for most toxicological endpoints, the relationships observed between chemical structure and skin sensitization potency is not linear, in particular for more diverse sets of data. It is, therefore, generally difficult to develop linear models for toxicity prediction with a large AD. Nonlinear machine learning approaches have proven particularly successful in this context but are more difficult to interpret (Gleeson et al. 2012).

The OECD has defined guidelines for (Q)SAR models for use in a regulatory environment (OECD 2014b). These include that the models should have a defined endpoint, an unambiguous algorithm, a defined domain of applicability, appropriate measures of goodness-of-fit, robustness and predictivity, and, if possible, should allow a mechanistic interpretation. The latter is also important for the validation of models because it reduces the risk of overfitting due to non-causal but correlated features (Luechtefeld, Rowlands, et al. 2018).

Whereas goodness-of-fit (how well the model accounts for the variance of the response in the training data) and robustness (how stable the model is when one or more instances of training data are removed) of a model can be evaluated internally during model development, predictivity (how well the model can predict new data) can only be evaluated

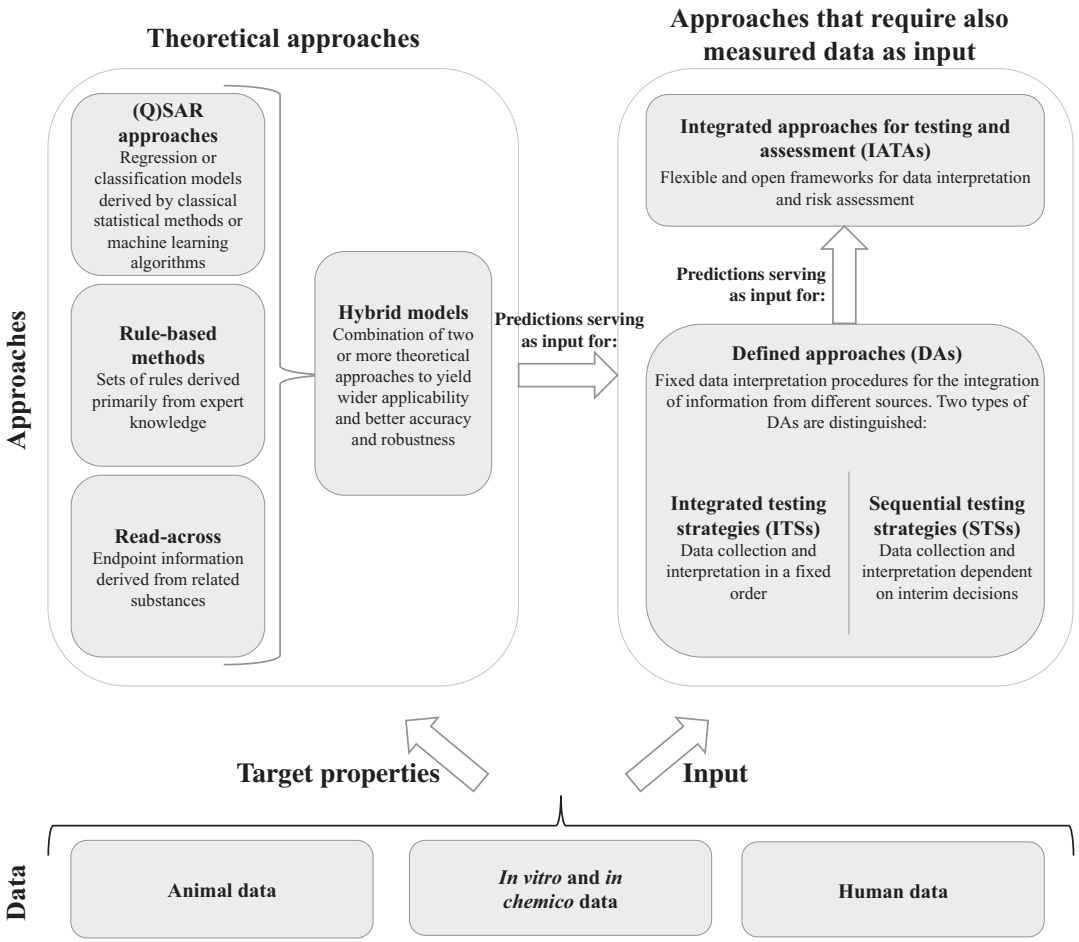


Figure 1. Overview of approaches for skin sensitization prediction that are either pure *in silico* models or can include a computational component.

externally on the basis of new data not used for model development (OECD 2014b). This external evaluation requires the available data to be split into a training and a test set prior to modeling, and the test set must be used for model evaluation only. Unfortunately, for a significant number of available models, no such external tests have been conducted, primarily because of the scarcity of data on skin sensitization. Instead, only results for cross-validation or, in the worst case, only the training data, are reported. The first may lead to an overestimation of model performance; the latter almost certainly will. But even in cases where external validation is carried out, analyses of the representativeness and diversity of the test data with respect to the training data, as well as the AD of a model, are all too often missed.

The AD of a model describes the property and/or structural space of substances for which a model can make reliable predictions. Consideration of the AD is therefore of the essence to the application of any model. The AD is based on the assumption that similar predictivity can be achieved for substances that are similar to those in the training data. It is therefore considered to depend on structural, physicochemical, and response information in the data used for training a model, with the selection of important parameters also depending on the modeling algorithm used (OECD 2014b). A large number of different methods are available for the definition of the AD. For an overview of these techniques, the reader is referred to the OECD's Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models (OECD 2014b).

Several thousand molecular descriptors are at our disposal today (Todeschini et al. 2009). They can be differentiated according to the type of information they encode. 0D descriptors encode properties that can be directly derived from the chemical formula (e.g. molecular weight, number of heavy atoms), whereas 1D descriptors capture the presence or absence of substructures in a molecule. 2D descriptors are derived from the molecular graph and encode the atom connectivity of molecules. Finally, 3D descriptors are derived from the 3D structure of molecules and capture, for example, the molecular surface area or quantum chemical properties such as HOMO-LUMO energies. For models to be mechanistically interpretable and robust, the use of small sets of physically meaningful descriptors is preferred. Therefore, for the development of skin sensitization models, experts often

Table 1. *In Silico* models for the prediction of skin sensitization potential or potency.

Models and software	Core components	Description	Availability	Online resources	Scientific literature
Linear (Q)SAR modeling approaches					
RAI/QMM models	Linear regression	Mostly quantitative linear models for predicting the skin sensitization potency (EC3 values) of a defined range of compounds. Make use of a few, physically meaningful descriptors of chemical reactivity and/or hydrophobicity	Linear equations available from the primary literature		For example Roberts and Williams1982; Promkatkaew et al.2014; Roberts, Schultz, et al.2017
Warne et al. model	Linear regression	Binary classifier for predicting skin and respiratory sensitization potentials	Linear equations available from the primary literature		Warne et al.2009
TOPKAT	Linear regression	Ternary classifiers for predicting the skin sensitization potency. Package including three original and two extensible models. The latter is based on a modified Bayesian learning method and uses seven descriptors	Commercial	BIOVIA. QSAR, ADMET, and Predictive Toxicology	Enslein et al.1997; BIOVIA2017a, 2017b
Toropova et al. model	Linear regression	Linear statistical models for predicting skin sensitization potency (EC3 values). Model based on hybrid optimal descriptors combining information calculated from hydrogen-suppressed graphs as well as from SMILES strings	Free		Toropova and Toropov2017
Nonlinear (Q)SAR modeling approaches					
Lu et al. model	Decision Tree	Binary classifier for predicting the skin sensitization potential of compounds based on a decision tree	Decision tree available from the primary literature		Lu et al.2011
Pred-Skin	Random forest	Binary/ternary classifiers for predicting the LLNA and human skin sensitization potential	Free web service	LabMol	Braga et al.2017
VEGA	Adaptive fuzzy partition algorithm	Binary classifier for predicting the skin sensitization potential	Free	VEGA HUB	Chaudhry et al.2010
Rule-based methods					
ToxAlerts	Structural alerts	Set of structural alerts compiled from different sources for assigning of compounds to reaction domains	Free web service	OChem	Sushko et al.2011, 2012
Toxtree	Decision tree based on structural alerts	Set of rules for assigning compounds to one of the five established skin sensitization reaction domains and to nonspecific protein binding alerts	Free, also as a web service	Toxtree	Enoch, Madden, et al.2008; Enoch et al.2011
Read-across					
VEGA	Read-across	Model for predicting the skin sensitization potency (EC3 values) of alkenes reacting through Michael addition	Free	VEGA	Enoch, Cronin, et al.2008
MuDRA	Multi-descriptor read-across	Consensus model of four read-across approaches performed with different descriptors or fingerprints for predicting skin sensitization potentials	Free	Chembench	Alves, Golbraikh, et al.2018
Hybrid models					
Derek Nexus	Nearest-neighbor approach for compounds triggering the same structural alert	Expert system based on 90 structural alerts. Verification of negative results by comparison with known false positives and by checking for structural features not present in the training data. Prediction of EC3 values based on a nearest-neighbor approach within compounds triggering the same structural alert	Commercial	Lhasa Limited	Barratt et al.1994; Langton et al.2006; Canipa et al.2017
OECD QSAR Toolbox	Plattform for read-across including rule-based profilers	Software for toxicity prediction that includes two profilers for skin protein binding and one profiler for general protein binding. Integration of autoxidation and skin metabolism through additional profilers is also possible	Free	OECD. The OECD QSAR Toolbox	
TIMES-SS	Structural alerts combined with 3D QSAR	Structural alerts discriminating three potency classes. Compounds not assigned to any of these potency classes are further assessed with QSAR models. Also includes modules for the prediction of autoxidation and volatility. Can be combined with the TIMES skin metabolism module	Commercial	OASIS-LMC. TIMES-SS Software	Dimitrov et al.2005; Mekenyan et al.2012
CASE Ultra		Includes models for predicting electrophilicity, protein binding, ARE activation in keratinocytes, activation of dendritic cells, LLNA outcomes and ACD induction in human and guinea pigs. Statistical analysis of positive and negative structural alerts and modulating structural features. Binary prediction of the skin sensitization potential plus potency prediction for several mechanistic groups of compounds	Commercial	MultiCASE	Klopman1992; Graham et al.1996; Chakravarti et al.2012; Saiakhov et al.2013

(*continued*)

Table 1. Continued.

Models and software	Core components	Description	Availability	Online resources	Scientific literature
Dearden et al. model	Structural alerts combined with local linear models	Structural alerts for assigning compounds to reaction domains. EC3 prediction within the reaction domains by local linear QSAR models	Equations available from the primary literature		Dearden et al.2015
CADRE-SS	Structural alerts combined with Monte Carlo approach, quantum mechanical calculations and linear QSAR modules	Assignment of compounds to three or five potency classes by linear QSAR integrating (i) a calculated permeability coefficient ($\log K_p$), (ii) the rule-based assignment of a reaction domain and (iii) reactivity based on quantum mechanical calculations	Commercial	ToxFix	Kostal and Voutchkova-Kostal2016
REACH*Across*™	Read-across combined with machine learning	Combination of a fingerprint-based similarity analysis with machine learning techniques for the binary prediction of the skin sensitization potential	Commercial	UL	Luechtefeld, Rowlands, et al.2018
Ellison et al. model	Consensus model of different *in silico* tools	Integrates results from the OECD QSAR Toolbox, Derek for Windows, the global model of CAESAR and the SMARTS rules of Enoch et al. (Enoch, Madden, et al.2008) for the binary prediction of the skin sensitization potential	Equations available from the primary literature. Requires additional commercial software		Ellison et al.2010)
SkinSensDB	Read-across combined with a decision tree or 2-out-of-3 majority voting	Provides access to two published ITS approaches for predicting the skin sensitization potential. These ITSs can be performed with non-animal test results provided by the SkinSensDB or with data derived by read-across within the platform	Free web service	SkinSensDB	Wang et al.2017; Tung et al.2018

choose to work with descriptors encoding properties related to the ability of a compound to penetrate the skin (e.g. molecular weight, molecular volume, and log*P*) and react with skin proteins (e.g. HOMO-LUMO energy gap, activation energies, and reaction rates). A plethora of descriptor selection methods also is available that allow the automated selection of small numbers of descriptors with high information content.

Linear models

Chemical class or mechanism-based models. One of the earliest examples of linear QSAR models for skin sensitization prediction is the relative alkylation index (RAI) model developed by Roberts and Williams (1982). This model allowed the quantitative prediction of the sensitization potential of sultones as a function of their reactivity toward proteins. In its most general form, the original RAI model can be formulated as

$$\mathrm{RAI} = \log D + a\ \log k + b\ \log P$$

with D being the molar dose, k the alkylation rate constant, P the partition coefficient between a standard polar/nonpolar solvent, and a and b being constant prefactors.

Since the publication of the initial RAI model, further RAI models, applicable to a defined set of structurally closely related chemicals, and quantitative mechanistic models (QMMs), applicable to a wider range of compounds that share similar reaction chemistry, have been developed. For a comprehensive review of these types of models, the reader is referred to the work of Patlewicz and Worth (2008). More recently, the RAI/QMM concept has been applied to predict the skin sensitization potency of epoxides (Roberts, Aptula, et al. 2017), aldehyde Schiff bases (Roberts, Schultz, et al. 2017), Michael acceptor electrophiles (Roberts and Natsch 2009; Wondrousch et al. 2010) and molecules undergoing aromatic substitutions (Roberts et al. 2011; Ouyang et al. 2014; Roberts and Aptula 2014). Although the original RAI model required the experimental measurement of P and k to derive the skin sensitization potential of a compound of interest, more recent models either incorporate calculated p values or neglect the parameter and derive k based on precalculated reactivity parameters (Roberts, Schultz, et al. 2017).

Enoch and Roberts (2013) reported a linear model for the prediction of the potency (pEC3 value) of Michael acceptors as a function of the available surface area at the site of reaction and the stability of the expected reaction intermediate (which correlates with the reaction rate k). The latter descriptor is calculated from the sum of the ground state energies of the query molecule and a probe, as well as the energy of the charged intermediate using density functional theory (DFT). The model was developed based on LLNA data for 33 Michael acceptors and predicted pEC3 values with an R^2 of 0.79 (after the removal of several outliers).

Linear approaches for the prediction of aromatic substitutions include a model by Promkatkaew et al. (2014), who found a moderate correlation ($r^2 = 0.64$) between the energy barriers (derived by modeling the reaction pathways with DFT) and pEC3 values of 12 sensitizers. Interestingly, no correlation was found between the HOMO−LUMO energy (which is frequently used to encode chemical reactivity) and

the pEC3, indicating that the HOMO−LUMO energy is not a relevant attribute for this class of compounds and reactions.

All of these RAI/QMM models have in common that they are based on only a few, mechanistically interpretable features, which minimizes the risk of overfitting. However, the models are derived from very small, focused data sets, which greatly limit their AD. This limitation can be mitigated through combination with other models, provided that the reaction domains and molecular classes of substances of interest can be correctly assigned. Many hybrid approaches combine local linear models, as described in the section "Hybrid *in silico* models."

Models applicable to a wider range of chemical classes and mechanisms. Linear models with a broader AD make use of larger and more diverse data sets and often use feature selection algorithms to identify a small subset of relevant descriptors. The descriptors selected by these automated procedures are not necessarily physically meaningful or easily interpretable. Manual refinements based on expert knowledge are therefore generally advised or even necessary. This type of model is also limited by the fact that the relationship between substances and their skin sensitization potency or potential is not linear when observed on a larger scale.

More broadly applicable linear (Q)SAR models include a categorical model for the prediction of skin and respiratory sensitization potential (Warne et al. 2009). This model is based on a data set of 119 compounds annotated with GPMT and LLNA, as well as animal and human inhalation data. Most of these data have been obtained from the Annex I of the Dangerous Substances Directive of the European Union (67/548/EEC). Linear regression was performed to select the eight most relevant descriptors (representing molecular orbital energies, differences thereof, and electronegativity) for the discrimination of skin sensitizers and non-sensitizers from a set of 59 descriptors. Although the model was able to correctly identify four of the five skin sensitizers (and both respiratory sensitizers) from a test set of 17 substances, the ability of the model to discriminate skin from respiratory sensitizers was insufficient. The poor discrimination between these types of sensitizers is likely linked to shared chemical properties.

TOPKAT includes several categorical models for skin sensitization. The original TOPKAT models are based on the work of Enslein et al. (1997). A global model combines two linear binary equations, one to distinguish non-sensitizers and sensitizers, and the other to further classify sensitizers into weak and strong categories. This global model is complemented by two local models, one covering aliphatic and single-benzene-ring-containing chemicals, and the second model covering the remaining aromatics. Historically, TOPKAT is one of the first models for skin sensitization to include a definition of the AD. For the original model, a specificity of 79% and a sensitivity of 82% were reported for an independent test set of 25 compounds. In addition to the original models, two newer, extensible models for the prediction of skin sensitization based on a modified Bayesian learning method are available (BIOVIA 2017a, 2017b). One of these models differentiates sensitizers and non-sensitizers, and the other model differentiates weak and strong sensitizers. Both extensible models make use of seven molecular descriptors, including log*P*, molecular weight, polar surface area, number of donors, acceptors and rotational bonds, and an atom type fingerprint. The classification model for sensitizers and non-sensitizers is trained on GPMT data for 392 compounds and achieved a ROC score of 0.77 in 10-fold cross-validation, whereas the strong vs weak sensitizer model was trained with GPMT data for 258 compounds and obtained a ROC value of 0.92 during leave-one-out cross-validation.

More recently, Toropova and Toropov (2017) used the Monte Carlo approach implemented in CORAL to derive four continuous linear regression models from a training set of 147 compounds annotated with measured EC3 data. The models were derived based on hybrid optimal descriptors combining information calculated directly from SMILES strings and from hydrogen-suppressed molecular graphs. The best performing model obtained an r^2 of 0.86 for the quantitative prediction of EC3 values for an external test set of 29 compounds. The skin sensitization potency of compounds was observed to increase with the presence of five-membered rings, aromatic six-membered rings, and double bonds.

Nonlinear models

In recent years a wide range of nonlinear approaches has been explored to model the complex relationships between substances and their skin sensitization potential and potency. In particular, machine learning algorithms can account for the nonlinear relationships observed in large and diverse data sets. With increasing amounts of data, manual curation by experts is often replaced with automated and less reliable data curation procedures, which can have a negative impact on the quality of data sets used for modeling.

Machine learning algorithms can deal with large numbers of descriptors. Often, a substantial number of descriptors are calculated and subjected to feature selection procedures prior to or as part of the model generation process. The use of large numbers of descriptors entails the risk of model overfitting. In addition, the inclusion of descriptors that are not physically meaningful adds to the black box character of complex machine learning models, making it difficult if not impossible to understand on which basis the algorithm assigns a substance to a certain biological outcome. Taken together, these issues often lead to the neglect or insufficient definition of the AD of models, which is not only problematic for the application of the individual models but also for the perception and reputation of computational methods in general.

Lu et al. (2011) used recursive partitioning to derive a decision tree model for the binary classification of sensitizers and non-sensitizers based on LLNA data for 295 compounds, including, among others, Michael acceptors, S_N2 and S_NAr electrophiles, Schiff base formers, and acyl transfer agents. Eight quantum chemical and physicochemical descriptors linked to chemical reactivity, hydrophobicity, and electrostatic interaction, as well as a fragment descriptor, served as the input for model building. The fragment descriptor encodes the presence or absence of eight substructural features by a single binary value. The final model correctly classified ~80% of all (25 and 37) compounds of two test sets.

Alves et al. (2015b) derived random forest models for the classification of skin sensitizers and non-sensitizers from a curated and balanced subset of 127 sensitizers and 127 non-sensitizers extracted from the NICEATM LLNA database. Two types of models were developed, one based on 0D, 1D and 2D descriptors calculated with Dragon (Talete S.r.l. Dragon) and the other based on 2D SiRMS descriptors (simplex representation of molecular structure, encoding molecular structure by tetratomic fragments of fixed composition, structure, chirality, and symmetry; Muratov et al. 2010). A consensus model based on models derived from either type of descriptor performed best on an external test set containing 152 sensitizers but only five non-sensitizers, with a CCR of 0.86. Because of the strict definition of the AD applied in this test, predictions were only made for 24% of the compounds of the test set. A consensus model with a less stringent definition of the AD reached slightly lower predictivity (CCR = 0.83) but higher coverage (50%) on the same test set. Five-fold cross-validation on balanced data resulted in comparable classification accuracies but significantly higher coverage (coverage 39% and 70%, depending on the definition of the AD). Based on this work, a free web service called Pred-Skin was developed, which allows the prediction of the skin sensitization potential of substances based on random forest models derived from Morgan2 fingerprints (Braga et al. 2017). Two of these models are binary classification models based on human skin sensitization data (for 109 compounds) and LLNA data (for 515 compounds). During five-fold cross-validation, these two models obtained CCRs of 0.80 and 0.84 for the two-thirds of all compounds that were within the AD. A third model discriminating three potency categories based on LLNA data obtained an accuracy of 0.76 and coverage of 78% under the same test scenario.

Yuan et al. (2009) developed a binary support vector machine (SVM) classifier for the prediction of the skin sensitization potential of substances based on LLNA and GPMT data on 108 and 61 organic compounds (including, among others, alkanes, aromatic hydrocarbons, alcohols, amines, acids, and esters), respectively. Particle swarm optimization (PSO) was used for the selection of important 2D molecular descriptors from a set of 926. The final model was based on six molecular descriptors corresponding to the number of chlorine atoms, the molecular electronic structure, molecular size, and hydrophobic properties. It obtained classification accuracies of 89% and 90% on LLNA and GPMT data for 54 and 31 compounds in the two test sets, respectively.

Within the EU-funded CAESAR project (CAESAR. CAESAR project), two global binary categorical models for the prediction of the skin sensitization potential of compounds were developed based on LLNA data compiled for 167 chemicals (Chaudhry et al. 2010). One of the models was derived from an in-house adaptive fuzzy partition algorithm. As part of this approach, a hybrid method combining a genetic algorithm with stepwise regression was used to select seven relevant 2 D descriptors calculated with Dragon (i.e. the number of nitrogen, double-bonded oxygen, and non-aromatic conjugated sp2 carbon atoms, as well as descriptors accounting for topological features, charge, and valence connectivity). The model obtained 90% classification accuracy on a test set of 42 compounds (8 non-sensitizers and 34 sensitizers) and is distributed as a component of VEGA (Benfenati et al. 2013; CAESAR. Skin sensitization model). The other model was derived with a multilayer perceptron neural network algorithm trained on a slightly modified training set. For the identical test set with a different threshold for the division between sensitizers and non-sensitizers (resulting in 21 non-sensitizers and sensitizers each), an accuracy of 71% was obtained with this model (which has not been implemented in VEGA).

Rule-based approaches

Knowledge-based expert systems have a long record of successful use in ADME (absorption, distribution, metabolism, and excretion) and toxicity prediction. A key component of these systems is dictionaries (sets of rules), which aim to encode existing empirical knowledge distilled from *in vitro* and *in vivo* data, as well as from clinical practice. These rules generally link structural fragments to mechanisms of skin sensitization but may also be more complex than simple structural alerts. For example, they may also take into account skin penetration, chemical reactivity, or steric accessibility. Rule-based methods are easily interpretable and inherently subjective.

The relevance of rule-based methods for skin sensitization prediction stems to a significant extent from the sparsity and limited reliability of the available data, which poses a bottleneck in model development. Whereas statistical approaches and machines require a significant number of instances to identify, support, and weigh patterns in the data, experts may be able to derive valid rules from a limited number of observations. In the absence of sufficient hard data, expert knowledge may allow artificial extension of the scope of rule-based approaches. For example, experts may implement rules stating that two chemical substructures behave similarly in a defined context (Enoch, Madden, et al. 2008).

On the downside, in certain cases, expert bias may hinder corrections of rule-based systems or even the further data collection. For example, molecules with a molecular weight above 500 Da have generally been assumed to be too large to diffuse into the epidermis and cause skin sensitization. As a consequence of this assumption, only a few compounds with a molecular weight above 500 Da have been evaluated regarding their skin sensitization potential. For example, among the of 211 compounds of the LLNA data set compiled by Gerberick et al. (2005), only two compounds have a molecular weight above 500 Da, one of which is a known skin sensitizer (Fitzpatrick et al. 2017a). It is only more recently that awareness about the skin sensitization potential of compounds with a molecular weight above 500 Da has been raised (Roberts et al. 2013; Alves et al. 2015a; Fitzpatrick et al. 2017a; Luechtefeld, Rowlands, et al. 2018).

In the context of skin sensitization prediction, rule-based systems are nowadays more often part of hybrid computational models than used as individual models (see section

"Hybrid in silico models"). However, several existing platforms allow the screening of substances of interest for the presence of structural features related to skin sensitization. One example is ToxAlerts (Sushko et al. 2011, 2012), which provides, among others, structural alerts for potential skin sensitizers based on sets of rules distilled from different sources (Barratt et al. 1994; Payne and Walsh 1994; Gerner et al. 2004; Kazius et al. 2005). The rule set which was originally implemented in the toxicity prediction model DEREK is also included in ToxAlerts. The current version of DEREK, Derek Nexus, includes additional modules for skin sensitization prediction, for which reason the software is discussed in the section "Hybrid *in silico* models."

Another example of a rule-based system for the assignment of substances to one or several skin sensitization reaction domains (the "Skin Sensitisation Reactivity Domain" module) has been implemented in Toxtree (Enoch, Madden, et al. 2008; Toxtree). The rule set was derived from LLNA data measured for 208 compounds and encodes substructures associated with the five established skin sensitizing reaction domains. The structural alerts also take metabolism and oxidation (but not bioavailability) into account. Toxtree also features a set of 104 structural alerts for protein binding related to acylation, Michael addition, Schiff base formation, S_N2 and S_NAr (Enoch et al. 2011). As protein binding is the first key event in the AOP for skin sensitization, these structural alerts might also be useful for the prediction of skin sensitization.

The OECD QSAR Toolbox (OECD. The OECD QSAR Toolbox) provides many profilers that may be used for building chemical categories for subsequent read-across (see section "Read-across") within the software package. In particular, the profilers for skin protein binding and general protein binding are of relevance to the prediction of the skin sensitization potential of compounds. Simulators of autoxidation and skin metabolism are also implemented in the OECD QSAR Toolbox and may be of value to the refinement of predictions. These simulators are also part of TIMES (OASIS-LMC; TIMES-SS Software), which includes additional capabilities for the assessment of biotransformations, information on the AD and metabolic maps.

In general, structural alerts alone are an insufficient predictor of the skin sensitization potential or potency (Alves, Muratov, et al. 2016). Toxtree and the profilers of OECD QSAR Toolbox, for example, are not intended to be used as predictors but rather as tools to assign substances of interest to reaction domains (Enoch, Madden, et al. 2008; OECD. The OECD QSAR Toolbox). Nevertheless, these tools are frequently investigated as potential predictors, i.e. such that any substance matching structural alerts related to one of the reaction domains is deemed a skin sensitizer (Urbisch, Honarvar, et al. 2016; Verheyen et al. 2017). However, comparative studies have shown that structural alerts may be able to improve predictions when used in combination with other approaches (Teubner et al. 2013; Verheyen et al. 2017; see the section "Comparative analyses of the performance of computational models for skin sensitization prediction").

Read-across

Read-across is an approach for the prediction of endpoint information based on available data on the same endpoint of related substances (Patlewicz et al. 2013; OECD 2014a; Schultz et al. 2015). This method is a pillar of risk assessment for many toxicological endpoints but does not necessarily involve computation. Read-across can either be performed as an analog approach, where – in the absence of a trend or regular pattern of biological properties – a missing property value of a compound of interest is predicted based on one or several other compounds with known values for this property, or as a grouping approach, in which predictions for a compound of interest are derived from several structurally related source compounds with similar properties or properties following a regular pattern (Patlewicz et al. 2017). Like rule-based methods, read-across approaches are generally easily interpretable and extendable with new data (e.g. in-house data).

An example of a local read-across tool for the prediction of the skin sensitization potency of alkenes reacting through Michael addition has been implemented in VEGA (Enoch, Cronin, et al. 2008). It is based on a database of Michael acceptors with measured EC3 values and DFT-derived electrophilicity indices (ω). For any compounds of interest, the potency is derived based on the EC3 values of compounds with similar ω.

Recently, Alves et al. (Alves, Golbraikh, et al. 2018) published a multi-descriptor read-across (MuDRA) consensus model that integrates read-across based on various types of chemical descriptors and molecular fingerprints. In a test on 217 skin sensitizers and non-sensitizers, of which 42% were within the AD and considered in the analysis, the consensus model obtained a binary classification accuracy of 0.78. It is not entirely clear whether the improved prediction accuracy of the consensus approach over the best-performing individual model justifies the more complex approach.

Hybrid in silico models

Hybrid *in silico* models combine two or more of the above-mentioned components with the aim to improve prediction accuracy and the applicability of computational methods. For example, the combination of complementary approaches such as a reactivity model based on quantum mechanical calculations with a rule-based approach can be particularly beneficial for potency prediction. Importantly, the agreement or disagreement of predictions from individual components is not necessarily an indicator of reliability. For example, overlaps in the training data, the knowledge base and/or modeling methods of the individual components can lead to correlations (Rorije et al. 2013; Fitzpatrick et al. 2018). Care must be taken to not wrongly interpret such correlations as founded indications of the high reliability of predictions.

In the context of skin sensitization prediction, the majority of hybrid models are composed of two or more dependent modules that may not be used individually. However, there are also hybrid models in existence that integrate

self-sufficient modules for skin sensitization prediction. These are commonly referred to as consensus models.

An example of a hybrid model that integrates several dependent models for the prediction of skin sensitization potency is Derek Nexus (formally Derek for Windows or DEREK; Barratt et al. 1994; Payne and Walsh 1994; Gerner et al. 2004; Kazius et al. 2005). The core component of Derek Nexus is an expert system based on 90 structural alerts for the prediction of skin sensitization potential that also includes functionality for verifying negative predictions by (i) comparing them to the structures of compounds that are known to be predicted as false-negatives by the model and (ii) scanning them for substructures not covered by the training data (Williams et al. 2016). An additional component predicts EC3 values for any compounds triggering a skin sensitization alert by calculating the weighted average of the EC3 values of 3–10 nearest neighbors (of an LLNA data set containing a total of 465 compounds) matching that alert (Canipa et al. 2017). The molecular similarity of individual pairs of molecules is evaluated based on an in-house radial molecular fingerprint. A likelihood level ranging from "certain" to "impossible" is provided together with the predicted EC3 value. For an external test set of 103 compounds, Derek Nexus correctly predicted the EC3 values of half of all tested compounds with less than a five-fold deviation from the LLNA-derived value. In addition, the software correctly assigns 64% of all tested compounds to one of the three categories of the Globally Harmonized System of Classification and Labelling (GHS) recommended by the United Nations Economic Commission for Europe (UNECE) for a standardized classification of skin sensitizers and non-sensitizers according to potency. The error rates differed significantly depending on the skin sensitization alert triggered within Derek Nexus. They were particularly high for metal and metal salts, as well as for substituted phenols and their precursors. In a recent evaluation (Chilton et al. 2018), Derek Nexus obtained a sensitivity of 54% and a specificity of 77% when used to discriminate between 302 skin sensitizers and 683 non-sensitizers measured with different animal testing systems. Derek Nexus can be combined with Meteor Nexus to also assess the skin sensitization potential of likely metabolites.

The OECD QSAR Toolbox (OECD. The OECD QSAR Toolbox) offers the combination of several rule-based profilers (see section "Rule-based approaches") with read-across to find adequate analogs or build chemical categories. The OECD QSAR Toolbox provides experimental data on various endpoints for a large number of substances.

TIMES-SS (Dimitrov et al. 2005; Mekenyan et al. 2012; OASIS-LMC, TIMES model for skin sensitization prediction) is a hybrid model for the semi-quantitative prediction of the skin sensitization potency of substances. The predictor is part of the TIMES platform for toxicity prediction. TIMES includes a large collection of models for the prediction of human endpoints and metabolism. It also includes modules for the prediction of autoxidation and volatility, which can support the prediction of skin sensitization (Patlewicz, Kuseva, Mehmed, et al. 2014). TIMES-SS was developed based on 875 substances annotated with GPMT, LLNA, and human and animal data. The model combines a skin metabolism simulator with several local models for the assignment of three sensitization classes. Substances of interest are analyzed through 420 hierarchical ordered transformations (sorted by probability of occurrence) that link a source to a product structural fragment. The transformations account for abiotic reactions, covalent interaction with proteins and phase I and II metabolic reactions. Whenever a covalent interaction with a skin protein is predicted to occur, the compound is classified as either a strong or a weak sensitizer (depending on the triggered alert), or it is further analyzed by one of the several local 3D QSAR models that differentiate between non-, weak, and strong sensitizers. These 3D QSAR models take parameters such as the HOMO and LUMO energies, the HOMO-LUMO energy gap, molecular weight, electronegativity, hydrophobicity, and acceptor superdelocalizability as input (Mekenyan et al. 2004).

The AD of TIMES-SS is defined by the value range of several physicochemical properties and the structural and mechanistic domain covered by the training data. A representative test set of 40 REACH-relevant chemicals was selected from the European Inventory of Existing Commercial Chemical Substances (EINECS), taking into account commercial availability and structural diversity. The 40 compounds (16 sensitizers and 24 non-sensitizers) were evaluated in subsequent LLNA experiments. TIMES-SS correctly classified 30 compounds (9 sensitizers and 21 non-sensitizers) of the 40 compounds (Patlewicz et al. 2007; Roberts, Patlewicz, et al. 2007).

Several models relevant to skin sensitization prediction have also been implemented in the QSAR software CASE Ultra (Klopman 1992; Graham et al. 1996; Chakravarti et al. 2012; Saiakhov et al. 2013). These include models for the prediction of electrophilicity, protein binding (first key event of the skin sensitization AOP; trained on 194 compounds), the activation of the antioxidant response element (ARE) in keratinocytes (second key event of AOP; trained on 185 compounds), the activation of dendritic cells (third key event of AOP; trained on 189 compounds), LLNA outcomes (trained on 587 compounds) and ACD induction in humans and guinea pigs (trained on 1032 compounds). All of these models are purely of statistical nature and were derived with an algorithm based on the MultiCASE methodology (Klopman 1992). The algorithm generates a large number of structural fragments from sets of molecules and identifies fragments of statistical relevance for an endpoint of interest. Within this process, a hierarchical approach is applied to divide the training set into logical subsets. In contrast to the structural alerts used in most rule-based approaches, CASE Ultra not only encodes structural fragments related to skin sensitization ("biophores" or positive alerts) but also takes into account fragments that are identified as hindering skin sensitization ("biophobes" or deactivating alerts). Where feasible, local models are developed for each group of compounds sharing the same positive alert by stepwise linear regression incorporating, among others, log*P*, local charges, vapor pressure or presence or absence of modulating structural fragments as descriptors. The developers of CASE Ultra report 10-fold cross-validation accuracies of 67% to 87% for the different models relevant to skin sensitization prediction. The best

performance was obtained with a model for the prediction of human and guinea pig ACD. A tool to perform the read-across analysis is also provided with CASE Ultra.

A quantitative hybrid model for the prediction of the skin sensitization potency of compounds that combines expert knowledge with a linear QSAR approach was developed by Dearden et al. (2015). The hybrid model was developed based on a curated set of 204 known sensitizers annotated with measured EC3 values (non-sensitizers were not considered) from the data sets of Gerberick et al. (2005) and Kern et al. (2010). The compounds were assigned to one of seven different (pro-) reaction domains (i.e. acyl transfer, (pro-) Michael addition, (pro-) Schiff base, S_N2, and oxidation potential). Local linear QSAR models were derived for four (pro-) reaction domains for which sufficient data were available. From an initial set of 1600 descriptors (including log*P*, water solubility, molar refractivity, surface area descriptors, vapor pressure, Gasteiger charges, E-State descriptors, and fragment descriptors), up to six descriptors were selected by a wrapper method of stepwise multiple linear regression (MLR) for the different local models. Although the potency of Michael acceptors was found to be well described by reactivity and (hydrophobic) surface area, the potency of substances in pro-Michael, acyl transfer, and the combination of Schiff base and pro-Schiff domains was found to correlate with several descriptors representing hydrogen bonding. The potency of Schiff bases correlated with polarity and molecular flexibility; the potency of molecules undergoing S_N2 reactions increased with hydrophobicity and decreased with electron-donating ability. A range of the values of the descriptors covered by the training data is given for each local model as an indicator of whether a compound of interest is within the AD of the model. An R^2 of 0.95 was reported for a set of 37 compounds covering the same chemical space as the training data. The compounds had previously been used for descriptor selection but not for model training. The model is applicable only to skin sensitizers that can be assigned based on expert knowledge to one of the reaction domains for which a local model was retrieved.

Another approach combining a linear QSAR method with an expert-curated set of rules has been implemented in the CADRE-SS model for predicting the skin sensitization potency of compounds (Kostal and Voutchkova-Kostal 2016; ToxFix). The three-class categorical hybrid model consists of three modules describing different steps in the sensitization process. In the first module, the permeability coefficient ($\log K_p$) is calculated by Monte Carlo simulation. The second module uses a set of rules, encoded in a similar way as those implemented in Toxtree, to assign the most likely reaction domain. Compounds for which no reaction domain could be assigned are assumed to be non-sensitizers. Any compounds predicted as sensitizers are passed on to the third module, which calculates the chemical reactivity of compounds based on ground-state, site-specific, or global physicochemical and quantum mechanical descriptors, depending on the reaction domain assigned. For each reaction domain, a linear model was developed that takes the predictions of modules one (skin permeability) and three (chemical reactivity) as input. In addition, a rule-based approach was implemented to account for the qualitative sensitizing potential of metal salts. CADRE-SS was trained on a set of 384 chemicals annotated with LLNA data. Confidence levels for the individual predictions are derived from the range of the descriptors values observed for the training set. Tested on a set of 100 compounds annotated with human data, animal data or both, the model correctly assigned more than 90% of these compounds to one of the three GHS skin sensitization potency categories. The authors emphasize that, in contrast to other *in silico* tools for the prediction of skin sensitization potential, CADRE-SS was applicable to all compounds of this test set.

Very recently, Luechtefeld et al. (Luechtefeld, Marsh, et al. 2018) reported two binary classifiers of different complexity for the prediction of the skin sensitization potential of compounds. Both classifiers are based on the combination of a fingerprint similarity analysis with machine learning. The basic model generates 2D vectors that describe the similarities of each of the substances in the database to the closest sensitizing and non-sensitizing neighbors. These are analyzed by logistic regression in a second, supervised modeling step. The basic model obtained binary classification accuracies of 68% during leave-one-out cross-validation on a data set of 4783 compounds. The more complex model takes dependencies between 19 different endpoints into account (described by 74 D vectors) and uses a random forest algorithm for prediction. This model was tested with five-fold cross-validation during which it obtained an accuracy of 84% on a data set of 7670 compounds. Because of its ability to handle missing data, the more complex model is more widely applicable. Both models are available within the proprietary platform REACH*Across*™ (Luechtefeld, Rowlands, et al. 2018; UL).

An example of a consensus model for the discrimination of sensitizers and non-sensitizers is a weight of evidence approach developed by Ellison et al. (2010) that integrates results from the OECD QSAR Toolbox, Derek for Windows, the SMARTS pattern-based approach of Enoch et al. (Enoch, Madden, et al. 2008), and the CAESAR global model. The ability of the model to predict binary LLNA data was tested on 44 compounds that were published shortly before the consensus model was developed, for which reason the authors assume that these compounds are not part of the training data of any of the individual models. The consensus model produced conclusive results for 26 of the 44 test compounds, with 76% correct binary classifications. For 7 of the 44 test compounds, the binary predictions from all four models were in agreement and correct. For 18 compounds, the predictions were inconclusive.

Very recently, Alves et al. (Alves, Capuzzi, et al. 2018) reported a naive Bayes binary classification model for the prediction of the skin sensitization potential that integrates predictions from several in silico models provided within the same publication. The model was trained on 138 compounds annotated with human data and obtained a CCR of 89%. The model has been published as a free KNIME workflow (Naive Bayes Skin Sensitization Model v. 1.1).

Whereas hybrid *in silico* models integrate different computational methods and models, there are also approaches in existence that integrate results from one or several testing

methods (mostly *in vitro* or *in chemico* assays) and *in silico* approaches to establish the skin sensitization potential or potency of a compound of interest. These are discussed in the section "Computational methods used in combination with non-animal testing results."

The SkinSensDB platform (Wang et al. 2017; Tung et al. 2018) not only provides data relevant to skin sensitization (LLNA, human, DPRA/PPRA, KeratinoSens™/LuSens, and h-CLAT) but also includes functionality for the prediction of the skin sensitization potential of compounds based on the integration of these experimental data. For compounds of interest, values for missing experimental data are derived by a read-across approach. Using the stored or derived non-animal testing data as input, two different integrated testing strategies can be utilized for the binary prediction of the skin sensitization potential in the LLNA and in humans. Depending on the selected minimum similarity threshold acceptable for the read-across approach, accuracies of up to 81% and 89% were obtained for the prediction of LLNA (∼350 compounds) and human (∼50 compounds) outcomes, respectively.

Comparative analyses of the performance of computational models for skin sensitization prediction

Comparing the performance of computational methods for the prediction of skin sensitization is a non-trivial task. Most models are derived from different, often undisclosed or inaccessible data sets, which prohibit the design of an independent, representative and universal benchmark data set. Nevertheless, several studies have been published in recent years that aim to compare the performance and applicability of current *in silico* models. When considered with the necessary caution, these reports provide relevant insights on the scope and limitations of the individual models.

Teubner et al. (2013), for example, compared the performance of seven *in silico* models for the prediction of the skin sensitization potential: VEGA, CASE Ultra, TOPKAT, Toxtree, Derek Nexus, TIMES-SS, and the OECD QSAR Toolbox profilers for protein binding, direct peptide depletion activity, and keratinocyte gene expression. A data set of 100 compounds (55 non-sensitizers and 45 sensitizers) meeting a number of conditions was compiled for testing. The compounds were required to have reliable animal or human data on their skin sensitization potential available (i.e. adequate for GHS classification) and a molecular weight of less than 500 Da. Importantly, to avoid overlaps with the (often inaccessible) training data, only compounds that were not part of a high production volume program and had not been reported in scientific publications in the context of skin sensitization were considered. The tested models correctly classified 23% to 100% of all non-sensitizers and 55% to 100% of all sensitizers that were within the AD of the individual models (i.e. 16% to 100% of the test compounds). The mechanistic models obtained slightly higher success rates than purely statistical models. TIMES-SS turned out to be the most accurate model (100% correct classification) but was, however, applicable to just 16% of the tested compounds. One of the main conclusions drawn by the authors was that the tested models identify skin sensitizers with sufficient accuracy only if they bind to skin proteins without transformation or through a well-established transformation route and do not contain any rare structural features, functional groups, or atoms. None of the models for potency prediction yielded a good correlation with the experimentally determined GHS sensitization subcategories. Overall, the authors concluded that the existing models are not sufficiently accurate and broadly applicable for a widespread application in skin sensitization prediction.

Ellison et al. (2010) developed a weight of evidence approach using results from Derek for Windows, Enoch's SMARTS rules (Enoch, Madden, et al. 2008), OECD QSAR Toolbox and CAESAR global models, which we discuss in section "Hybrid *in silico* models." For an LLNA data set of 19 sensitizers and 25 non-sensitizers, the individual models yielded accuracies between 57% (CAESAR global models) and 70% (Derek for Windows). Because the testing data was published only recently before the actual tests were conducted, Ellison et al. assumed that there were no overlaps between the training and testing data.

Urbisch et al. (Urbisch, Honarvar, et al. 2016) tested the performance of the OECD QSAR Toolbox and TIMES-SS on a data set of 213 and 111 compounds annotated with LLNA and human skin sensitization data, respectively. Base accuracies for the OASIS profiler (71% and 70%) and the OECD profiler (69% and 67%) – both implemented in the OECD QSAR Toolbox – as well as for TIMES-SS (77% and 71%) were reported according to LLNA and human data, respectively. By a combination of the methods with modules or profilers for predicting metabolism and autoxidation, the accuracy of the best models based on the OECD QSAR Toolbox and TIMES-SS reached 84% and 94%, respectively, in predicting LLNA outcomes. Interestingly, the predictivity of these models was lower for human data, with models based on the OECD QSAR Toolbox and TIMES-SS reaching accuracies of up to 82% and 76%, respectively.

Verheyen et al. (2017) evaluated the ability of VEGA, CASE Ultra, Toxtree, Derek Nexus, and the skin sensitization profiler of the OECD QSAR Toolbox to discriminate skin sensitizers from non-sensitizers. A test set of 160 substances (82 sensitizers and 78 non-sensitizers) annotated with binary animal or human data on skin sensitization was compiled from public sources such as the Priority List of Hazardous Substances (published by the Agency for Toxic Substances and Disease Registry; ATSDR 2018) and the OECD's eChemPortal (OECD, eChemPortal). The major findings of this study are in agreement with those of Teubner et al. The models correctly classified between 48% and 78% of the test compounds to which the models were applicable, with rule-based models obtaining better results than the statistical models. Predictions could be obtained for 38% to 100% of the test compounds, with rule-based models having higher coverage rates than statistical models. The authors agreed with Teubner et al. that coverage and predictivity of current computational models are not satisfactory.

Rorije et al. (2013) studied and compared the ability of the h-CLAT and five *in silico* models (i.e. MultiCASE, Derek Nexus, TIMES-SS, TOPKAT, and the SMARTS rules of Enoch

et al.; Enoch, Madden, et al. 2008) to predict LLNA outcomes. An (incomplete) data matrix of 1045 compounds annotated with binary human, LLNA, GPMT, *in vitro*, and/or *in chemico* data served as the data basis for this analysis. The authors concluded that neither the performance of h-CLAT nor that of any of the *in silico* models is currently sufficient for standalone risk assessment (even though the *in vitro* model performed better than any of the *in silico* models and reached levels of predictivity close to those of GPMT by LLNA and vice versa). The performance indicators of the individual models indicate that a combination of non-animal testing approaches with *in silico* methods (see section "Computational methods used in combination with non-animal testing results") could yield models, which predict LLNA outcomes with an accuracy comparable with the predictivity of GPMT by LLNA and vice versa. For the combinations of these methods, however, the authors found lower than expected accuracies, which are likely related to dependencies between the individual *in silico* tools caused by overlaps of the knowledge bases or training data.

More recently, Fitzpatrick et al. (2018) evaluated the ability of VEGA, Derek Nexus, and TIMES-SS to predict binary LLNA and GPMT outcomes and compared their performance to the correlation between GPMT and LLNA outcomes. On a test set of 1295 unique compounds derived from eChemPortal, the overall accuracies obtained by VEGA, Derek Nexus, and TIMES-SS were 44%, 71%, and 67%. On a smaller test set of 515 unique compounds derived from the NICEATM LLNA data set, the models obtained accuracies between 57% and 61%. For both data sets, accuracies increased when only considering substances that are within the AD of the models but remained significantly lower than LLNA/GPMT predictivity, which was in the range of 80% to 85%. The low accuracy of VEGA was caused by a high sensitivity combined with a low specificity (which is in agreement with previous findings; Rorije et al. 2013). The authors detected 83 compounds for which all three models produced wrong predictions. This may be an indication of dependencies and bias shared among the models and is in the line with previous reports (Rorije et al. 2013), which found that the integration of several *in silico* models does not necessarily lead to the increase in accuracy that would be expected if all models were independent of each other.

Computational methods used in combination with non-animal testing results

When used on their own, modern theoretical and experimental AATs do not yet reach the regulatory acceptance requirements for skin sensitization risk assessment (Casati et al. 2018). Therefore, organizations such as the OECD encourage the development of integrated approaches to testing and assessment (IATAs), which can be described as human expert-led, non-formalized weight of evidence approaches amalgamating results obtained from different experimental models and theoretical approaches (e.g. OECD 2017b, 2017c).

IATAs are designed to be flexible and open for interpretation. New data are introduced in the decision process by an iterative procedure. Defined approaches (DAs) to testing and assessment, on the contrary, integrate information following fixed data interpretation procedures (DIP). As such they do not require or allow expert judgment but provide a defined algorithm that draws conclusions from defined input variables (Kleinstreuer et al. 2018). Two types of DAs are established in skin sensitization prediction: integrated testing strategies (ITSs) and sequential testing strategies (STSs) (Ezendam et al. 2016). Whereas ITSs combine information from multiple sources to reach a conclusion, STSs collect information in a stepwise manner involving interim decisions. Either type of DAs can be integrated into IATAs.

Obviously, the validity of the predictions made by IATAs and DAs depends on the quality of the individual inputs (Leontaridou et al. 2017). Alves et al. (Alves, Capuzzi, et al. 2018) recently showed that the binary outcomes of some of the assays most commonly used in IATAs (i.e. h-CLAT, DPRA, and KeratinoSens^{TM}) can be predicted with QSAR models with adequate accuracy. Similarly, Wijeyesakere et al. (2018) reported on a rule-based approach that yielded 89% correct predictions of binary DPRA outcomes of 162 substances. Both of these studies show that further improvement of these models could allow the use of calculated assay outcomes as input variables for IATAs and DAs in the future.

Approaches integrating existing data with experimental and/or computational approaches have been reviewed in several recent publications (Rovida et al. 2015; Ezendam et al. 2016; Jaworska 2016; Kleinstreuer et al. 2018). Here, we focus on those using computational methods to support data integration or that have been developed with the support of computational approaches.

An example of a computer-assisted IATA for skin sensitization risk assessment is the model developed by Patlewicz et al. (Patlewicz, Kuseva, Kesova, et al. 2014). The expert user is guided through a schematic workflow that collects the available experimental information on skin sensitization potential or potency (i.e. results from animal or non-animal experiments), skin irritation, genotoxicity, and physicochemical properties. As part of this process, the expert user is often requested to interpret the information collected as part of the workflow or to generate additional experimental data. Parts of the described workflow have been implemented as a software prototype called IATA-SS. IATA-SS obtained a binary classification accuracy of 74% on a test set of 100 compounds (consisting of 45 sensitizers and 55 non-sensitizers) that was previously compiled by Teubner et al. (2013).

Most DAs reported for skin sensitization prediction are ITSs. One of the earliest types of ITSs are the so-called "2-out-of-3" approaches (and variants thereof), which are based on majority voting. These include the ITS of Urbisch et al. (Urbisch, Honarvar, et al. 2016), which integrates results from either the DPRA, OECD QSAR Toolbox, or TIMES-SS with results from LuSens and h-CLAT. Interestingly, binary classification accuracies did not differ substantially for the different combinations of data. For example, the integration of the OECD QSAR Toolbox with LuSens and h-CLAT reproduced human data in 89% of all cases (covering 92 of the 111 compounds) and LLNA data in 91% of all cases (covering 141 of the 213 compounds). The integration of TIMES-SS with

LuSens and h-CLAT resulted in better accuracy (up to 100%) with coverages of only 13 to 20 compounds. Recently, "2-out-of-3" approaches integrating results from several non-animal testing methods have been challenged by "2-out-of-2" approaches, for which comparable accuracies were reached (Otsubo et al. 2017; Roberts and Patlewicz 2018). These findings are in line with the general concern that the added value of majority voting can be limited if the individual assays differ in their performance (Johansson and Gradin 2017) or provide redundant data in terms of the biological mechanisms assessed.

ITSs of higher complexity make use of statistical methods or machine learning, often with the aim to also allow a (semi-) quantitative prediction of skin sensitization potencies. For example, Jaworska et al. (2011, 2013, 2015) developed a Bayesian integrated testing strategy that takes calculated physico-chemical properties related to bioavailability, *in silico* potency predictions (performed with TIMES-SS) and results from *in vitro* (KeratinoSens™ and h-CLAT) and *in chemico* (DPRA) assays as input in a weight of evidence assessment. The latest published version of this model (ITS-3; Jaworska et al. 2015) was trained on LLNA data for a diverse set of 147 fragrances, preservatives, dyes, dye precursors, halogenated alkanes, and solvents. The model was able to predict the correct potency category (of four) for 53 of the 60 compounds of a test set. All wrong predictions were caused by misclassifications into a directly neighboring class. A recent study showed for version 2 of this ITS that the TIMES-SS (a commercial product) can be substituted by the structural alerts set implemented in the protein binding for skin sensitization profiler of the OECD QSAR toolbox (which is free software and the structural alerts are related to those covered by TIMES-SS) without a substantial decrease in prediction accuracy (Fitzpatrick and Patlewicz 2017).

Luechtefeld et al. (2015) trained dose-informed random forest/hidden Markov classification models on categorical LLNA data compiled for 145 substances (mainly derived from the Jaworska data set; see section "Data sets"). Up to 10 descriptors derived from different *in vitro* and *in chemico* assays, as well as descriptors calculated with Dragon and predictions of skin sensitization performed with TIMES-SS, served as input variables. For the best-performing models, the three most important input variables originated from *in vitro* and *in chemico* tests. The consideration of TIMES-SS predictions as input variables had no advantage over the use of Dragon descriptors. Accuracies for predicting the correct or neighboring potency category (of the four categories) were around 92% during stratified shuffle split cross-validation. The correct category was assigned to up to 65% of the compounds, depending on the input variables used.

Asturiol et al. (2016) derived classification trees from LLNA data (partly accompanied by human data) collected for 269 substances (using 80% of the data for training and 20% for testing). *In vitro* (KeratinoSens™ and h-CLAT) and *in chemico* data (DPRA), as well as molecular descriptors (calculated with Dragon) and *in silico* predictions (performed with Toxtree, OECD QSAR Toolbox, Derek Nexus, VEGA, TIMES-SS, and ADMET Predictor) were used as inputs. Interestingly, in contrast to the work of Luechtefeld et al., the prediction of protein binding by TIMES-SS turned out to be the most discriminating node for both of the two best-performing decision tree models, and neither of these models made use of *in vitro* or *in chemico* data. The best-performing binary classification model obtained an accuracy of 83% on the test set.

Zang et al. (2017) developed several classification models for the prediction of three categories of skin sensitization potency based on 94 compounds annotated with LLNA or 63 compounds annotated with human data. The models were derived using various machine learning algorithms (i.e. classification and regression tree, linear discriminant analysis, logistic regression, and support vector machine) following either a one-tiered approach (directly assigning one of three potency classes to a compound) or a two-tiered approach (first dividing compounds into non-sensitizers and sensitizers and then differentiating between weak and strong sensitizers). Six physicochemical properties (i.e. log*P*, water solubility, vapor pressure, melting point, boiling point, and molecular weight) and the results from three non-animal testing methods (i.e. DPRA, h-CLAT and KeratinoSens™) served as input for model building. The best-performing models resulted from a two-tiered SVM approach that took all input variables into account. These models obtained accuracies of 88% and 81% for the prediction of LLNA and human outcomes on two test sets of 63 and 24 compounds, respectively.

Strickland et al. (2017) used different subsets of the same types of input as Zang et al. (i.e. the three non-animal testing methods and six physicochemical properties) in combination with results obtained from the four protein binding profilers implemented in the OECD QSAR Toolbox to develop models based on logistic regression and a SVM for the prediction of the human skin sensitization potential. The models were trained on a set of 72 compounds for which results from DPRA, KeratinoSens™, h-CLAT, and LLNA, as well as human data are available. A random forest model was used for feature selection. Cysteine depletion measured by the DPRA was ranked as the most important feature, followed by other non-animal testing outcomes and the predictions from the OECD QSAR Toolbox. Of the six physicochemical properties, only the use of log*P* resulted in better performance of the linear regression and SVM model. The best models obtained an accuracy of 92% on a test set of 24 compounds.

Natsch et al. (2015) developed two models based on linear regression for the prediction of pEC3 values based on 244 compounds annotated with pEC3 values and tested on 68 compounds. One of the models used the reaction rate with peptides, Nrf2-induction, and cytotoxicity in KeratinoSens™ as the most distinguishing input variables, leading to an adjusted r^2 of 0.62. Interestingly, this model performed better for weak and moderate sensitizers than for strong or extreme sensitizers. The other model (a domain-based model) first groups compounds by their reaction mechanism as predicted by TIMES-SS and as measured by experimental adduct formation, and subsequently applies local regression for the different reaction domains. This model led to the better prediction for well-populated domains but resulted in poor predictivity for less populated ones. Most recently, Natsch et al. (2018) applied the domain-based model within an IATA to 22 existing and 7 new fragrance substances to derive EC3 values.

Within that application, the local model for aldehydes was slightly modified compared with the original publication. For 15 compounds with congruent human and LLNA data, an R^2 of 0.67 was reported for potency prediction. For each prediction on a substance of interest, uncertainty was assessed by applying the model to a structurally similar molecule with available LLNA data (from a database of more than 400 compounds) and comparing predicted and experimental potency. The no expected sensitization induction level (NESIL) was thereby derived from predicted EC3 values.

Hirota et al. (2017) utilized artificial neural networks for the prediction of LLNA EC3 values (binned into four potency classes) by integrating results obtained from the h-CLAT, DPRA, and KeratinoSens™ with predictions from Toxtree and TIMES. Several models taking into account different combinations of input variables were trained on 134 and tested on 28 compounds annotated with LLNA data. Models taking into account predictions from TIMES (accuracy 71%) or Toxtree alerts (accuracy 64%) obtained higher accuracies than the two models based solely on experimental data (43% and 50%).

ITSs may also be used to address the issue of a limited predictivity of LLNA data for human skin sensitization. For example, Alves et al. (Alves, Capuzzi, et al. 2016) combined LLNA outcomes with QSAR results to predict the skin sensitization potential in humans. Several binary QSAR models using radial basis function interpolation and self-consistent regression were developed for this purpose using a data set of 109 compounds annotated with human and LLNA data. A consensus model combining 10 of the best-performing QSAR models in each fold resulted in the correct classification of 71% of all compounds into skin sensitizers and non-sensitizers during five-fold cross-validation. In comparison, the LLNA obtained a CCR of only 63% on this data. When combining QSAR predictions with LLNA outcomes by only considering compounds for which both approaches produced concordant results, the CCR improved to 82% (five-fold cross-validation), at the cost of the applicability of the approach, which was reduced to 52% of the tested compounds.

Most DAs for skin sensitization prediction can be identified as ITS, but STSs are also in existence. For example, van der Veen et al. (2014) proposed an STS for the qualitative prediction of human skin sensitization potential. The three-tiered, independent Bayesian approach integrates results from *in silico* tools (MultiCASE, CAESAR, Derek Nexus, and OECD QSAR Toolbox) with those from *in chemico* and *in vitro* assays (peptide binding, gene signature, KeratinoSens™, and h-CLAT). The model was developed based on a set of 41 compounds annotated with human and LLNA data and designed to reflect various potency classes. In addition, compounds known to cause false-positive or false-negative results in LLNA (compared with human data) were included in the data set. Depending on the results of each tier, results from one to three non-animal testing methods were requested by the approach. Within the first tier, QSAR models were applied and, only if they resulted in an equivocal call, peptide binding was also tested. Depending on the result of this tier, either Keratinosens™ or gene signature was tested in the second tier of the approach. The third tier, comprising h-CLAT results, was only performed when the first and second tier were not in agreement. These interim decision steps not only aim for a reduction of experiments but to account for the predictive performance of the different included methods. For human data, the three-tiered approach obtained classification accuracies of 92% and higher. In contrast, LLNA data only predicted 78% of the human data correctly.

A further example of an STS is a binary classification model for skin sensitization potential based on a decision tree that integrates predictions from Derek Nexus with results from a maximum of two of the five *in chemico* and *in vitro* assays (i.e. DPRA, KeratinoSens™, LuSens, h-CLAT, U-SENS™) (Macmillan et al. 2016). As part of this STS, for any compound of interest that is within the AD of the assays, a two-tiered majority voting approach is applied that takes into account predictions from Derek Nexus and one or two assays. If the outcome of the first assay is in accordance with the predictions from Derek Nexus, no additional assay is used for majority voting to reduce the number of required experiments. In the case of discordant results, however, the second assay decides the overall result. If a substance is outside of the AD of both assays, only Derek Nexus is used for prediction. The STS was tested with 20 different combinations of *in vitro* and *in chemico* assays as input variables on a data set of 213 compounds annotated with LLNA and, where available, also human data. The models reached a classification accuracy of ~80% to 90%, depending on the assay used (median accuracy 85%). The authors note that the substances evaluated in this study are not independent of the models since 20% of them have been part of the Derek Nexus training set and an unknown portion of them is assumed to be included in the training data of the assays. Nevertheless, the authors state that removal of Derek Nexus training data from the test set did not significantly alter the final results.

Kleinstreuer et al. (2018) evaluated six DAs for the prediction of the skin sensitization potential and potency of compounds based on the Cosmetics Europe data set (see section "Data sets"). The best models reached binary classification accuracies of up to 83 and 80% for LLNA and human data, respectively (in comparison, prediction of human skin sensitization based on LLNA data was correct for only 68% of all tested compounds).

Outlook and conclusions

Today, a large number of *in silico* models for the prediction of the skin sensitization potential and potency of substances are in existence. The models are based on a variety of approaches, each with its own advantages and disadvantages. Currently, there is no single model or algorithm in existence that consistently outperforms all others.

Rule-based approaches, read-across, and linear statistical models score with good interpretability, and the latter may also provide new mechanistic insights. Machine learning approaches are generally more difficult to interpret or even have the character of a black box, but they are generally most suitable for modeling complex nonlinear relationships

such as those observed for most toxicity endpoints, including skin sensitization. Two major strategies that have been followed successfully to increase the prediction accuracy of models are the combination of different computational methods (hybrid models) and the amalgamation of theoretical methods and experimental data (DAs and IATAs). Both strategies have the potential to maximize accuracy and applicability by combining information from different sources. When working with these approaches and interpreting predictions, it is, however, important to carefully consider possible correlations between sources of information.

Molecular descriptors play a crucial role in the performance, applicability, and interpretability of models. Ideally, models should be based on small sets of physically meaningful descriptors to enable the interpretation of models and minimize the risk of overfitting. In the context of machine learning, in particular, feature selection algorithms are commonly utilized to select important features from large sets of descriptors. This can yield better performing models but generally at the cost of interpretability. Among models for the prediction of skin sensitization, a prevalence of descriptors associated with the toxicological endpoint on a mechanistic level is observed. These descriptors include structural fragments or alerts that correspond to the five established reaction domains, as well as descriptors linked to chemical reactivity (e.g. molecular orbital energies) or skin penetration (e.g. log*P* or molecular volume).

Whereas the existing molecular descriptors and modeling techniques have come of age the limited availability of reliable and relevant data remains a bottleneck for the development of more accurate and widely applicable *in silico* predictors of skin sensitization, particularly of potency.

Human data on skin sensitization remain extremely rare, are mostly NOAELs that are difficult to interpret and vary in quality. LLNA outcomes have been reported for a total of more than 1000 substances. These are mostly binary data; potency information is available in the public domain for only a subset of a few hundred substances. A recent study has shown that LLNA data populate regions in chemical space not covered by any other type of skin sensitization data (Alves, Capuzzi, et al. 2018). For these and other reasons, most *in silico* models are derived from collections of LLNA data. Although this increases the AD of models, it also caps predictivity of human health to that of animal experiments, which themselves are clearly limited (Alves, Capuzzi, et al. 2018; Hoffmann et al. 2018). High-quality data sets on skin sensitization, such as those compiled by Hoffmann et al. (2018) or Natsch et al. (2013), are available but small in size. They generally include EC3 values accompanied by information on non-animal testing results and human evidence. These data sets are not sufficiently large for model training but can be of high value as benchmark data sets for theoretical and experimental approaches alike.

Much of the measured data on skin sensitization is proprietary company data. Because of the pressing problem of a lack of data required to advance theoretical and experimental AATs alike, new avenues are being explored that could allow the distribution of proprietary data for model development without contravening the interests of their owners. These strategies include the use of machine learning algorithms on encrypted data (Luechtefeld, Rowlands, et al. 2018), as well as the allocation of data by so-called honest brokers. The latter has already resulted, for example, in the contribution of data from nine Lhasa member organizations to the evaluation of Derek Nexus (Chilton et al. 2018).

For computational models to be acceptable for use for regulatory purposes, they should comply with the guidelines for linear (Q)SAR models (OECD 2014b). Ideally, the data used for model building and testing should be fully disclosed to ensure reproducibility and allow a detailed understanding of the AD, the verification of data, and the testing of models with external data while excluding overlaps. In recent years, significant efforts have been made to develop *in silico* models for the prediction of skin sensitization that satisfy the requirements for use in a regulatory environment. A growing acceptance of these methods (together with other AATs) in risk assessment is observed. For example, in April 2018 the EPA released a draft for the acceptance of AATs for predicting skin sensitization, reasoning that substantial scientific evidence supports the use of these new methodologies (U.S. EPA 2018). In addition, the ECHA has recently promoted the use of AATs in REACH applications (ECHA 2017).

A major determinant for the acceptance of AATs for regulatory purposes is their validation with robust protocols complying with defined international standards. However, a substantial number of models, even of those published recently, are still not properly validated. All too often, only evaluation results from cross-validation are reported, or, in the worst case, from predictions on the training data.

Independent studies comparing the performance of models are an important cornerstone on the way to establishing *in silico* models and other AATs as a major pillar of risk assessment, but these studies are hindered by the scarcity of data available for testing and the often undisclosed training data. The development of well-characterized, high-quality data sets is therefore essential for the robust, comparative evaluation of theoretical and experimental models alike.

Just like any modern AAT, *in silico* models are not yet sufficiently reliable and broadly applicable to be used as a single prediction method for risk assessment. They are also not capable of predicting skin sensitization caused by mixtures, and few approaches are applicable to metals. However, despite all challenges, several studies have shown that *in silico* models have the capacity to outperform animal testing experiments, which have been accepted for regulatory purposes for decades. With the increasing availability of experimental data and advances in modeling techniques, computational methods and other AATs are expected to reach levels of accuracy and applicability that will make them a primary tool for risk assessment in the foreseeable future. In particular, integrated approaches combining *in vivo*, *in vitro*, *in chemico,* and *in silico* data hold the promise to evolve into powerful models for the prediction of the skin sensitization potential and potency of substances in humans with previously unmet accuracy and reach.

Acknowledgments

The authors thank Andreas Schepky from Beiersdorf AG Hamburg, Germany, and Christina de Bruyn Kops and Neann Mathai, both from the Center of Bioinformatics (ZBH) of the University of Hamburg, Germany, for discussion and proofreading of the manuscript. The authors also thank the three anonymous reviewers for the valuable feedback received, which helped the authors to improve this manuscript.

Declaration of interest

Beiersdorf AG is a personal care company based in Hamburg, Germany. Beiersdorf's consumer business focusses on skin care products with brands such as Nivea and Eucerin. The company has a strong background in the development of alternatives to animal testing for about 30 years, being involved in the development of e.g. the *in vitro* 3T3 NRU Photoxicity test and in KeratinoSens™ ring studies. Within Cosmetics Europe, the European trade association of cosmetics and personal care industry, Beiersdorf is contributing to cooperative efforts to foster the development and implementation of alternatives to animal testing.

HITeC e.V. is the Research and Technology Transfer Center of the Department of Informatics at the University of Hamburg. HITeC is a registered, nonprofit association, which is supported by members of the Department of Informatics at the University of Hamburg. The association is affiliated with the University of Hamburg.

AW is funded by Beiersdorf AG through HITeC e.V. and JKi is supported by the Bergen Research Foundation (BFS) [BFS2017TMT01]. BFS gives grants toward research and research supporting activities at the University of Bergen (UiB) and Haukeland University Hospital (HUS), and other Norwegian research institutions that cooperate with institutions in Bergen. The foundation also gives grants to support research at UiB and HUS at the interface between basic research and clinical research.

Neither JKü, AW, nor JKi have participated in legal proceedings related to the contents of the paper. The authors have sole responsibility for the writing and content of the paper.

ORCID

Anke Wilm http://orcid.org/0000-0003-2891-1407
Jochen Kühnl http://orcid.org/0000-0001-8421-9381
Johannes Kirchmair http://orcid.org/0000-0003-2667-5877

References

Adler S, Basketter D, Creton S, Pelkonen O, van Benthem J, Zuang V, Andersen KE, Angers-Loustau A, Aptula A, Bal-Price A. 2011. Alternative (non-animal) methods for cosmetics testing: current status and future prospects-2010. Arch Toxicol. 85:367–485.

Alves VM, Capuzzi SJ, Braga RC, Borba JVB, Silva AC, Luechtefeld T, Hartung T, Andrade CH, Muratov EN, Tropsha A. 2018. A perspective and a new integrated computational strategy for skin sensitization assessment. ACS Sustainable Chem Eng. 6:2845–2859.

Alves VM, Capuzzi SJ, Muratov E, Braga RC, Thornton T, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. 2016. QSAR models of human data can enrich or replace LLNA testing for human skin sensitization. Green Chem. 18:6501–6515.

Alves VM, Golbraikh A, Capuzzi SJ, Liu K, Lam WI, Korn DR, Pozefsky D, Andrade CH, Muratov EN, Tropsha A. 2018. Multi-Descriptor Read Across (MuDRA): a simple and transparent approach for developing accurate quantitative structure-activity relationship models. J Chem Inf Model. 58:1214–1223.

Alves V, Muratov E, Capuzzi S, Politi R, Low Y, Braga R, Zakharov AV, Sedykh A, Mokshyna E, Farag S, et al. 2016. Alarms about structural alerts. Green Chem. 18:4348–4360.

Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. 2015a. Predicting chemically-induced skin reactions. Part II: QSAR models of skin permeability and the relationships between skin permeability and skin sensitization. Toxicol Appl Pharmacol. 284:273–280.

Alves VM, Muratov E, Fourches D, Strickland J, Kleinstreuer N, Andrade CH, Tropsha A. 2015b. Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. Toxicol Appl Pharmacol. 284: 262–272.

Anderson SE, Siegel PD, Meade BJ. 2011. The LLNA: a brief review of recent advances and limitations. J Allergy. 2011:424203.

Api AM, Parakhia R, O'Brien D, Basketter DA. 2017. Fragrances categorized according to relative human skin sensitization potency. Dermatitis. 28:299–307.

Aptula AO, Roberts DW. 2006. Mechanistic applicability domains for non-animal-based prediction of toxicological end points: general principles and application to reactive toxicity. Chem Res Toxicol. 19:1097–1105.

Asturiol D, Casati S, Worth A. 2016. Consensus of classification trees for skin sensitisation hazard prediction. Toxicol In Vitro. 36:197–209.

ATSDR. 2018. Substance priority list. [accessed 2018 Aug 14]. https://www.atsdr.cdc.gov/SPL/.

Barratt MD, Basketter DA, Chamberlain M, Payne MP, Admans GD, Langowski JJ. 1994. Development of an expert system rulebase for identifying contact allergens. Toxicol In Vitro. 8:837–839.

Basketter DA, Alépée N, Ashikaga T, Barroso J, Gilmour N, Goebel C, Hibatallah J, Hoffmann S, Kern P, Martinozzi-Teissier S, et al. 2014. Categorization of chemicals according to their relative human skin sensitizing potency. Dermatitis. 25:11–21.

Basketter D, Clewell H, Kimber I, Rossi A, Blaauboer B, Burrier R, Daneshian M, Eskes C, Goldberg A, Hasiwa N, et al. 2012. A roadmap for the development of alternative (non-animal) methods for skin sensitization testing. In: A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing. ALTEX. 29:3–32.

Benfenati E, Manganaro A, Gini G. 2013. VEGA-QSAR: AI inside a platform for predictive toxicology. In: Baldoni M, Chesani F, Mello P, Montali M, editors. 2013. Proceedings of the workshop "Popularize Artificial Intelligence 2013"; December 5 2013; Turin, Italy. CEUR Workshop Proceedings Vol 1107.

Benigni R, Bossa C, Tcheremenskaia O. 2016. A data-based exploration of the adverse outcome pathway for skin sensitization points to the necessary requirements for its prediction with alternative methods. Regul Toxicol Pharmacol. 78:45–52.

Bergers LIJC, Reijnders CMA, van den Broek LJ, Spiekstra SW, de Gruijl TD, Weijers EM, Gibbs S. 2016. Immune-competent human skin disease models. Drug Discov Today. 21:1479–1488.

BIOVIA. QSAR, ADMET and predictive toxicology. [accessed 2018 Aug 22]. http://accelrys.com/products/collaborative-science/biovia-discovery-studio/qsar-admet-and-predictive-toxicology.html.

BIOVIA. 2017a. BIOVIA toxicity prediction model – skin sensitiser vs non sensitiser. (Q)SAR Model Reporting Format Database. [accessed 2018 Jan 16]. https://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-46-0042/document?media=application%2Fpdf.

BIOVIA. 2017b. BIOVIA toxicity prediction model – weak vs strong sensitiser. (Q)SAR Model Reporting Format Database. [accessed 2018 Jan 16]. https://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-46-0043/document?media=application%2Fpdf.

Braga RC, Alves VM, Muratov EN, Strickland J, Kleinstreuer N, Trospsha A, Andrade CH. 2017. Pred-Skin: a fast and reliable web application to assess skin sensitization effect of chemicals. J Chem Inf Model. 57: 1013–1017.

CAESAR. CAESAR project. [accessed 2018 Feb 6]. http://www.caesar-project.eu.

CAESAR. Skin sensitization model. [accessed 2018 Jan 25]. http://www.caesar-project.eu/index.php?page=results&section=endpoint&ne=2.

Canipa SJ, Chilton ML, Hemingway R, Macmillan DS, Myden A, Plante JP, Tennant RE, Vessey JD, Steger-Hartmann T, Gould J, et al. 2017. A quantitative in silico model for predicting skin sensitization using a nearest neighbours approach within expert-derived structure-activity alert spaces. J Appl Toxicol. 37:985–995.

Casati S, Aschberger K, Barroso J, Casey W, Delgado I, Kim TS, Kleinstreuer N, Kojima H, Lee JK, Lowit A, et al. 2018. Standardisation of defined approaches for skin sensitisation testing to support regulatory use and international adoption: position of the International Cooperation on Alternative Test Methods. Arch Toxicol. 92:611–617.

Chakravarti SK, Saiakhov RD, Klopman G. 2012. Optimizing predictive performance of CASE Ultra expert system models using the applicability domains of individual toxicity alerts. J Chem Inf Model. 52:2609–2618.

Chaudhry Q, Piclin N, Cotterill J, Pintore M, Price NR, Chrétien JR, Roncaglioni A. 2010. Global QSAR models of skin sensitisers for regulatory purposes. Chem Cent J. 4(Suppl 1):S5.

Chembench. Accelerating chemical genomics research. [accessed 2018 July 25]. https://chembench.mml.unc.edu/mudra/.

Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, et al. 2014. QSAR modeling: where have you been? Where are you going to? J Med Chem. 57:4977–5010.

Chilton ML, Macmillan DS, Steger-Hartmann T, Hillegass J, Bellion P, Vuorinen A, Etter S, Smith BPC, White A, Sterchele P, et al. 2018. Making reliable negative predictions of human skin sensitisation using an in silico fragmentation approach. Regul Toxicol Pharmacol. 95: 227–235.

Chipinda I, Hettick JM, Siegel PD. 2011. Haptenation: chemical reactivity and protein binding. J Allergy. 2011:839682.

CORAL. Free software for QSAR and nanoQSAR. [accessed 2018 Aug 13]. www.insilico.eu/coral.

Cottrez F, Boitel E, Auriault C, Aeby P, Groux H. 2015. Genes specifically modulated in sensitized skins allow the detection of sensitizers in a reconstructed human skin model. Development of the SENS-IS assay. Toxicol In Vitro. 29:787–802.

Cottrez F, Boitel E, Ourlin JC, Peiffer JL, Fabre I, Henaoui IS, Mari B, Vallauri A, Paquet A, Barbry P, et al. 2016. SENS-IS, a 3D reconstituted epidermis based model for quantifying chemical sensitization potency: reproducibility and predictivity results from an inter-laboratory study. Toxicol In Vitro. 32:248–260.

Dearden JC, Hewitt M, Roberts DW, Enoch SJ, Rowe PH, Przybylak KR, Vaughan-Williams GD, Smith ML, Pillai GG, Katritzky AR. 2015. Mechanism-Based QSAR Modeling of Skin Sensitization. Chem Res Toxicol. 28:1975–1986.

Dimitrov SD, Low LK, Patlewicz GY, Kern PS, Dimitrova GD, Comber MHI, Phillips RD, Niemela J, Bailey PT, Mekenyan OG. 2005. Skin sensitization: modeling based on skin metabolism simulation and formation of protein conjugates. Int J Toxicol. 24:189–204.

Dumont C, Barroso J, Matys I, Worth A, Casati S. 2016. Analysis of the Local Lymph Node Assay (LLNA) variability for assessing the prediction of skin sensitisation potential and potency of chemicals with non-animal approaches. Toxicol In Vitro. 34:220–228.

ECHA. Homepage. [accessed 2018 July 17]. https://echa.europa.eu/en.

ECHA. 2017. The use of alternatives to testing on animals for the REACH Regulation. [accessed 2018 Aug 27]. https://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2017_en.pdf.

Ellison CM, Madden JC, Judson P, Cronin MTD. 2010. Using in silico tools in a weight of evidence approach to aid toxicological assessment. Mol Inform. 29:97–110.

Enoch SJ, Cronin MTD, Schultz TW, Madden JC. 2008. Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via Michael addition. Chem Res Toxicol. 21:513–520.

Enoch SJ, Ellison CM, Schultz TW, Cronin MTD. 2011. A review of the electrophilic reaction chemistry involved in covalent protein binding relevant to toxicity. Crit Rev Toxicol. 41:783–802.

Enoch SJ, Madden JC, Cronin MTD. 2008. Identification of mechanisms of toxic action for skin sensitisation using a SMARTS pattern based approach. SAR QSAR Environ Res. 19:555–578.

Enoch SJ, Roberts DW. 2013. Predicting skin sensitization potency for Michael acceptors in the LLNA using quantum mechanics calculations. Chem Res Toxicol. 26:767–774.

Enslein K, Gombar VK, Blake BW, Maibach HI, Hostynek JJ, Sigman CC, Bagheri D. 1997. A quantitative structure-toxicity relationships model for the dermal sensitization guinea pig maximization assay. Food Chem Toxicol. 35:1091–1098.

EURL ECVAM. KeratinoSens assay for the testing of skin sensitizers [accessed 2018 Aug 29]. https://tsar.jrc.ec.europa.eu/test-method/tm2010-03.

EUR-Lex. 2009. Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. [accessed 2018 Jul 4]. https://eur-lex.europa.eu/eli/reg/2009/1223/oj.

EURL ECVAM. LuSens assay. [accessed 2018 Mar 28]. https://tsar.jrc.ec.europa.eu/test-method/tm2011-10.

Ezendam J, Braakhuis HM, Vandebriel RJ. 2016. State of the art in non-animal approaches for skin sensitization testing: from individual test methods towards testing strategies. Arch Toxicol. 90:2861–2883.

Fitzpatrick JM, Patlewicz G. 2017. Application of IATA – a case study in evaluating the global and local performance of a Bayesian network model for skin sensitization. SAR QSAR Environ Res. 28:297–310.

Fitzpatrick JM, Roberts DW, Patlewicz G. 2017a. What determines skin sensitization potency: myths, maybes and realities. The 500 molecular weight cut-off: an updated analysis. J Appl Toxicol. 37:105–116.

Fitzpatrick JM, Roberts DW, Patlewicz G. 2017b. Is skin penetration a determining factor in skin sensitization potential and potency? Refuting the notion of a LogKow threshold for skin sensitization. J Appl Toxicol. 37:117–127.

Fitzpatrick JM, Roberts DW, Patlewicz G. 2018. An evaluation of selected (Q)SARs/expert systems for predicting skin sensitisation potential. SAR QSAR Environ Res. 29:439–468.

Garner LA. 2004. Contact dermatitis to metals. Dermatol Ther. 17: 321–327.

Gerberick GF, Ryan CA, Kern PS, Schlatter H, Dearman RJ, Kimber I, Patlewicz GY, Basketter DA. 2005. Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods. Dermatitis. 16:157–202.

Gerner I, Barratt MD, Zinke S, Schlegel K, Schlede E. 2004. Development and prevalidation of a list of structure-activity relationship rules to be used in expert systems for prediction of the skin-sensitising properties of chemicals. Altern Lab Anim. 32:487–509.

Gleeson MP, Modi S, Bender A, Robinson RLM, Kirchmair J, Promkatkaew M, Hannongbua S, Glen RC. 2012. The challenges involved in modeling toxicity data in silico: a review. Curr Pharm Des. 18:1266–1291.

Goebel C, Aeby P, Ade N, Alépée N, Aptula A, Araki D, Dufour E, Gilmour N, Hibatallah J, Keller D, et al. 2012. Guiding principles for the implementation of non-animal safety assessment approaches for cosmetics: skin sensitisation. Regul Toxicol Pharmacol. 63:40–52.

Goebel C, Diepgen TL, Blömeke B, Gaspari AA, Schnuch A, Fuchs A, Schlotmann K, Krasteva M, Kimber I. 2018. Skin sensitization quantitative risk assessment for occupational exposure of hairdressers to hair dye ingredients. Regul Toxicol Pharmacol. 95:124–132.

Goebel C, Kosemund-Meynen K, Gargano EM, Politano V, von Bölcshazy G, Zupko K, Jaiswal N, Zhang J, Martin S, Neumann D, Rothe H. 2017. Non-animal skin sensitization safety assessments for cosmetic ingredients – What is possible today? Curr Opin Toxicol. 5:46–54.

Graham C, Gealy R, Macina OT, Karol MH, Rosenkranz HS. 1996. QSAR for allergic contact dermatitis. Quant Struct-Act Relat. 15:224–229.

Hartung T. 2013. Look back in anger – what clinical studies tell us about preclinical work. ALTEX. 30:275–291.

Hartung T. 2017. Opinion versus evidence for the need to move away from animal testing. ALTEX. 34:193–200.

Helgee EA, Carlsson L, Boyer S, Norinder U. 2010. Evaluation of quantitative structure-activity relationship modeling strategies: local and global models. J Chem Inf Model. 50:677–689.

Hennen J, Aeby P, Goebel C, Schettgen T, Oberli A, Kalmes M, Blömeke B. 2011. Cross talk between keratinocytes and dendritic cells: impact on the prediction of sensitization. Toxicol Sci. 123:501–510.

Heratizadeh A, Werfel T, Schubert S, Geier J. IVDK. 2018. Contact sensitization in dental technicians with occupational contact dermatitis. Data of the Information Network of Departments of Dermatology (IVDK) 2001-2015. Contact Dermatitis. 78:266–273.

Hirota M, Ashikaga T, Kouzuki H. 2017. Development of an artificial neural network model for risk assessment of skin sensitization using human cell line activation test, direct peptide reactivity assay, KeratinoSens™ and in silico structure alert parameter. J Appl Toxicol. 38:514–526.

Hoffmann S. 2015. LLNA variability: an essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. ALTEX. 32:379–383.
Hoffmann S, Kleinstreuer N, Alépée N, Allen D, Api AM, Ashikaga T, Clouet E, Cluzel M, Desprez B, Gellatly N, et al. 2018. Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database. Crit Rev Toxicol. 48:344–358.
ICCVAM. 2011. Test method evaluation report: usefulness and Limitations of the Murine Local Lymph Node Assay for potency categorization of chemicals causing allergic contact dermatitis in humans. [accessed 2018 Apr 16]. https://ntp.niehs.nih.gov/iccvam/docs/immunotox_docs/llna-pot/tmer.pdf.
ICCVAM. 2013. NICEATM Murine Local Lymph Node Assay (LLNA) Database. [accessed 2017 Nov 24]. https://ntp.niehs.nih.gov/pubhealth/evalatm/test-method-evaluations/immunotoxicity/nonanimal/index.html#NICEATM-Murine-Local-Lymph-Node-Assay-LLNA-Database.
Jaworska J. 2016. Integrated testing strategies for skin sensitization hazard and potency Assessment—State of the Art and Challenges. Cosmetics. 3:16.
Jaworska J, Dancik Y, Kern P, Gerberick F, Natsch A. 2013. Bayesian integrated testing strategy to assess skin sensitization potency: from theory to practice. J Appl Toxicol. 33:1353–1364.
Jaworska J, Harol A, Kern PS, Gerberick GF. 2011. Integrating non-animal test information into an adaptive testing strategy – skin sensitization proof of concept case. ALTEX. 28:211–225.
Jaworska JS, Natsch A, Ryan C, Strickland J, Ashikaga T, Miyazawa M. 2015. Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. Arch Toxicol. 89:2355–2383.
Johansson H, Gradin R, Forreryd A, Agemark M, Zeller K, Johansson A, Larne O, van Vliet E, Borrebaeck C, Lindstedt M. 2017. Evaluation of the GARD assay in a blind Cosmetics Europe study. ALTEX. 34:515–523.
Johansson H, Albrekt AS, Borrebaeck CA, Lindstedt M. 2013. The GARD assay for assessment of chemical skin sensitizers. Toxicol In Vitro. 27:1163–1169.
Johansson H, Gradin R. 2017. Skin sensitization: challenging the conventional thinking – a case against 2 out of 3 as integrated testing strategy. Toxicol Sci. 159:3–5.
Johansson H, Lindstedt M. 2014. Prediction of skin sensitizers using alternative methods to animal experimentation. Basic Clin Pharmacol Toxicol. 115:110–117.
Johansson H, Lindstedt M, Albrekt A-S, Borrebaeck CAK. 2011. A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. BMC Genomics 12:399.
Johansson H, Rydnert F, Kühnl J, Schepky A, Borrebaeck CAK, Lindstedt M. 2014. Genomic allergen rapid detection in-house validation–a proof of concept. Toxicol Sci. 139:362–370.
Karlberg AT, Bergström MA, Börje A, Luthman K, Nilsson JLG. 2008. Allergic contact dermatitis-formation, structural requirements, and reactivity of skin sensitizers. Chem Res Toxicol. 21:53–69.
Kazius J, McGuire R, Bursi R. 2005. Derivation and validation of toxicophores for mutagenicity prediction. J Med Chem. 48:312–320.
Kern PS, Gerberick GF, Ryan CA, Kimber I, Aptula A, Basketter DA. 2010. Local lymph node data for the evaluation of skin sensitization alternatives: a second compilation. Dermatitis. 21:8–32.
Kimber I, Basketter DA, Gerberick GF, Ryan CA, Dearman RJ. 2011. Chemical allergy: translating biology into hazard characterization. Toxicol Sci. 120Suppl 1:S238–S268.
Kimber I, Frank Gerberick G, Basketter DA. 2017. Quantitative risk assessment for skin sensitization: success or failure? Regul Toxicol Pharmacol. 83:104–108.
Kleinstreuer NC, Hoffmann S, Alépée N, Allen D, Ashikaga T, Casey W, Clouet E, Cluzel M, Desprez B, Gellatly N, et al. 2018. Non-animal methods to predict skin sensitization (II): an assessment of defined approaches *. Crit Rev Toxicol. 48:359–374.
Klopman G. 1992. MULTICASE 1. A hierarchical computer automated structure evaluation program. Quant Struct-Act Relat. 11:176–184.
Kostal J, Voutchkova-Kostal A. 2016. CADRE-SS, an in silico tool for predicting skin sensitization potential based on modeling of molecular interactions. Chem Res Toxicol. 29:58–64.
LabMol. Pred-Skin 2.0. [accessed 2018 Aug 22]. http://labmol.com.br/predskin/.
Langton K, Patlewicz GY, Long A, Marchant CA, Basketter DA. 2006. Structure-activity relationships for skin sensitization: recent improvements to Derek for Windows. Contact Dermatitis. 55:342–347.
Leontaridou M, Gabbert S, Van Ierland EC, Worth AP, Landsiedel R. 2016. Evaluation of non-animal methods for assessing skin sensitisation hazard: a Bayesian Value-of-Information analysis. Altern Lab Anim. 44:255–269.
Leontaridou M, Urbisch D, Kolle SN, Ott K, Mulliner DS, Gabbert S, Landsiedel R. 2017. The borderline range of toxicological methods: quantification and implications for evaluating precision. ALTEX. 34:525–538.
Lhasa Limited. Derek Nexus. [accessed 2018 Jan 26]. https://www.lhasalimited.org/products/derek-nexus.htm.
Luechtefeld T, Maertens A, McKim JM, Hartung T, Kleensang A, Sá-Rocha V. 2015. Probabilistic hazard assessment for skin sensitization potency by dose-response modeling using feature elimination instead of quantitative structure-activity relationships. J Appl Toxicol. 35:1361–1371.
Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu H, Hartung T. 2016. Analysis of publically available skin sensitization data from REACH registrations 2008-2014. ALTEX. 33:135–148.
Luechtefeld T, Marsh D, Rowlands C, Hartung T. 2018. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicol Sci. 165:198–212.
Luechtefeld T, Rowlands C, Hartung T. 2018. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. Toxicol Res. 7:732–744.
Lu J, Zheng M, Wang Y, Shen Q, Luo X, Jiang H, Chen K. 2011. Fragment-based prediction of skin sensitization using recursive partitioning. J Comput Aided Mol Des. 25:885–893.
Lushniak BD. 2004. Occupational contact dermatitis. Dermatol Ther. 17:272–277.
Macmillan DS, Canipa SJ, Chilton ML, Williams RV, Barber CG. 2016. Predicting skin sensitisation using a decision tree integrated testing strategy with an in silico model and in chemico/in vitro assays. Regul Toxicol Pharmacol. 76:30–38.
Martin SF, Esser PR, Weber FC, Jakob T, Freudenberg MA, Schmidt M, Goebeler M. 2011. Mechanisms of chemical-induced innate immunity in allergic contact dermatitis. Allergy. 66:1152–1163.
Mehling A, Eriksson T, Eltze T, Kolle S, Ramirez T, Teubner W, van Ravenzwaay B, Landsiedel R. 2012. Non-animal test methods for predicting skin sensitization potentials. Arch Toxicol. 86:1273–1295.
Mekenyan O, Dimitrov S, Pavlov T, Dimitrova G, Todorov M, Petkov P, Kotov S. 2012. Simulation of chemical metabolism for fate and hazard assessment. V. Mammalian hazard assessment. SAR QSAR Environ Res. 23:553–606.
Mekenyan O, Dimitrov S, Pavlov T, Veith G. 2004. A systematic approach to simulating metabolism in computational toxicology. I. The TIMES heuristic modelling framework. Curr Pharm Des. 10:1273–1293.
MultiCASE. CASE Ultra: QSAR expert system. [accessed 2018 Aug 22]. http://www.multicase.com/case-ultra.
Muratov EN, Artemenko AG, Varlamova EV, Polischuk PG, Lozitsky VP, Fedchuk AS, Lozitska RL, Gridina TL, Koroleva LS, Sil'nikov VN, et al. 2010. Per aspera ad astra: application of Simplex QSAR approach in antiviral research. Future Med Chem. 2:1205–1226.
Naive Bayes Skin Sensitization Model v. 1.1. [accessed 2018 Aug 16]. https://figshare.com/articles/Naive_Bayes_Skin_Sensitization_Model/5758644.
Natsch A, Emter R, Gfeller H, Haupt T, Ellis G. 2015. Predicting skin sensitizer potency based on in vitro data from KeratinoSens and kinetic peptide binding: global versus domain-based assessment. Toxicol Sci. 143:319–332.
Natsch A, Emter R, Haupt T, Ellis G. 2018. Deriving a no expected sensitization induction level for fragrance ingredients without animal

testing: an integrated approach applied to specific case studies. Toxicol Sci. 165:170–185.
Natsch A, Ryan CA, Foertsch L, Emter R, Jaworska J, Gerberick F, Kern P. 2013. A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. J Appl Toxicol. 33:1337–1352.
Nendza M, Gabbert S, Kühne R, Lombardo A, Roncaglioni A, Benfenati E, Benigni R, Bossa C, Strempel S, Scheringer M, et al. 2013. A comparative survey of chemistry-driven in silico methods to identify hazardous substances under REACH. Regul Toxicol Pharmacol. 66:301–314.
OASIS-LMC. TIMES model for skin sensitization prediction. [accessed 2018 Jan 30]. http://oasis-lmc.org/products/models/human-health-endpoints/skin-sensitization.aspx.
OASIS-LMC. TIMES-SS Software. [accessed 2018 Jan 30]. http://oasis-lmc.org/products/software/times.aspx.
OCHEM. Online chemical modeling environment. [accessed 2018 Mar 28]. https://ochem.eu/home/show.do.
OECD. eChemPortal. [accessed 2018 July 17]. https://www.echemportal.org/.
OECD. The OECD QSAR Toolbox. [accessed 2018 July 4]. http://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm.
OECD. 2012. The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins. [accessed 2018 Apr 17]. http://www.oecd.org/env/the-adverse-outcome-pathway-for-skin-sensitisation-initiated-by-covalent-binding-to-proteins-9789264221444-en.htm.
OECD. 2014a. OECD series on testing and assessment. Guidance on grouping of chemicals, second edition. [accessed 2018 Apr 17]. https://www.oecd-ilibrary.org/environment/guidance-on-grouping-of-chemicals-second-edition_9789264274679-en.
OECD. 2014b. OECD series on testing and assessment. Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. [accessed 2018 Apr 17]. http://www.oecd.org/env/guidance-document-on-the-validation-of-quantitative-structure-activity-relationship-q-sar-models-9789264085442-en.htm.
OECD. 2015a. Test No. 442C: In chemico skin sensitisation. [accessed 2018 Apr 17] http://www.oecd.org/env/test-no-442c-in-chemico-skin-sensitisation-9789264229709-en.htm.
OECD. 2015b. Test No. 442D: In vitro skin sensitisation. [accessed 2018 Apr 17]. http://www.oecd.org/env/test-no-442d-in-vitro-skin-sensitisation-9789264229822-en.htm.
OECD. 2017a. Test No. 442E: In vitro skin sensitisation. [accessed 2018 Apr 17]. http://www.oecd.org/env/test-no-442e-in-vitro-skin-sensitisation-9789264264359-en.htm.
OECD. 2017b. Guidance document on the reporting of defined approaches and individual information sources to be used within integrated approaches to testing and assessment (IATA) for skin sensitisation. [accessed 2018 July 4]. https://www.oecd-ilibrary.org/docserver/9789264274822-en.pdf.
OECD. 2017c. Guidance document on the reporting of defined approaches to be used within integrated approaches to testing and assessment. [accessed 2018 July 4]. https://www.oecd-ilibrary.org/docserver/9789264274822-en.pdf.
Otsubo Y, Nishijo T, Miyazawa M, Saito K, Mizumachi H, Sakaguchi H. 2017. Binary test battery with KeratinoSens™ and h-CLAT as part of a bottom-up approach for skin sensitization hazard prediction . Regul Toxicol Pharmacol. 88:118–124.
Ouyang Q, Wang L, Mu Y, Xie X-Q. 2014. Modeling skin sensitization potential of mechanistically hard-to-be-classified aniline and phenol compounds with quantum mechanistic properties. BMC Pharmacol Toxicol. 15:76.
Patlewicz G, Aptula AO, Roberts DW, Uriarte E. 2008. A minireview of available skin sensitization (Q)SARs/expert systems. QSAR Comb Sci. 27:60–76.
Patlewicz G, Ball N, Booth ED, Hulzebos E, Zvinavashe E, Hennes C. 2013. Use of category approaches, read-across and (Q)SAR: general considerations. Regul Toxicol Pharmacol. 67:1–12.
Patlewicz G, Casati S, Basketter DA, Asturiol D, Roberts DW, Lepoittevin JP, Worth AP, Aschberger K. 2016. Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? Regul Toxicol Pharmacol. 82:147–155.
Patlewicz G, Dimitrov SD, Low LK, Kern PS, Dimitrova GD, Comber MIH, Aptula AO, Phillips RD, Niemelä J, Madsen C, et al. 2007. TIMES-SS – A promising tool for the assessment of skin sensitization hazard. A characterization with respect to the OECD validation principles for (Q)SARs and an external evaluation for predictivity. Regul Toxicol Pharmacol. 48:225–239.
Patlewicz G, Helman G, Pradeep P, Shah I. 2017. Navigating through the minefield of read-across tools: a review of in silico tools for grouping. Comput Toxicol. 3:1–18.
Patlewicz G, Kuseva C, Kesova A, Popova I, Zhechev T, Pavlov T, Roberts DW, Mekenyan O. 2014. Towards AOP application-implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization. Regul Toxicol Pharmacol. 69:529–545.
Patlewicz G, Kuseva C, Mehmed A, Popova Y, Dimitrova G, Ellis G, Hunziker R, Kern P, Low L, Ringeissen S, et al. 2014. TIMES-SS-recent refinements resulting from an industrial skin sensitisation consortium. SAR QSAR Environ Res. 25:367–391.
Patlewicz G, Worth A. 2008. Review of Data Sources, QSARs and Integrated Testing Strategies for Skin Sensitisation. [accessed 2018 June 17]. https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/doc/EUR_23225_EN.pdf.
Payne MP, Walsh PT. 1994. Structure-activity relationships for skin sensitization potential: development of structural alerts for use in knowledge-based toxicity prediction systems. J Chem Inf Comput Sci. 34:154–161.
Politano VT, Api AM. 2008. The Research Institute for Fragrance Materials' human repeated insult patch test protocol. Regul Toxicol Pharmacol. 52:35–38.
Promkatkaew M, Gleeson D, Hannongbua S, Gleeson PM. 2014. Skin sensitization prediction using quantum chemical calculations: a theoretical model for the SNAr domain. Chem Res Toxicol. 27:51–60.
Raunio H. 2011. In silico toxicology - non-testing methods. Front Pharmacol. 2:33.
Reisinger K, Hoffmann S, Alépée N, Ashikaga T, Barroso J, Elcombe C, Gellatly N, Galbiati V, Gibbs S, Groux H, et al. 2015. Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. Toxicol In Vitro. 29:259–270.
Reuter H, Spieker J, Gerlach S, Engels U, Pape W, Kolbe L, Schmucker R, Wenck H, Diembeck W, Wittern K-P, et al. 2011. In vitro detection of contact allergens: development of an optimized protocol using human peripheral blood monocyte-derived dendritic cells. Toxicol In Vitro. 25:315–323.
Roberts DW. 2013. Allergic contact dermatitis: is the reactive chemistry of skin sensitizers the whole story? A response. Contact Dermatitis. 68:245–249.
Roberts DW, Aptula A, Api AM. 2017. Structure-potency relationships for epoxides in allergic contact dermatitis. Chem Res Toxicol. 30:524–531.
Roberts DW, Aptula AO. 2014. Electrophilic reactivity and skin sensitization potency of SNAr electrophiles. Chem Res Toxicol. 27:240–246.
Roberts DW, Aptula AO, Patlewicz G. 2007. Electrophilic chemistry related to skin sensitization. Reaction mechanistic applicability domain classification for a published data set of 106 chemicals tested in the mouse local lymph node assay. Chem Res Toxicol. 20:44–60.
Roberts DW, Aptula AO, Patlewicz GY. 2011. Chemistry-based risk assessment for skin sensitization: quantitative mechanistic modeling for the SNAr domain. Chem Res Toxicol. 24:1003–1011.
Roberts DW, Mekenyan OG, Dimitrov SD, Dimitrova GD. 2013. What determines skin sensitization potency-myths, maybes and realities. Part 1. The 500 molecular weight cut-off. Contact Dermatitis. 68:32–41.
Roberts DW, Natsch A. 2009. High throughput kinetic profiling approach for covalent binding to peptides: application to skin sensitization potency of Michael acceptor electrophiles. Chem Res Toxicol. 22:592–603.
Roberts DW, Patlewicz G. 2018. Non-animal assessment of skin sensitization hazard: Is an integrated testing strategy needed, and if so what should be integrated? J Appl Toxicol. 38:41–50.
Roberts DW, Patlewicz G, Dimitrov SD, Low LK, Aptula AO, Kern PS, Dimitrova GD, Comber MIH, Phillips RD, Niemelä J, et al. 2007.

TIMES-SS – a mechanistic evaluation of an external validation study using reaction chemistry principles. Chem Res Toxicol. 20:1321–1330.
Roberts DW, Schultz TW, Api AM. 2017. Skin sensitization QMM for HRIPT NOEL data: aldehyde Schiff-base domain. Chem Res Toxicol. 30: 1309–1316.
Roberts DW, Williams DL. 1982. The derivation of quantitative correlations between skin sensitisation and physio-chemical parameters for alkylating agents, and their application to experimental data for sultones. J Theor Biol. 99:807–825.
Rorije E, Aldenberg T, Buist H, Kroese D, Schüürmann G. 2013. The OSIRIS weight of evidence approach: ITS for skin sensitisation. Regul Toxicol Pharmacol. 67:146–156.
Rovida C, Alépée N, Api AM, Basketter DA, Bois FY, Caloni F, Corsini E, Daneshian M, Eskes C, Ezendam J, et al. 2015. Integrated testing strategies (ITS) for safety assessment. ALTEX. 32:25–40.
Russell WMS, Burch RL. 1959. The principles of humane experimental technique. Michigan: Universities Federation for Animal Welfare.
Saiakhov R, Chakravarti S, Klopman G. 2013. Effectiveness of CASE Ultra expert system in evaluating adverse effects of drugs. Mol Inform. 32: 87–97.
Schmidt M, Raghavan B, Müller V, Vogl T, Fejer G, Tchaptchet S, Keck S, Kalis C, Nielsen PJ, Galanos C, et al. 2010. Crucial role for human Toll-like receptor 4 in the development of contact allergy to nickel. Nat Immunol. 11:814–819.
Schultz TW, Amcoff P, Berggren E, Gautier F, Klaric M, Knight DJ, Mahony C, Schwarz M, White A, Cronin MT. 2015. A strategy for structuring and reporting a read-across prediction of toxicity. Regul Toxicol Pharmacol. 72:586–601.
SkinSensDB. [accessed 2018 Mar 29]. http://cwtung.kmu.edu.tw/skinsensdb/.
Steiling W. 2016. Safety evaluation of cosmetic ingredients regarding their skin sensitization potential. Cosmet Toiletries. 3:14.
Strickland J, Zang Q, Paris M, Lehmann DM, Allen D, Choksi N, Matheson J, Jacobs A, Casey W, Kleinstreuer N. 2017. Multivariate models for prediction of human skin sensitization hazard. J Appl Toxicol. 37: 347–360.
Sushko I, Novotarskyi S, Körner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, et al. 2011. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J Comput Aided Mol Des. 25:533–554.
Sushko I, Salmina E, Potemkin VA, Poda G, Tetko IV. 2012. ToxAlerts: a web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. J Chem Inf Model. 52:2310–2316.
Talete S.r.l. Dragon. [accessed 2018 Apr 16]. http://www.talete.mi.it/products/dragon_description.htm.
Teubner W, Mehling A, Schuster PX, Guth K, Worth A, Burton J, van Ravenzwaay B, Landsiedel R. 2013. Computer models versus reality: how well do in silico models currently predict the sensitization potential of a substance. Regul Toxicol Pharmacol. 67:468–485.
Thyssen JP, Giménez-Arnau E, Lepoittevin JP, Menné T, Boman A, Schnuch A. 2012. The critical review of methodologies and approaches to assess the inherent skin sensitization potential (skin allergies) of chemicals. Part I. Contact Dermatitis. 66: 11–24.
Thyssen JP, Linneberg A, Menné T, Johansen JD. 2007. The epidemiology of contact allergy in the general population-prevalence and main findings. Contact Dermatitis. 57:287–299.
Todeschini R, Consonni V, Mannhold R. 2009. Molecular descriptors for chemoinformatics: Volume I: Alphabetical listing/Volume II: Appendices, references. Weinheim: Wiley-VCH.
Toropova AP, Toropov AA. 2017. Hybrid optimal descriptors as a tool to predict skin sensitization in accordance to OECD principles. Toxicol Lett. 275:57–66.
ToxFix. CADRE-SS skin sensitization model. [accessed 2018 Jan 30]. http://toxfix.com/skin-sensitization.php.
Toxtree. [accessed 2018 Apr 16]. http://toxtree.sourceforge.net/.
Tung CW, Wang CC, Wang SS. 2018. Mechanism-informed read-across assessment of skin sensitizers based on SkinSensDB. Regul Toxicol Pharmacol. 94:276–282.
UL. REACHAcross™ – satisfy REACH regulations and ECHA submissions with automated read-across. [accessed 2018 July 7]. https://www.ulreachacross.com/index.html.
Urbisch D, Becker M, Honarvar N, Kolle SN, Mehling A, Teubner W, Wareing B, Landsiedel R. 2016. Assessment of pre- and pro-haptens using nonanimal test methods for skin sensitization. Chem Res Toxicol. 29:901–913.
Urbisch D, Honarvar N, Kolle SN, Mehling A, Ramirez T, Teubner W, Landsiedel R. 2016. Peptide reactivity associated with skin sensitization: the QSAR Toolbox and TIMES compared to the DPRA. Toxicol in Vitro. 34:194–203.
Urbisch D, Mehling A, Guth K, Ramirez T, Honarvar N, Kolle S, Landsiedel R, Jaworska J, Kern PS, Gerberick F, et al. 2015. Assessing skin sensitization hazard in mice and men using non-animal test methods. Regul Toxicol Pharmacol. 71:337–351.
U.S. EPA. 2018 Interim science policy: Use of alternative approaches for skin sensitization as a replacement for laboratory animal testing. [accessed 2018 Aug 30]. https://www.regulations.gov/document?D=EPA-HQ-OPP-2016-0093-0090.
van der Veen JW, Rorije E, Emter R, Natsch A, van Loveren H, Ezendam J. 2014. Evaluating the performance of integrated approaches for hazard identification of skin sensitizing chemicals. Regul Toxicol Pharmacol. 69:371–379.
VEGA HUB. VEGA. [accessed 2018 Mar 28]. https://www.vegahub.eu/portfolio-item/vega-qsar/.
Verheyen GR, Braeken E, Van Deun K, Van Miert S. 2017. Evaluation of in silico tools to predict the skin sensitization potential of chemicals. SAR QSAR Environ Res. 28:59–73.
Wang CC, Lin YC, Wang SS, Shih C, Lin YH, Tung CW. 2017. SkinSensDB: a curated database for skin sensitization assays. J Cheminform. 9:5.
Warne MA, Nicholson JK, Lindon JC, Guiney PD, Gartland KPR. 2009. A QSAR investigation of dermal and respiratory chemical sensitizers based on computational chemistry properties. SAR QSAR. Environ Res. 20:429–451.
Wijeyesakere SJ, Wilson DM, Settivari R, Auernhammer TR, Parks AK, Sue Marty M. 2018. Development of a profiler for facile chemical reactivity using the open-source Konstanz information miner. Appl In Vitro Toxicol. 4:202–213.
Williams RV, Amberg A, Brigo A, Coquin L, Giddings A, Glowienke S, Greene N, Jolly R, Kemper R, O'Leary-Steele C, et al. 2016. It's difficult, but important, to make negative predictions. Regul Toxicol Pharmacol. 76:79–86.
Winkler GC, Perino C, Araya SH, Bechter R, Kuster M, Lovsin Barle E. 2015. Classification of dermal sensitizers in pharmaceutical manufacturing. Regul Toxicol Pharmacol. 72:501–505.
Wondrousch D, Böhme A, Thaens D, Ost N, Schüürmann G. 2010. Local electrophilicity predicts the toxicity-relevant reactivity of Michael acceptors. J Phys Chem Lett. 1:1605–1610.
Yuan H, Huang J, Cao C. 2009. Prediction of skin sensitization with a particle swarm optimized support vector machine. Int J Mol Sci. 10: 3237–3254.
Zang Q, Paris M, Lehmann DM, Bell S, Kleinstreuer N, Allen D, Matheson J, Jacobs A, Casey W, Strickland J. 2017. Prediction of skin sensitization potency using machine learning approaches. J Appl Toxicol. 37: 792–805.

3.2 (Q)SAR modeling and machine learning

Computational methods as discussed in our review article [P1] bear the potential to provide fast and reliable predictions on skin sensitization potential and potency without the need to conduct time consuming and expensive experiments. Especially among the most recently reviewed approaches, the application of ML algorithms in (Q)SAR approaches has proven a promising route within the field: ML allows for high predictivity and applicability, can be combined with automated measures of reliability, be easily updated when new data appear, and be made available within intuitive programs or web applications that do not require expert knowledge from the user. In the present chapter, we introduce basic concepts of (Q)SAR and ML modeling and illustrate their methodological background, like the computational encoding of molecules and the definition of the applicability domain (AD) of the models. One popular way of mathematically defining the AD of ML models (in contrast to other subjective definitions), is to embed ML models in a conformal prediction (CP) framework, which ensures the reliability of a model at a defined confidence level. The basic concept of CP will also be introduced in section 3.2.4 of the present chapter.

3.2.1 (Q)SAR modeling

(Q)SAR approaches are based on the assumption that the biological activity of a substance is related to its molecular structure and, hence, to the molecular descriptors representing it [44]. Early (Q)SAR models tried to leverage the relationship between a small number of preselected chemical descriptors and the observed biological activity by applying simple regression or classification methods. Due to their relative simplicity, these (Q)SAR models were usually only applicable to a small group of chemicals. The earliest example of such a model in the field of skin sensitization prediction was developed by Roberts and Williams in the 1980s [45]: A linear equation derives the relative alkylation index (RAI) (which serves as a quantitative measure of skin sensitization potential) of sultones from only three chemical descriptors. (More details on the RAI concept and the models derived from it can be found in section "Chemical class or mechanism-based models" of our review article [P1].) Until today, the RAI concept has successfully been applied to a variety of small classes of similar or similar reacting molecules. However, these models have never achieved a degree of generalization that allows their application without expert knowledge on the potential reaction pathways of a molecule of interest.

To circumvent the limitations given by linear modeling, current (Q)SAR approaches make use of a variety of advanced ML algorithms. These algorithms are capable of capturing also complex (non-linear) relationships between molecular descriptors and biological endpoints. Moreover, automated feature selection

methods prior to or integrated in the ML algorithm allow for the unbiased selection of input features from large sets of descriptors (see subsection 3.3.3).

3.2.2 Machine learning approaches

ML approaches are computational algorithms that automatically learn from available training data in order to transfer the findings to new, unseen data. ML approaches are thereby also capable of capturing relationships within the data that are too complex for human detection. At the same time, when combined with insufficient data, complex ML models are prone to the danger of overfitting, which occurs when the model learns even from the noise of the training data and can therefore not generalize to unseen data.

Depending on the absence or presence of a target variable, ML approaches can be divided into unsupervised and supervised ML algorithms, respectively. Unsupervised ML models are trained on unlabeled data (i.e. data without a target variable) and aim to detect patterns that best separate or group the data. A common application of unsupervised ML is data clustering, which groups similar samples within each cluster. Common unsupervised clustering algorithms are k-means [46] , which separates the data based on the distance of each sample to the cluster centers, or hierarchical clustering [47], which iteratively merges similar objects into clusters until a single cluster remains and the hierarchy of the clusters can be observed in a dendrogram. Unsupervised ML algorithms are also successfully applied for dimensionality reduction of unlabeled data. This allows, for example, visualization of high dimensional data in a two dimensional space while still capturing a large amount of information. Common examples for these approaches are uniform manifold approximation and projection (UMAP) [48] and principal component analysis (PCA) [49]. Within the UMAP approach, data are projected into a lower dimensional space while maintaining the pairwise distance between samples. Within PCA, the multidimensional descriptor space is transposed into a new basis of orthogonal vectors by maximizing the explained variance of the first components (see Figure 3.1). In the case of skin sensitization prediction, unsupervised ML algorithms offer the advantage that they can be applied to data for which no skin sensitization data are available. This allows, for example, the comparison between the LLNA data and common reference data sets containing approved drugs, cosmetics, or pesticides in order to analyze the chemical space coverage of the available data. Because unsupervised ML algorithms are not biased by the experimentally assigned class labels, they might also be utilized to detect potential outliers, whose experimental class label might be questionable.

In contrast to unsupervised ML algorithms, supervised ML algorithms map each sample to a specific label or target variable. This information allows the extraction of patterns in existing data that can be used to predict the target

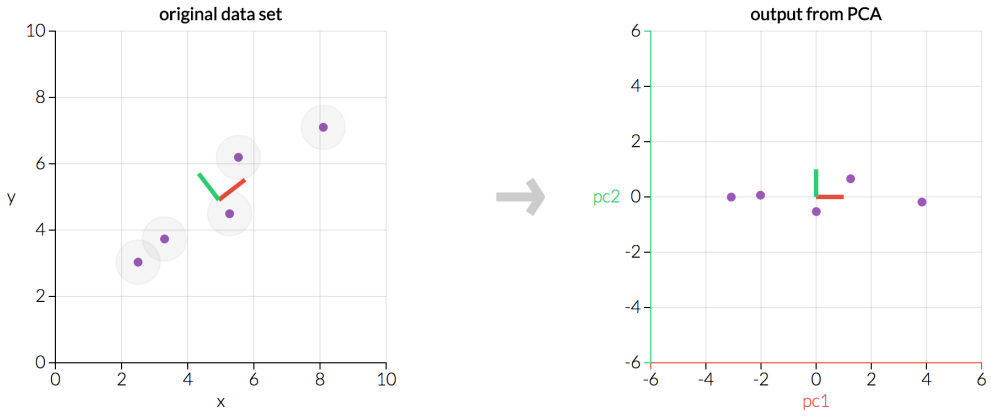


Figure 3.1: Example of a PCA of a two dimensional data set [50]: With the PCA the data set can be transposed into one dimension with only small information loss.

variable of unseen data. Depending on the modeling task (classification or regression), the available data and the descriptor space, the most favorable supervised ML algorithms may be different. Commonly applied supervised ML algorithms are linear regression, decision trees, random forest (RF) [51], k-nearest neighbor (KNN) [52] or support vector machine (SVM) [53] algorithms, among which linear regression is one of the simplest algorithms. It describes the target variable as a linear combination of the descriptors. Because of its simplicity, it is easy to interpret but fails in describing non-linear relations. Moreover, it is prone to overfitting when the complexity of the model (i.e. the number of descriptors with non-zero coefficients) is too high in comparison to the number of available training instances. Another supervised ML algorithm with low complexity is the KNN algorithm. It is based on the assumption that similar instances will also have similar target variables. This algorithm predicts the class label of an instance of interest based on the most common label among its k nearest neighbors (see Figure 3.2). The number k of considered neighbors is a freely defined parameter and the neighbors of each instance are identified based on their distance (e.g. the Euclidean distance) in the descriptor space. KNN models are highly sensitive to the number and characteristics of the descriptors, and careful feature selection and scaling may be therefore needed prior to their application. Irrelevant features may induce high amounts of noise in the prediction and descriptors with a larger absolute span would have more influence on the prediction than others. Special care has to be also undertaken for unbalanced data, since otherwise the minority class will be underrepresented in the descriptor space and hence have lower likelihood of being predicted. Visual investigation of the neighboring instances considered for a decision can assist human interpretation of the KNN model and its predictions. Another basic ML algorithm is the decision tree: In a decision tree, the label of a

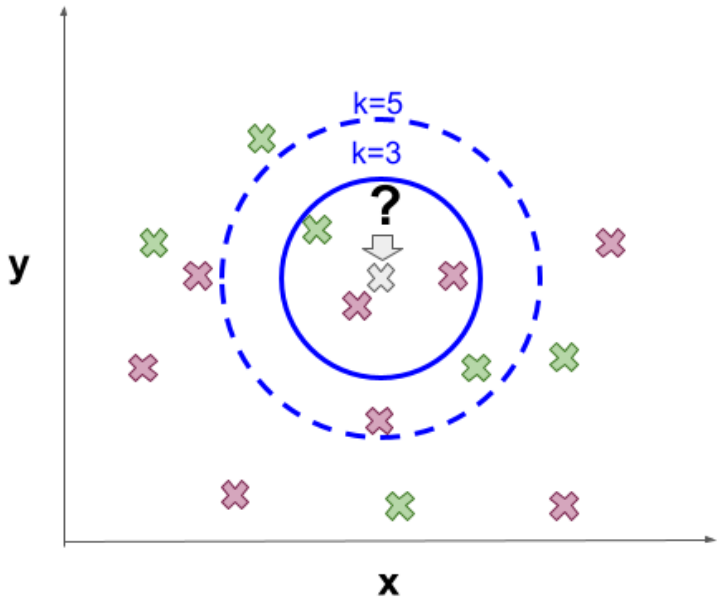


Figure 3.2: Schematic example of a binary KNN classifier. In the case of the solid line, the three nearest neighbors are considered. In the case of the dashed line, the prediction is drawn from the class labels of the five nearest neighbors. In both cases the majority of neighbors belong to the purple class, so the compound of interest is predicted to belong to this class, too.

sample is predicted by following a branch of (mostly binary) decisions from node to node until a leaf node (see Figure 3.3). The decision tree is capable of capturing also simple but non-linear relations while still preserving a high degree of interpretability. However, the disadvantage of this algorithm is its tendency to overfit the training data. Complex relationships can also be described by a RF model (see Figure 3.4). This model combines several single decision trees that are trained on randomly selected subsets of descriptors and training instances. A final prediction is derived from majority voting over all trees. In the case of binary classification, the predicted probability is defined as the mean over the results from the single trees, which is equivalent to the percentage of trees that predict class 1. This predicted probability is often used as a measure of reliability for the final prediction (section 3.2.3). Compared to single decision trees, RF has the advantage of still describing complex relationships while being less prone to overfitting. However, these advantages come at the cost of higher computational effort and decreased interpretability. In a SVM model, the descriptor space spanned by the training instances is divided by a multidimensional hyperplane that separates instances with differing class labels by the largest possible margin (see Figure 3.5). In this case, the distance of a test substance to the hyperplane can serve as a measure of reliability of the prediction. The characteristics of a SVM model strongly depend on its kernel (i.e. the mathematical function utilized for separating the samples). A kernel

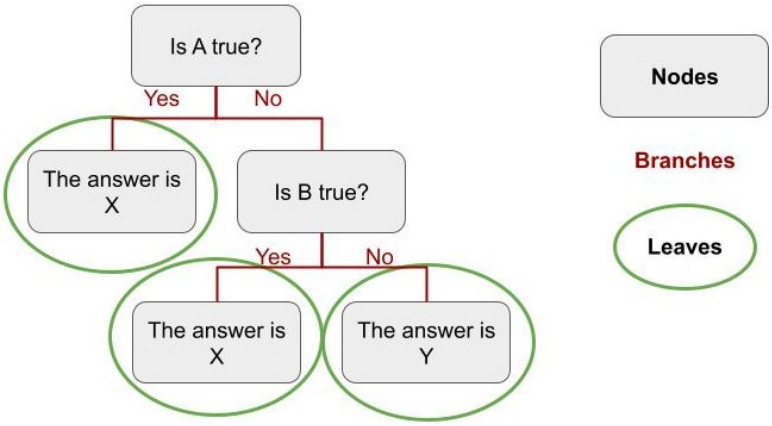


Figure 3.3: Example of a simple decision tree. Leaves, nodes and branches of the tree are marked.

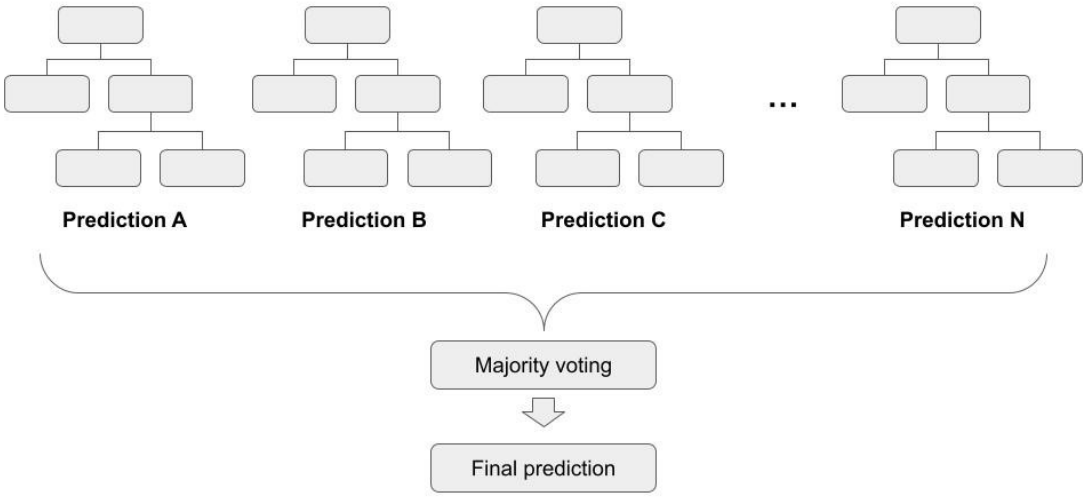


Figure 3.4: Scheme of a RF model. Final decision is drawn by a majority voting over all N trees included in the model.

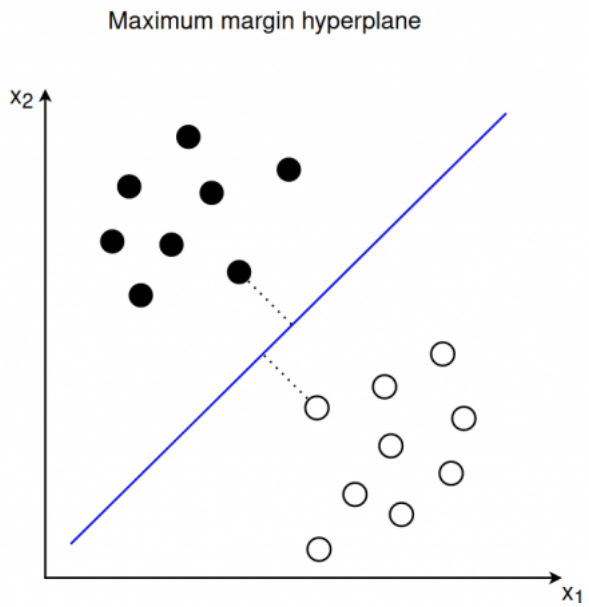


Figure 3.5: Schematic example of a two-dimensional SVM classifier: The blue line separates the open and the filled data points from each other with the maximum margin [54].

of higher complexity should achieve better performance on complex data, but is, at the same time, more prone to overfitting. The interpretability of SVM models is, as for other more complex ML models, relatively limited.

3.2.3 Applicability domain

The reliability of the predictions returned by a ML model is not evenly distributed in chemical space [55]. The AD of a model should always be defined to differentiate reliable predictions (in domain) from unreliable ones (out of domain). Two different main concepts for the definition of the AD have been described [56]: novelty detection and confidence estimation. Novelty detection defines the AD by the similarity of a query compound to the training data of the model. This concept is based on the observation that the model's reliability decreases with decreasing similarity of the query compounds to the training data. The characteristics of the novelty detection methods hence depend on the definition of similarity applied in each case. In contrast to novelty detection methods, confidence estimation approaches make direct use of the information returned by the model (e.g. prediction probability returned by a RF model or distance to the decision threshold returned by a SVM model). By doing this, confidence estimation approaches can also identify predictions that are well covered by the chemical space of the training data, but within a region where prediction reliability is hampered by the vicinity of samples with conflicting class

labels. In direct comparison, confidence estimation has proven advantageous in AD definition when compared to novelty detection approaches [56].

3.2.4 Conformal Prediction

A confidence estimation method alternative to the classical AD definition is CP, an algorithm processing the reliability of predictions returned by a ML model. It offers the advantage of mathematically defining the AD without the need of defining arbitrary thresholds as cutoffs. As long as the randomness assumption of the samples holds true (an assumption that is also part of classical ML methods), a CP model will always return predictions with the user-defined reliability $1 - \varepsilon$ [57, 58]. The desired error significance ε is mathematically anchored in the model itself and can be defined by the user for the specific task.

The basic implementation of CP is inductive CP (see Figure 3.6 A). In the inductive CP framework, the available training data are divided into a proper training set and a calibration set. A ML model (see section 3.2.2 for an overview on different ML models) is trained on the proper training set only and then applied to both the calibration and test sets. The probability estimates returned by the ML model are processed within a nonconformity function to calculate an α-value or nonconformity score for every instance of the calibration and test set. In classification models, the inverse probability error function (Equation 3.1) or margin error function (Equation 3.2) are the most commonly used nonconformity functions:

$$\alpha = \hat{P}(y_i|x) \tag{3.1}$$

$$\alpha = 0.5 - \frac{\hat{P}(y_i|x) - max_{y \neq y_i}\hat{P}(y|x)}{2} \tag{3.2}$$

with $\hat{P}(y_i|x)$ being the predicted class probability of class i and $max_{y \neq y_i}\hat{P}(y|x)$ being the maximal class probability for any other class. Due to the consideration of the class probabilities of the other classes, the margin error function is also well suited for multiclass classification. In the case of regression, the absolute error function (Equation 3.3) or the signed error function (Equation 3.4) are commonly applied:

$$\alpha = |y_i - \hat{y}_i| \tag{3.3}$$

$$\alpha = y_i - \hat{y}_i \tag{3.4}$$

whereas y is the predicted value returned by the regression model.
Based on these α-values, the p-values (or calibrated probabilities) of the test instances can be calculated as the relative rank of the test instance's α-value within the sorted list of α-values from the calibration set. A test sample will be assigned a specific class label whenever the corresponding p-value exceeds the desired significance level ε. Depending on the p-value of an instance and the significance level, none, one, or multiple class labels can be assigned to it. If no class label is assigned, the sample is considered to be outside of the AD

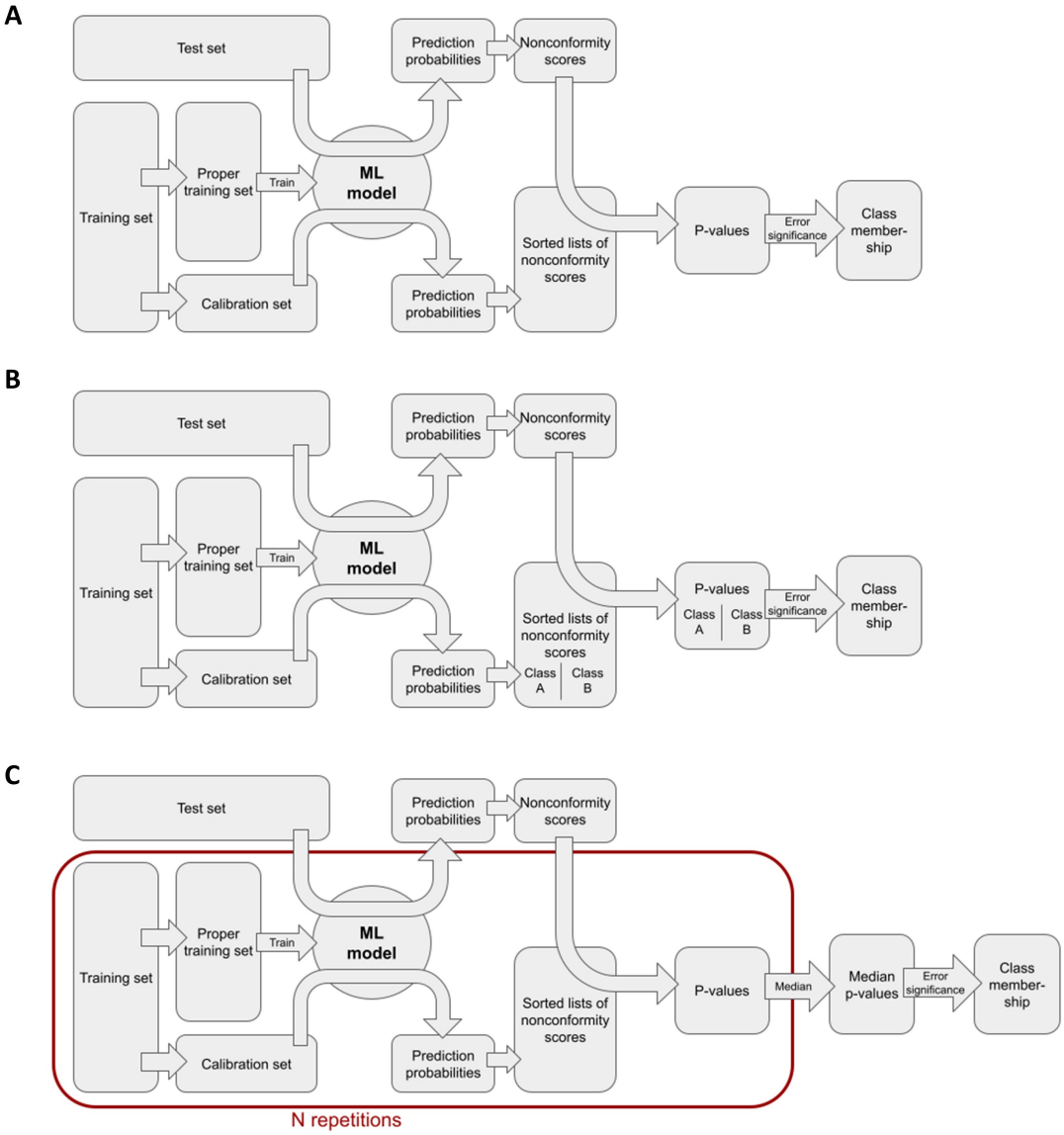


Figure 3.6: Schematic workflow of different versions of CP: (A) inductive CP, (B) Mondrian CP, and (C) aggregated CP.

of the CP model.

Different modeling requirements can be addressed by different variants of CP [59]. The most common variants are Mondrian CP and aggregated CP. Mondrian CP (see Figure 3.6 B) is best suited for modeling imbalanced data sets [60]. In this CP variant, individual lists of α-values are created for each class and used to derive class-specific p-values. Hence, the relative ranks of the test α-values do not depend on the prevalence of each class in the training data. The aggregated CP variant (see Figure 3.6 C) was designed to reduce the effects of not using the complete training set for the ML model and is therefore well suited for small data sets. In aggregated CP, the splitting of the training set into proper training and calibration data is repeated several times [61]. The final p-values are then derived by averaging the p-values over all repetitions (most commonly by the median, but also the maximal value or the mean for example can be used). With this approach the number of data points not used for modeling is reduced compared to an inductive CP workflow. This comes along with a strong increase in computational effort depending on the numbers of iterations N selected.

3.3 Computational representation of molecules

3.3.1 Representation of molecular structure

In order to be processed within a computational program, molecular structures need to be transferred into a machine readable format. Two and three dimensional molecular structures can, for example, be stored in high precision in Mol files, structure data (SD) files or XYZ files (in the 2D case, usually the Z component is artificially set to 0) [62]. In a XYZ file, each atom in a molecule is described by its element symbol and its Cartesian coordinates. Atom connectivity is implicitly given by the distances resulting from the coordinates. In a Mol file, the coordinate information is complemented with information on connectivity and molecular and atomic properties, like charges and isomers. Several Mol files can be combined in a single SD file. The SD format also allows for the inclusion of additional information about each molecule, like alternative names, experimental properties, or data sources. While these formats are capable of capturing the molecular structure with high precision, they come along with comparably large file sizes and the correspondingly higher computational effort for processing them. Moreover, the exact molecular structure is often not known and may change with the molecular environment (e.g. temperature, aggregation state, solvation state, etc.), so that such highly precise formats are mainly useful only if a detailed structure optimization of the compounds has been undertaken.

A lean alternative to formats capturing the exact structure of a molecule is given by various line notation formats like international chemical identifier (InChI) [63]

or simplified molecular-input line-entry system (SMILES) [64]. Both formats capture the 2D (and partialy 3D) structure of a molecule by describing atom types and connectivity between atoms. An approximated 2D structure can be derived from this information by most molecule-processing programs without explicit description of the exact atom positions. While InChI strings define the protonation state of each atom (from which the bond types can be deducted), SMILES strings explicitly define bond orders. Both formats share similar advantages and disadvantages: They both result in small and flexible structure representations and can be read by computational programs in a shorter time frame compared to the previously described, more information-rich formats like Mol, SD or XYZ files. At the same time, InChI and SMILES strings can only be interpreted by specialized programs and can cause a loss of information as they are not able to capture some structural features. Compared to the InChI notation, the SMILES notation is easier to interpret for human investigators.

3.3.2 Molecular descriptors

For a ML model to be applied to molecular data, molecules need to be encoded in a machine readable format. This can be realized by so-called molecular or chemical descriptors. The derivation and selection of a discriminative set of molecular descriptors is crucial for development of a predictive (Q)SAR model. To achieve this goal, a variety of different types of molecular descriptors is available [65].

Molecular descriptors can be classified by the dimensionality of the molecular structure that is captured, as also pointed out in our review article [P1]. While 0D descriptors cover information derived from the chemical formula (e.g. atom counts), 1D and 2D descriptors cover the absence or presence of certain substructures and the atom connectivity, respectively. Finally, 3D descriptors can be derived from the three dimensional structure of the molecule only (e.g. quantum chemical properties or volume descriptors).

Binary information on the 0D, 1D and 2D structure of a molecule (e.g. absence or presence of certain atom types, bond types, charges, or chemical subgroups) can be assembled into so called fingerprints (i.e. Boolean bit strings), which can be further divided into structural keys and hashed fingerprints. In structural keys like molecular access system (MACCS) keys [66] (composed of 166 bits) or PubChem fingerprints [67] (with 880 bits), each bit can be directly associated with the absence or presence of exactly one molecular substructure or fragment. Detailed lists for the translation between fingerprint and molecular substructure are available and allow human interpretation. In contrast to structural keys, hashed fingerprints may encode the absence or presence of several different fragments on the same bit. Hence, hashed fingerprints allow for the representation of a higher number of features with the same number of bits compared to structural keys. At the same time they lose interpretability, since hashed fingerprints cannot be unambiguously linked to specific functional groups.

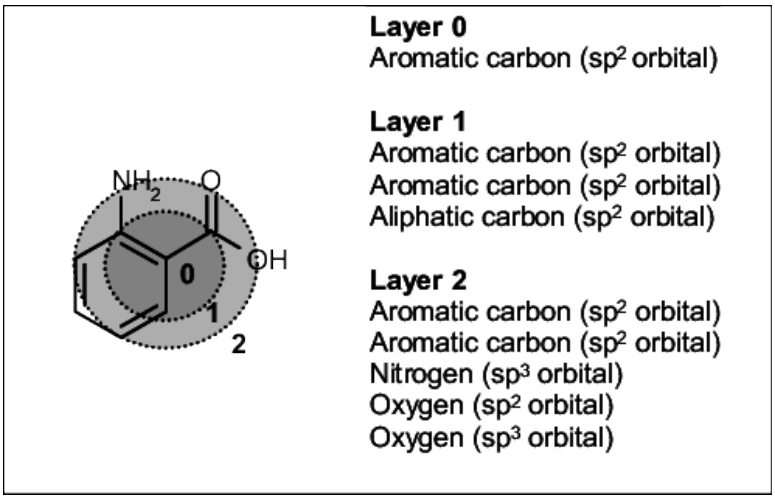


Figure 3.7: Example of the generation of a circular fingerprint [71].

Hashed fingerprints can be divided into topological or path-based fingerprints (e.g. Daylight fingerprint [68]), circular fingerprints (e.g. extended-connectivity fingerprints (ECFP) [69], or Morgan fingerprints [70]), depending on the method used for the enumeration of the encoded fragments. While topological fingerprints encode the molecular paths that can be found starting from each atom in the molecule, circular fingerprints (see Figure 3.7) encode the radial environment of each atom up to a defined bond order. This cutoff bond order X is usually indicated in the name of the fingerprint (i.e. ECFPX or MorganX). In addition, fingerprints can be specialized due to specific molecular structural features. For example, the chemistry development kit (CDK) extended fingerprint [72] (which is also available via the pharmarceutical data exploration laboratory (PaDEL) software [73] and is called PaDEL extended fingerprint in this thesis) is a path based fingerprint that includes additional bits for different numbers of rings and numbers of rings in fused ring systems. Analogously, the 79 bit CDK E-state fingerprint (called PaDEL estate fingerprint through this thesis) encodes a molecule by the electrotopological state indices of its atoms. Several popular descriptors and fingerprints can be created with the popular software package RDKit, too [74].

Whereas fingerprints directly encode the structure of a molecule, other sets of descriptors include calculated physicochemical properties that are mathematically derived from the molecular formula or structure. Prominent examples of such non-binary descriptor sets are the descriptor sets provided by molecular operating environment (MOE) [75], PaDEL [73, 76], and Dragon [77] software. Each of these descriptor sets includes a variety of descriptors starting from simple atom and ring types and counts to calculated molecular properties or eigenvalues. A large overlap between the descriptors covered by these programs can be observed. In this thesis, we call these sets of descriptors physicochemi-

cal descriptors. While MOE (>400 descriptors) and Dragon (5,270 descriptors) are proprietary tools, PaDEL (1,875 descriptors) is available as an open source license and the application is free of charge. Several of the descriptors provided are calculated from the 3D structure of the molecule. This requires structure optimization prior to descriptor calculation. Since this demands high computational resources and the success strongly varies depending on the molecule and the methods selected for optimization, the usage of such 3D descriptors is only recommended if a significant increase in model's performance is expected.

3.3.3 Feature standardization and feature selection

Non-binary descriptors like physicochemical descriptors can cover very different ranges of values. To compare the weights given to a descriptor by a modeling algorithm and for several feature selection and modeling algorithms to work properly, feature standardization is needed in advance. Most commonly, this is conducted by projecting the features to have their mean set to zero and their variance set to unit variance at the same time. The standardized feature vector x' is derived by subtracting the mean of the feature vector $\overline{x}$ from the original feature vector x and division by the standard deviation σ:

$$x' = \frac{x - \overline{x}}{\sigma} \tag{3.5}$$

Not all descriptors are evenly suited for different modeling tasks. A careful selection of the descriptors by rational consideration or computational feature selection methods is advantageous. While the presence of additional relevant descriptors can improve a model's performance, the presence of irrelevant or redundant descriptors increases modeling time as well as the danger of overfitting. Feature selection methods can be distinguished by the way they are applied to the data of interest into filter, wrapper, and embedded methods [78]. A filter method selects a subset of descriptors by optimizing a measure independent of a ML algorithm. Prominent measures that are suitable for such filter methods include mutual information, Pearson product moment, or correlation coefficient. In contrast to filter methods, wrapper methods make use of a ML algorithm to evaluate the suitability of a subset of descriptors for the modeling task: A model is trained on every subset of descriptors and gets evaluated on holdout data. Due to the high number of models to be built, these methods are computationally more demanding than filter methods, but the results are optimized for the modeling task. For wrapped feature selection, any ML algorithm that is suitable for the specific modeling task can be applied. Finally, embedded feature selection methods combine aspects of both the filter and the wrapper method. The most prominent example of embedded feature selection is least absolute shrinkage and selection operator (LASSO) regression [79]. It minimizes the result of a linear cost function by minimizing the coefficients of the different descriptors. Descriptors that encode irrelevant

or redundant information will hereby yield coefficients close to zero and can be discarded from the final model. When applying LASSO regression for feature selection one should keep in mind that in the case of descriptors encoding redundant information one descriptor could potentially be replaced by the other in the final model. The model selects which descriptor to keep and which to dismiss after optimizing the performance on the data it is trained on. A human investigator might prefer a different decision.

3.4 Model performance evaluation

A variety of different measures is available for evaluation of models performance. Most of the performance measures are calculated from the counts of true negative (TN), false negative (FN), true postive (TP) and false positive (FP) predictions, which can be retrieved from the confusion matrix of a binary classifier as depicted in Figure 3.8.

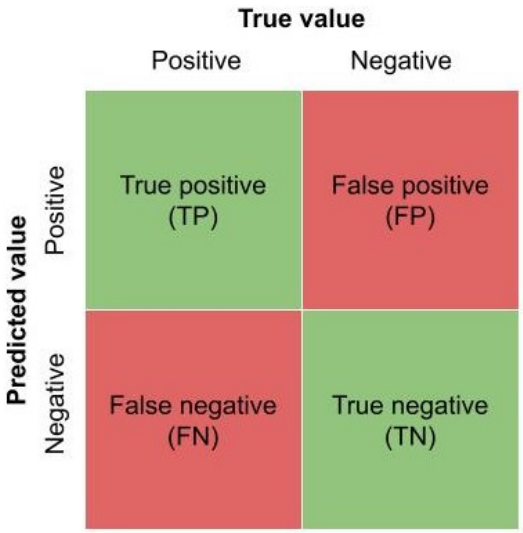


Figure 3.8: Confusion matrix of a binary classifier.

Typical measures to characterize classical models are: accuracy (ACC), Matthews correlation coefficient (MCC), F_1 score, correct classification rate (CCR), sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV) and coverage. A CP model is usually characterized by the two measures validity and efficiency.
The ACC is defined as the percentage of correct predictions within all predictions. For a binary model this is defined as:

$$ACC = \frac{TN + TP}{TN + FN + TP + FP} \tag{3.6}$$

The ACC is the most intuitive and therefore also the most common performance measure for prediction models. Nevertheless, it does not represent the whole

picture of the model's performance, since it also depends on the fraction of each class. Two alternative measures of reliability try to circumvent this downside: The MCC, the F_1 score and the CCR, which is also known as balanced ACC:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.7}$$

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3.8}$$

$$CCR = \frac{sensitivity + specificity}{2} \tag{3.9}$$

With sensitivity and specificity reflecting the ability of the model to successfully identify active or inactive compounds:

$$sensitivity = \frac{TP}{TP + FN} \tag{3.10}$$

$$specificity = \frac{TN}{TN + FP} \tag{3.11}$$

The percentage of correct predictions among all positive or among all negative predictions is called PPV and NPV:

$$PPV = \frac{TP}{TP + FP} \tag{3.12}$$

$$NPV = \frac{TN}{TN + FN} \tag{3.13}$$

For a classical model, coverage is defined as the percentage of test compounds for which a prediction within the AD of the model is returned.

$$Coverage = \frac{TP + TN + FP + FN}{Total\ number\ of\ test\ samples} \tag{3.14}$$

For CP models, the most characteristic performance measure is validity. It is defined as the percentage of predictions that include the true prediction. Since a CP model can also predict more than one class to be true for a compound, this also includes unambiguous predictions with the true class being predicted among others. A model is considered valid, as long as the validity is close to the expected value of $1 - \varepsilon$. Analogously to the definition of coverage of a classical model, the efficiency of a CP model is defined as the percentage of distinct predictions that are returned under the selected error significance ε. Validity and efficiency can also be returned class wise (i.e. exclusively for molecules assigned experimentally to a specific activity class). A CP model considered valid should also be valid for any single class covered by the specific classifier.

3.5 Approaches and data sets for the prediction of skin sensitization potential and potency that are reported after the publication of P1

Since our review article [P1] was published in 2018, several new developments in the field of computational prediction of skin sensitization potential and potency could be observed. On the one hand, Di et al. [16] collected, curated and published a new LLNA data set comprising binary and quinary LLNA class labels for 1007 compounds. In 2019 we could show that this data set includes 440 substances that are not present in the LLNA data set compiled by Alves et al. in 2018 [2]. On the other hand, several new computational models or updated versions of existing models have since been published. While most of these models are based on existing collections of public data that have been used previously or on proprietary data that cannot be investigated, Di et al. [16] published models trained and tested on a new collection of LLNA data. In addition to this, several tools to integrate existing knowledge and results from non-animal testing approaches as well as from computational tools have become available. A new development within this field is the relatively high number of AOP related reasoning frameworks that do not return an unambiguous prediction, but are meant as structuring and assisting frameworks for expert decision-making.

3.5.1 New models for the prediction of skin sensitization potential

Several new computational models aim for the prediction of binary skin sensitization potential in humans or animals. These will be shortly summarized in the present subsection:

Di et al. [16] developed and compared 81 binary and ternary QSAR classifiers for the prediction of binary and ternary skin sensitization potential. The models are based on a new collection of 1007 LLNA data annotated with binary and quinary skin sensitization potential. The high number of classifiers originates from using all combinations between nine modeling algorithms (i.e. SVM, decision tree, gradient boosting, RF, tree ensemble, probabilistic neural network, multilayer perceptron, and fuzzy rules) and nine different molecular fingerprints calculated with PaDEL software. The best binary and ternary models achieve an ACC of 0.81 and 0.71 on the test set, respectively. In addition to these global models, several local models only incorporating substances from one out of five activity classes are presented. Their ACC on the test set ranges up to 0.89 and 0.85 (for two and three potency classes, respectively) for molecules which are assigned by the authors to undergo Michael addition reactions (131 substances in total). An AD is defined and applied for all models. Closer analysis of the ternary models by ourselves suggests, that the non-published class-wise performance measures of the ternary models might be unacceptably

low [2].

A free online tool called STopTox was published by Borba et al. [80] in 2020. It provides binary ML models for six toxicity endpoints. One of these models is a RF model trained on 1000 LLNA data points with Morgan EFCP4 fingerprints (2048 bits) as descriptors. The model reaches a CCR of 0.70 in 5-fold external cross-validation (CV).

Predicted binary outcomes for LLNA as well as for human skin sensitization potential and the outcome of three non-animal testing approaches (DPRA, KeratinoSens, h-CLAT) are also provided by the now updated version 3 of the Pred-Skin webserver [81], which we previously discussed in our 2018 review, albeit with an older version [40]. Predicted results from the five models are integrated within a naive Bayes model to predict skin sensitization potential in humans. While this approach takes advantage of a larger data basis originating from different data sources, it does not require any additional testing for a new substance to be predicted.

3.5.2 New models for the prediction of skin sensitization potency

Other approaches aim for the prediction of skin sensitization potency as given by the EC3 value. Such quantitative prediction is advantageous, since it eliminates the need to exclude all sensitizers from a product, and allows for the use of sensitizing molecules in a safe dose adjusted to the predicted potency. Nevertheless, such quantitative models for skin sensitization prediction (if tested correctly) do not reach sufficient performance for an all-encompassing application. A main reason for this might be the relatively small number of data points available for modeling and testing and the variance in the data available. Two approaches to circumvent these limitations have been presented recently: the reduction of the model's applicability to a small and well defined class of molecules (local model) [82] and the re-projection of predicted EC3 values into two or more potency classes [83].

Gleeson et al. [82] developed a local model to predict skin sensitization potency for molecules of the Schiff base domain. Reaction energies of 22 molecules with Lysine sidechains were calculated with quantum mechanical methods. Based on a training set of 14 molecules, a linear correlation between the reaction energy combined with clogP and the pEC3 was derived. On the test set of 8 and 6 molecules $r^2 = 0.49$ and $r^2 = 0.62$ was derived, respectively. Weaknesses of the model are explained by the neglection of protein-specific steric and electronic effects and possible differences in the immune response caused by different adducts. In addition, the authors point out, that reactions could take place from tautomeric states, sides of reaction, or metabolites not considered in the

quantum mechanical calculation.

Kim et al. [83] reported on a linear correlation between the EC3 value and several physicochemical properties of 212 skin sensitizers and 38 non-sensitizers investigated. Highest correlation of the EC3 could be found to surface tension, melting point, or boiling point with correlation coefficients of 0.65, 0.69 or 0.44, respectively. P-values of the correlation are lower than 0.00001 for all three of these properties. More recently [84], the authors evaluated the capacity of these physicochemical properties to distinguish two (ACC=0.73) or three (ACC=0.64) potency classes on a slightly larger data set of 305 sensitizers and 57 non-sensitizers. Results were compared with the prediction of skin sensitization potential from Toxtree (ACC=0.70), Vega (ACC=0.81) and Danish EPA QSAR (ACC=0.56).

3.5.3 New models that integrate results from non-animal testing approaches

A promising route for reliable skin sensitization prediction is the integration of experimentally derived non-animal data (i.e. in vitro or in chemico testing results) as descriptors for a computational model [81, 85–87]. While these kind of models show promising increases in applicability and reliability compared to purely computational models, they suffer from the disadvantage of needing every substance to be tested in more or several assays prior to the prediction. This might be suitable for the evaluation of preselected and promising substances but not for a first and comprehensive screening.

Natsch et al. [29], compared the predictivity of the two out of three DA (an approach that considers a substance as skin sensitizer as soon as at least two out of three non-animal testing approaches addressing the three first key events from the skin sensitization AOP have tested positive). Already with this simple rule based approach, balanced accuracies between 0.76 and 0.94 are reported for the prediction of binary skin sensitization potential in humans. In the same meta-study, the balanced ACC of the LLNA for human skin sensitization potential is reported with between 0.58 and 0.88 depending on the underlying data set.

Silva et al. [85], combined experimental non-animal data with computationally derived molecular properties, fingerprints, and reactivity descriptors for linear regression. The models to predict two, three and six potency classes achieved accuracies of 1.00, 0.99 and 0.98, respectively. The models have to be viewed with caution since they are built and tested on a low number of data points (between 81 and 90 depending on the test case) in combination with a relatively high number of descriptors (up to 386 depending on the test case). In addition, the best performing test case includes a protein adduct formation descriptor obtained by the TIMES-SS package software. Training data of TIMES-SS are

not published and thus do not allow for a comparison with the test data evaluated here.

SkinSensDB [88,89], a free web server providing data on DPRA/PPRA, KeratinoSens/LuSens and h-CLAT as well as on human skin sensitization potential, also includes functionalities to assist read-across assessment of human skin sensitization potential [89]. Based on these data, Tung et al. published SkinSensPred in 2019. SkinSensPred is an ensemble tree-based multitask learning model that predicts human skin sensitization potential by leveraging the outcome of DPRA/PPRA, KeratinoSens/LuSens, h-CLAT and human skin sensitization potential as four simultaneous learning tasks [90]. The model can be assessed on the SkinSensDB web server, too.

In 2019, Li et al. [86] developed SVM models with accuracies up to 0.91 and 0.69 to predict binary and ternary skin sensitization potential in humans, respectively. The models incorporate a data-rebalancing ensemble learning algorithm and make use of non-animal testing results as well as on six (mostly experimentally derived) physicochemical properties as descriptors and are trained and tested on 96 and 32 substances from the Cosmetics Europe data base.
In 2021, Ambe et al. [87] added chemical information to the descriptor set used by Li et al. and achieved an r^2 of 0.75 in predicting EC3 values with a CatBoost based regression model on the Cosmetics Europe data set.

The Bayesian network integrated testing strategy (BN ITS-3) developed by Jaworska et al. [91] in 2015 also makes use of experimental non-animal results in combination with physicochemical parameters and structure-based predictions by TIMES. This model was evaluated in 2020 by Otsubo et al. [92] on a test set of 175 substances. The study derived ACCs of 0.93 and 0.66 for binary and ternary classification, respectively.

3.5.4 Recent studies evaluating and comparing existing skin sensitization prediction tools

Several common tools for skin sensitization prediction were compared in 2020: Golden et al. [93] evaluated the performance of eight computational tools for the prediction of binary skin sensitization potential (Toxtree, PredSkin, OECD QSAR Toolbox, REACHAcross, Danish QSAR Database, TIMES-SS, and Derek Nexus) with respect to the experimental outcome in human studies from two different data sources. These are the data set of Basketter et al. which comprises 131 highly curated substances (107 sensitizers and 24 non-sensitizers) annotated with human skin sensitization potential and the HSDB data set which comprises data on 375 substances on a screening level. Overall, most models investigated yielded ACCs between 0.70 and 0.80, which is comparable to the ACC for the LLNA predicting skin sensitization in humans (ACC between 0.74 and 0.82

depending on the data source). Closer investigation of the mispredicted compounds lead to the assumption that a combination of models could be beneficial.

Several models and data sources for the prediction of skin sensitization potential were reviewed by Ta et al. in 2021 [94]. However, this review has to be taken with caution, since not all models are labelled and cited correctly. In 2021, Santín et al. [15] specifically focused on artificial intelligence (AI) for toxicity prediction. Prediction of skin sensitization potential and potency is also mentioned with two examples in the article, but not covered in detail.

3.5.5 New objectives and reasoning workflows for expert judgement

Currently, computational methods can also be used to structure information and guide expert judgement. Several such reasoning frameworks have recently become available for the skin sensitization endpoint [39, 95, 96]. Their structure mainly follows the AOP for skin sensitization. Since those frameworks do not return unambiguous predictions, quantitative evaluation of their performance is not possible.

Expert judgment can also be assisted by increasing the usability of existing computational tools. This can either be realized by providing additional programs for the interpretability and plausibility estimation of existing predictions [97] or by the unification of several computational tools into one platform [98]. With SpheraCosmolife [98], a new platform for the risk assessment of cosmetic products has become available. It provides results from several VEGA models for endpoints relevant for cosmetic products, which also includes the one for the prediction of skin sensitization potential we previously discussed in our review article [P1]. Overall, the platform aims to assist with finding a safe dose depending on the product's application.

Computational methods can not only be applied to identify potentially problematic compounds, but can also help expand the understanding of mechanisms of the processes underlying skin sensitization. In 2019, Di et al. [99] applied computational tools to identify 33 dermatitis-related targets and 12 dermatitis-related pathways that might play a vital role in the induction of skin sensitization. Such a mechanistic understanding might also help to promote computational skin sensitization prediction tools in general, since they can help to interpret and gauge existing predictions.

4. Aims of the present work

Skin sensitization is an important endpoint for the development and registration of new chemicals and consumer products. Ethical considerations as well as regulatory requirements engender a shift from animal experiments towards an advanced non-animal safety assessment. Compared to non-animal testing approaches, computational approaches bear an advantage with respect to costs, testing time, and testing facilities and can be applied at an early stage of product development as well as alongside testing approaches. For computational methods to be accepted in a regulatory context, stringent requirements on prediction accuracy, measure of reliability and data quality are vital.
Within this thesis, we developed different computational models for the prediction of skin sensitization potential of small substances while focusing on increasing the model's value for risk assessment and registration. Herein we extensively addressed the following questions:

1. **Which computational models for the prediction of skin sensitization potential and potency exist and what are their advantages and limitations?** The theoretical prediction of skin sensitization potential and potency has a long history, starting from the first RAI models developed in the 1980s. Today, a variety of different approaches for the computational prediction of skin sensitization potential and potency are at hand. Nevertheless, no single prediction tool can be considered as a sufficient standalone method for reliable skin sensitization prediction in a regulatory context. Within a comprehensive review article, we investigated, structured, and qualitatively evaluated the available models and their underlying methods. We could point out common or potential pitfalls and formulate aims for possible future modeling approaches.

2. **Which skin sensitization data are available and what is the largest high quality data set we can compile?** The predictivity and reliability of a ML model strongly depend on the quantity and quality of the available data. For the prediction of skin sensitization potential and potency, quality and quantity of the data are highly limited. The best trade-off between quality and quantity of skin sensitization data can be found when utilizing LLNA data. With the aim of optimizing the data basis for further model development and the evaluation process, we created and curated the so far largest publicly available high quality LLNA data set. Importantly, the manual investigation of every single molecule from the data set in

common public available data sets ensured the highest possible standards for the structures associated with entries in our data set.

3. **How relevant is our data set for the chemical space of cosmetics, approved drugs, and pesticides?** A ML model can only be predictive for molecules covering the same chemical space as the training data. In order to prove the relevance of our models for molecules labeled as pesticides, approved drugs, or cosmetics, we closely investigated the chemical space covered by these reference data sets as well as by our LLNA data set using pair-wise similarity analysis and PCA.

4. **Can binary skin sensitization potential measured with the LLNA be predicted by ML algorithms?** With the aim of finding the best combination of ML algorithms, hyperparameters, and descriptors, we developed, optimized, and evaluated hundreds of ML models for the prediction of skin sensitization potential. The best performing models were closely investigated regarding their performance and the descriptors selected. Based on practical considerations, two of the best performing models were selected as the primary Skin Doctor models. They have been made applicable for public use via our web service new e-resource for drug discovery (NERDD) as the Skin Doctor suite for skin sensitization prediction.

5. **Can we quantify the reliability of our ML models and flag the most unreliable predictions?** The reliability of a ML model is not equally distributed over chemical space. Areas not (well) covered by training instances as well as areas covered by molecules with diversified class labels might have lower performance than that of the model's average. For the practical usability of a ML model, it is of highest importance to flag and dismiss such unreliable predictions and to return the model's reliability based on the reliable predictions only. To reach this goal, we defined an AD as well as two measures of reliability for our best performing models and demonstrated their applicability on our test set. For each of the reliability measurements, two different thresholds were suggested to either only remove the most unreliable predictions or to further increase model's predictivity at the cost of a decreased coverage. Both measures of reliability as well as the AD have been integrated into the Skin Doctor suite web service.

6. **Can we quantify the reliability of every single prediction returned by our model in a mathematically proven way by enveloping the model into a CP framework?** For a ML model to be applied in risk assessment, it is highly advantageous to not only know the overall performance of the model, but also the expected reliability of a single prediction for a compound of interest. With the aim of retrieving this information

for every single prediction returned by our models, we enveloped one of the best performing models from the Skin Doctor suite into an aggregated Mondrian CP workflow. This procedure also supersedes the need to define arbitrary thresholds for the separation of reliable and unreliable predictions since its measure of reliability is mathematically proven and derived from the model itself. The final CP model was published on our web server under the name Skin Doctor CP.

7. **Can we transfer our findings from binary classification of skin sensitization potential to a ternary classification of non-sensitizers, weak to moderate sensitizers, and strong to extreme sensitizers?** While most models available for skin sensitization prediction (including the ones developed by us) only address binary skin sensitization potential, it is extremely desirable for risk assessment to also quantify the skin sensitization potency of a compound, since this would allow for the usage of less potent sensitizers in safe doses. To support this need, two different approaches to shift our aggregated Mondrian CP model from a binary to a ternary classifier were developed and closely investigated regarding their opportunities and pitfalls. A close comparison with a ternary model for skin sensitization prediction published by another group on a subset of our data set concluded on the weaknesses that originate from the sparse data basis and that cannot be fully compensated by our ML approach.

8. **Is it possible to substitute the non-intuitive descriptors used for modeling by a small and biologically meaningful set of alternative descriptors to increase model's interpretability?** For a ML model to be accepted by risk assessors and regulatory instances, human interpretability is desirable. While in theory this aim can be supported by a RF modeling algorithm, it is not supported by the large and partly unintuitive sets of descriptors selected for our original Skin Doctor models. With the aim of increasing the interpretability of our models, we substituted the non-intuitive fingerprints by a set of bioactivity descriptors calculated as the results from 375 CP model predicting the outcomes of variety of assays. A strict feature selection process reduced the number of descriptors to the lowest possible value without losing too much information. The 10 bioactivity descriptors selected for the final model were also investigated regarding their biological connotation as well as their possible link to the skin sensitization AOP.

5. Methods

To reach the aims outlined in chapter 4 of this thesis, we applied a variety of computational methods, which will be outlined in the present chapter. For more details on the computational methods applied, we refer to the Methods sections in the corresponding papers [P2], [P3] and [P4], which are provided and discussed in chapter 6 of this thesis.

5.1 Data resources

Within the present thesis, skin sensitization potential and potency of substances are described by LLNA data. The LLNA data utilized within the present work originate from two public available LLNA data sets: the data set collected by Alves et al. [40] and the data set collected by Di et al. [16]. Both data sets have been merged by us in 2019 while removing any compound with contradicting class label. The resulting data set comprises 1416 unique compounds with binary and partly quinary skin sensitization class labels and was the basis for the original Skin Doctor models. Details on the creation of this data set can be found in the corresponding publication [P2].

A further manual data curation step was introduced and described in 2020 by us (for details see publication [P3]). Based on the CAS number or any other identifier available from the original data sources [16, 40], we visually inspected the corresponding entries in common chemical databases or catalogs. We manually removed all molecules from our data set for which we could not confirm the chemical structure derived from our primary data sources. This includes compounds with CAS numbers that lead to metal complexes, metal salts or polymers. Molecules without a defined structure as well as molecules that are linked to a structure different from the one in our primary data sources are also removed from our data set. The same is true for multi-component structures without a defined primary component. This additional data curation step reduced the size of the data set from 1416 to 1285 compounds while drastically increasing the quality of the data contained. The refined data set can be regarded as the largest well curated LLNA data set that is available in the public domain at present. It is the basis for our Skin Doctor CP model [P3] as well as for the analogous model based on bioactivity descriptors published in 2021 [P4].

In addition to the modeling data, several reference data sets have been utilized for the investigation of chemical space. These include different versions of data sets comprising cosmetics [100,101], approved drugs [102,103] or pesticides [101, 104]. In publication [P4], we developed 375 models for calculating bioactivity descriptors. Assay data for model building are derived from the publication of Alves et al. [40] and different sources listed in [105]. Additional information on the data resources utilized in this work is provided in the Methods sections of the respective publications [P2], [P3], and [P4].

5.2 Processing of molecular structures

All data sets utilized within the present thesis have undergone the identical automatized data curation and standardization pipeline comprising the following steps: (i) the removal of counter ions and neutralization of the remaining entity as described in the work of Stork et al. [106], (ii) standardization of tautomers (utilizing the "TautomerCanonicalizer" method implemented in MolVS [107]), (iii) removal of stereochemical information (which is not processed by the descriptors applied and is not present for all original structures), (iv) representation of molecules by canonical SMILES, and (v) deduplication based on canonical SMILES (In cases of multiple instances with identical class label, we only kept one. In cases of conflicting class labels, both entries were removed.).
Data sets utilized within publication [P4] have in addition undergone the data curation pipeline as described in [105]. This adds (vi) the removal of molecules containing any element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I and (vii) the removal of molecules with less than four heavy atoms.

5.3 Molecular descriptors

Within the first Skin Doctor models [P2] we evaluated different sets of descriptors and combinations thereof for their capacity to predict binary skin sensitization potential. Since the MACCS key fingerprint (for details on this set of descriptors, see section 3.3.2) consisting of only 166 bits turned out to be the best trade-off between complexity and model's performance, this was selected as descriptor for the Skin Doctor CP models [P3]. In our latest study [P4] we derived calculated bioactivity descriptors from 375 CP models. Those models are based on Morgan fingerprints with a radius of 2 and a length of 2048 bits. An overview of the descriptors investigated within this thesis can be found in Table 5.1.
All non-binary descriptors have been standardized with the StandardScaler function in scikit-learn [109]. This includes the shift of the mean of the descriptor to zero as well as scaling to unit variance.
In [P4] feature selection was performed with LASSO regression. The LASSO classifier was optimized with the scikit-learn LassoCV function (Linear_model module; random_state = 43, cv = 10, max_iter = 3000, n_alphas = 200).

Table 5.1: Descriptors used for model building and evaluation within this thesis.

Name within this thesis	**Description**	**Number of descriptors/ length of the fingerprint**	**Calculated with**
MOE 2D	0D, 1D and 2D descriptors	206	MOE [75]; all descriptors labeled as "2D descriptors" in the MOE software
PaDEL	0D, 1D and 2D descriptors	1444	PaDEL [73, 76]; this is the full set of 0D, 1D and 2D descriptors implemented in PaDEL software
MACCS	MACCS key fingerprint	166	RDKit [74]
Morgan2	Morgan fingerprint with a radius of 2 and a length of 2048 bits	2048	RDKit [74]
OASIS	OASIS skin sensitization protein binding fingerprint	5 bit fingerprint	OECD Toolbox [108]
PaDEL_est	PaDEL estate fingerprint	79	PaDEL [73, 76]
PaDEL_ext	PaDEL extended fingerprint	1024	PaDEL [73, 76]
Bioactivity descriptors	Calculated p-values for bioactivity in different assays	750	calculated in house

More detail on molecular descriptors, scaling and feature selection methods utilized within this work, can be found in the corresponding publications [P2], [P3] and [P4].

5.4 Modeling algorithms and hyperparameter optimization

Within the present work, several ML models were trained and tested on LLNA data employing either a RF or SVM modeling algorithm implemented in scikit-learn [109]. To ensure that models are tested on unseen data, our first LLNA data set derived in [P2] was split into training (80%) and testing (20%) data by stratified splitting with the train_test_split function of the model_selection module of scikit-learn [109] (data shuffling prior to data set splitting enabled). To enable comparison between the different versions of our models, the split into test and training set was kept constant through all our approaches, meaning that the training and test set of the refined LLNA data set applied in [P3] and [P4] are subsets of the training and test set of our first LLNA data set derived in [P2].

For the first Skin Doctor models [P2], a variety of RF and SVM models have been developed and evaluated. The hyperparameters of the modeling algorithms (RF: n_estimators and max_features; SVM: C and γ) were optimized within a 10-fold CV on the training data only. In publication [P3] optimal hyperparameters derived from [P2] (n_estimators = 1000, max_features = "sqrt", random_state = 43) were applied to the RF model underlying the CP workflow. In publication [P4] n_estimators of the RF model was set to 500 and all other hyperparameters kept as default values.
For a detailed view on the ML methods utilized, we refere to the Methods sections of the corresponding publications [P2], [P3] and [P4].

5.5 Reliability measures and definition of the applicability domain

For the first Skin Doctor models published in [P2] we developed two measures of reliability and one definition of the AD. As meaningful measures of reliability, the distance of a prediction to the decision threshold and the number of consecutive nearest neighbors with the same activity as predicted (measured by the Tanimoto similarity calculated on Morgan2 fingerprints) have been defined and tested in 10-fold CV and evaluated on the holdout data. For any model investigated, the mean Tanimoto similarity to the five nearest neighbors (measured from Morgan2 fingerprints) has proven as a conclusive measure of the AD. In a strict setting, any compound with a mean similarity smaller than 0.50 was considered out of domain. In a softened definition of the AD, the cutoff was set to 0.75. Nevertheless, the exact number of the cutoff as

well as the selection of the descriptor to define the AD and the number of consecutive nearest neighbors with same activity as predicted is to some degree arbitrary and not directly anchored within the specific model it is applied to. To compensate these downsides, we replaced the reliability measures as well as the definition of the AD by a CP workflow from 2020 on in publications [P3] and [P4]. The Methods sections of publications [P2] provides more detailed information on the implementation of the AD and the reliability measures of the corresponding work.

5.6 Conformal prediction

In publication [P3] and [P4] we employed aggregated Mondrian CP to ensure defined predictivity of our models.
Within our CP workflow, each training set was further divided into calibration (20%) and proper training set (80%) by stratified splitting with the train_test_split function of scikit-learn [109] (model_selection module; data shuffling enabled). Within our publications [P3] and [P4], this split was repeated 100 or 20 times, respectively, with different random_states applied.
Scikit-learn RF models (for technical details on the RF models, see section 5.4) were trained on each proper training set and applied to the calibration and test set. Non-conformity scores were calculated from the margin error function (Equation 3.2) for each class separately (following the aggregated CP protocol). The p-values of each substance from the test set were derived as the relative ranks of the corresponding non-conformity scores within the sorted lists of non-conformity scores from each calibration set. The final p-values for each substance are derived as the median p-values from all 100 or 20 random splits into proper training and calibration set, respectively. Within the present Mondrian CP workflow, non-conformity scores and p-values are treated separately for each activity class. More detailed information on the computational implementation of the CP models is provided in the Methods sections of the corresponding publications [P3] and [P4].

6. Results

Assessing skin sensitization is important for the development and approval of new substances and consumer products. Historically, the skin sensitization potential of chemicals was determined using animal models. Currently, it is preferable to predict skin sensitization using non-animal alternatives such as in chemico and in vitro assays and computational models. Compared to testing approaches, computational methods usually offer an advantage in testing time and financial expenses and can therefore also be applied to a large number of compounds during the early stages of development of drugs and cosmetics. Computational models are particularly beneficial, as they are able to utilize and integrate a variety of experimental data to make predictions. They can thereby also be utilized to seek deeper understanding into the mechanistic background of the skin sensitization AOP.

Today, a variety of computational tools to predict skin sensitization are available, but none of them is capable of fully replacing experimental approaches for risk assessment or regulatory approval. In this thesis, three different approaches to further promote computational methods for skin sensitization prediction have been presented.

6.1 Prediction of binary skin sensitization potential – evaluation of different combinations of machine learning algorithms and descriptor sets

Computational methods are a promising approach to predict the skin sensitization potential and potency of chemicals, minimizing the need of time consuming and expensive laboratory or clinical tests. Nevertheless, for computational methods to be applicable for risk assessment or regulatory approval, a high degree of well-defined predictivity is required. This includes methods to detect and flag predictions made by the model which may be unreliable.

In the following study, the largest known LLNA data set, to date, was compiled. This data set was used to train and test a variety of ML models to predict the binary skin sensitization potential. To demonstrate the relevance of these models, an extensive analysis was carried out to compare the chemical space covered by the LLNA data set to that of three reference data sets containing approved drugs, pesticides, and cosmetics. This comparison was conducted us-

ing PCA and pairwise-similarity analyses.

In this work, a total of 58 different combinations of ML algorithms (i.e. SVM or RF) and descriptor sets (one or two out of eight descriptor sets selected) were used for model building. An extensive grid search was conducted to find the optimal hyperparameters for the model for each of these combinations. The optimized models were then compared to each other. A solid definition of the AD and two additional measures of reliability were determined for the five best performing models (as measured by the MCC). The AD and measures of reliability were determined using a 10-fold CV and verified on the test set.

Finally, two complementary models and the corresponding AD and reliability measures have been made accessible through a public web service known as the Skin Doctor suite.

P2 **Wilm, A.**, Stork, C., Bauer, C., Schepky, A., Kühnl, J., Kirchmair, J., Skin doctor: Machine learning models for skin sensitization prediction that provide estimates and indicators of prediction reliability, *International Journal of Molecular Sciences*, **20**(19) (2019) 4833

Available at:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6801714/

Author Contributions: A. Wilm, J. Kühnl and J. Kirchmair concepualized the work. A. Wilm developed the methodologies, with contributions by C. Stork, C. Bauer, J. Kühnl and J. Kirchmair; A. Wilm developed the underlying software, with contributions from C. Stork. A. Wilm performed the data curation, validated the models, visualized and performed formal analysis of the research results. A. Wilm investigated the data and research results, with contributions by C. Stork, C. Bauer, J. Kühnl and J. Kirchmair. Resources were provided by A. Schepky, J. Kühnl and J. Kirchmair; A. Wilm wrote the manuscript, with contributions by C. Stork, C. Bauer, A. Schepky, J. Kühnl and J. Kirchmair. A. Schepky, J. Kühnl and J. Kirchmair supervised the work. A. Schepky, J. Kühnl and J. Kirchmair provided or aquired funding. All authors have read and agreed to the published version of the manuscript.

International Journal of
Molecular Sciences

MDPI

Article

Skin Doctor: Machine Learning Models for Skin Sensitization Prediction that Provide Estimates and Indicators of Prediction Reliability

Anke Wilm [1,2], Conrad Stork [1], Christoph Bauer [3,4], Andreas Schepky [5], Jochen Kühnl [5] and Johannes Kirchmair [1,3,4,*]

1 Center for Bioinformatics, Universität Hamburg, 20146 Hamburg, Germany; wilm@zbh.uni-hamburg.de (A.W.); stork@zbh.uni-hamburg.de (C.S.)
2 HITeC e.V, 22527 Hamburg, Germany
3 Department of Chemistry, University of Bergen, 5020 Bergen, Norway; christoph.bauer@uib.no
4 Computational Biology Unit (CBU), University of Bergen, 5020 Bergen, Norway
5 Front End Innovation, Beiersdorf AG, 20253 Hamburg, Germany; andreas.schepky@beiersdorf.com (A.S.); jochen.kuehnl@beiersdorf.com (J.K.)
* Correspondence: kirchmair@zbh.uni-hamburg.de; Tel.: +49-40-42838-7303

Received: 27 August 2019; Accepted: 18 September 2019; Published: 28 September 2019

Abstract: The ability to predict the skin sensitization potential of small organic molecules is of high importance to the development and safe application of cosmetics, drugs and pesticides. One of the most widely accepted methods for predicting this hazard is the local lymph node assay (LLNA). The goal of this work was to develop in silico models for the prediction of the skin sensitization potential of small molecules that go beyond the state of the art, with larger LLNA data sets and, most importantly, a robust and intuitive definition of the applicability domain, paired with additional indicators of the reliability of predictions. We explored a large variety of molecular descriptors and fingerprints in combination with random forest and support vector machine classifiers. The most suitable models were tested on holdout data, on which they yielded competitive performance (Matthews correlation coefficients up to 0.52; accuracies up to 0.76; areas under the receiver operating characteristic curves up to 0.83). The most favorable models are available via a public web service that, in addition to predictions, provides assessments of the applicability domain and indicators of the reliability of the individual predictions.

Keywords: skin sensitization potential; prediction; in silico models; machine learning; local lymph node assay (LLNA); cosmetics; drugs; pesticides; chemical space; applicability domain

1. Introduction

Repeated exposure to reactive chemicals with skin-sensitizing properties can cause allergic contact dermatitis (ACD) [1], an adverse cutaneous condition with a prevalence of ~20% among the general population [2] and even higher prevalence among workers with chronic occupational exposure [3]. Understanding the skin sensitization potential of small organic molecules is therefore of essence to the development and safe application of chemicals, including cosmetics and drugs.

Historically, animal tests have effectively been the only method for determining the skin sensitization potential and potency of substances. The local lymph node assay (LLNA) is currently considered to be the most advanced animal testing system [4]. In recent years, ethical considerations and regulatory requirements have led to an intensification of the search for alternatives to animal testing, in particular in the cosmetics industry [5]. New in vitro and in chemico methods have been developed and evaluated [6–9], and computational approaches are starting to be recognized as

important alternatives to animal testing [8–11]. The non-redundant combinatorial use of said methods in defined approaches that assess several key events of the adverse outcome pathway (AOP) for skin sensitization shows promising predictive capacity [12] and is currently evaluated in risk assessment case studies.

The bottleneck in the development of in silico tools for the prediction of skin sensitization is not related to technology but to the scarcity of available high-quality experimental data for model development. Three strategies have been pursued to address this problem. The first one is to increase the amount and coverage of data by employing data mining techniques to retrieve information from various types of assays and sources [13,14]. Although this has been discussed as a promising strategy to increase the applicability of models, it has also prompted controversial discussions regarding the quality and relevance of the data [15,16]. The second strategy is to develop focused models based on small, focused data sets of high-quality [17–21]. The third strategy is to pursue a middle way that aims for a favorable balance between quantity and quality of the data. The LLNA data available in the public domain are generally regarded as the most suitable source of information for this strategy [22–27].

The two largest curated collections of LLNA outcomes in the public domain are the data collections of Alves et al. [28] and Di et al. [22]. The data were obtained from reliable sources and subjected to deduplication procedures that reject discordant records. The data set of Alves et al. includes (mainly) binary LLNA outcomes recorded for 1000 compounds. In addition, it contains human data and outcomes from different types of in vitro and in chemico assays, although for substantially fewer substances. Based on these data, the authors developed machine learning models for different assay types and also a consensus model, all of which are available via an online platform ("PredSkin") [19]. Their model for the prediction of binary LLNA outcomes reached a correct classification rate (CCR) of 0.77 during five-fold external cross-validation.

The data set published by Di et al. contains 1007 substances annotated with LLNA potency classes [22]. Based on a subset of approximately 400 compounds for which an explicit reaction mechanism could be derived with a structural alerts tool for protein binding implemented in the OECD Toolbox [29], Di et al. developed a variety of models for the binary and ternary prediction of the skin sensitization potential. These models included local models for four reaction domains as well as global models. The best binary global model was reported to obtain an accuracy (ACC) of 0.84 during cross-validation and an ACC of 0.81 on a test set.

Major challenges in the application of machine learning approaches for risk assessment are related to the complexity of models that goes along with limited mechanistic interpretability. For these types of models, transparency with respect to the applicability domain as well as the provision of confidence estimates for individual predictions are of utmost importance to risk assessors, who ultimately are the main stakeholders of these methods.

In this context, and building on the works of Alves et al. and Di et al., this study pursues four main objectives to advance in silico capabilities for the prediction of the skin sensitization potential: (i) the development of a detailed understanding of the chemical space covered by the available LLNA data with respect to the chemical space of cosmetics, approved drugs and pesticides, (ii) the identification of the most suitable (sets of) molecular descriptors for modeling, (iii) the maximization of the applicability of the models by increasing the size and coverage of the data set used for model development, (iv) the definition of robust measures of the models' applicability domain as well as the provision of indicators for the reliability of individual predictions, and (v) the provision of the most suitable models via a public web service.

2. Results

2.1. Characterization of the LLNA Data Sets

In order to develop a detailed understanding of the relevance of the available LLNA data to modeling the skin sensitization potential of xenobiotics, we analyzed the composition and molecular

diversity of the LLNA data sets of Alves et al. and Di et al. In addition, we assessed how well the individual LLNA data sets cover the chemical space of cosmetics, approved drugs and pesticides.

2.1.1. Data Set Composition

Whereas the data set compiled by Alves et al. is balanced (481 sensitizers; 519 non-sensitizers), the data set of Di et al. contains almost twice as many non-sensitizers (n = 629) as sensitizers (n = 364; Table 1). Roughly 40% of all compounds (567) are present in both data sets (Table 2). The LLNA data set compiled by Alves et al. contains 7% of all substances listed in the cosmetics data set; coverage is lower for approved drugs and pesticides (4% and 5%, respectively). The percentages are similar for the LLNA data set of Di et al.: 5% overlaps with cosmetics, 3% with approved drugs and 4% with pesticides. Merging the two LLNA data sets increases the number of unique compounds to 1416 and the overlaps with cosmetics, approved drugs and pesticides to 8%, 5% and 5%, respectively.

2.1.2. Coverage of Chemical Space

Whereas only few of the cosmetics, approved drugs and pesticides listed in the reference data sets are included in the LLNA data sets, principal component analysis (PCA) shows that the areas in chemical space most densely populated with these xenobiotics are actually well-covered by the merged LLNA data set (Figure 1). Nevertheless, scattered data points radiating from the area of high data density towards the bottom and the top right corner of the PCA score plot indicate the existence of drugs and cosmetic compounds without closely related substances listed in the merged data set.

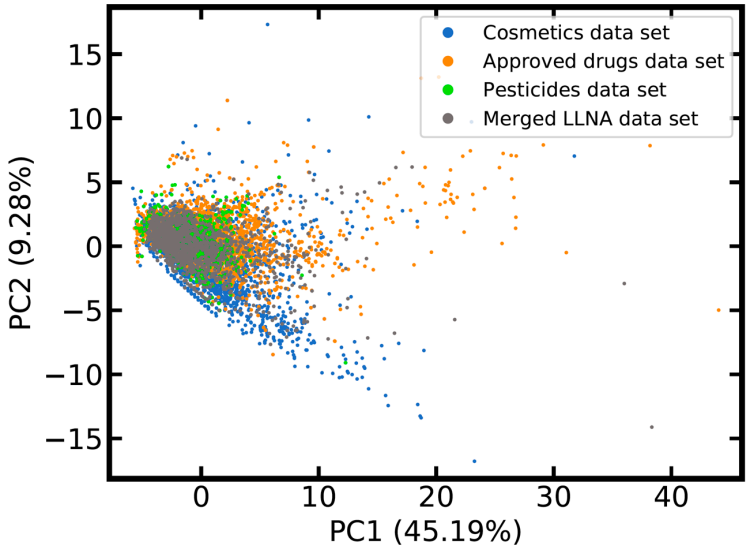


Figure 1. Score plot comparing the chemical space of compounds of the merged LLNA data set, cosmetics, approved drugs and pesticides. The plot is derived from a principal component analysis (PCA) based on 53 intuitive and physically meaningful molecular descriptors such as molecular weight and clogP (see Methods and Table S1 for details). Data points located in the lower parts of the PCA score plot are primarily cosmetics with long aliphatic and often halogenated chains; towards the top right corner of the diagram these are primarily large drug molecules with strong aromatic components. The variance explained by the first two principal components is reported in the axis titles. Four compounds of the cosmetics reference set and eight compounds of the approved drugs reference set are not shown because they are off the chosen limits of the plot (these are complex and large molecules, with a molecular weight of 2800 Da and higher).

Table 1. Overview of all data sets used in this work.

	LLNA Data Set Compiled by Alves et al.	**LLNA Data Set Compiled by Di et al.**	**Merged LLNA Data Set**	**Cosmetic Substances and Ingredients Data Set**	**Approved Drugs Data Set**	**Pesticides Data Set**
Data source	Chembench [30][1]	Supporting information of Di et al. [22]	LLNA data sets of Alves et al. and Di et al.	CosIng Database [31]	"Approved Drugs" subset of DrugBank [32,33][2]	EU Pesticides Database [34]
Number of compounds prior to data preprocessing	1000	1007	1993	5937	2352	1383
Number of compounds after data preprocessing	1000	993[3]	1416[4] (1132/284)[5]	4643[6]	2155[7]	812[8]
Number of sensitizers	481	364	572 (457/115)[5]	*n*/a	*n*/a	*n*/a
Number of non-sensitizers	519	629	844 (675/169)[5]	*n*/a	*n*/a	*n*/a
Number of Murcko scaffolds	312	354	453	856	1158	329
Proportion of compounds without a Murcko scaffold	0.32	0.29	0.31	0.42	0.13	0.24
Proportion of singleton scaffolds	0.77	0.79	0.78	0.72	0.82	0.81

[1] Chapel Hill, NC, United States. [2] Edmonton, Alberta, Canada. [3] Thirteen compounds were removed as part of the deduplication procedure; one compound was removed because of conflicting activity assignments. [4] Five hundred and sixty-seven compounds were removed as part of the deduplication procedure; ten compounds were removed because of conflicting activity assignments. [5] Number of compounds in the training set/test set prior to descriptor calculation. [6] One hundred and four compounds were removed by the salt filter because the main component could not be unambiguously identified; 26 compounds were removed due to invalid input structure; 1164 compounds were removed as part of the deduplication procedure. [7] Thirty-one compounds were removed by the salt filter because the main component could not be unambiguously identified; 166 compounds were removed as part of the deduplication procedure. [8] The SMILES notation of 893 compounds present in the EU Pesticides Database were automatically retrieved with the Chemical Identifier Resolver [35]. Six compounds were removed by the salt filter because the main component could not be identified; 13 compounds were removed due to invalid input structure; 62 compounds were removed as part of the deduplication procedure. Abbreviations: LLNA, local lymph node assay.

Table 2. Overlaps between the compounds contained in the LLNA data sets and the cosmetics, approved drugs and pesticides data sets.

	Number of Compounds	**Data Set Compiled by Alves et al.**	**Data Set Compiled by Di et al.**	**Merged LLNA Data Set**
Cosmetics	4643	324	252	387
Approved Drugs	2155	88	68	97
Pesticides	812	43	34	44

Abbreviations: LLNA, local lymph node assay.

In addition to PCA analysis, the coverage of cosmetics, approved drugs and pesticides by the merged LLNA data set was quantified based on the distribution of maximum pairwise similarities. As shown in Figure 2, the merged LLNA data set covers cosmetics much better than approved drugs and pesticides: over 30% of all cosmetics are represented by the respective nearest neighbor in the merged LLNA data set with a minimum Tanimoto coefficient of 0.6, whereas this is the case for only 10% and 13% of all approved drugs and pesticides, respectively.

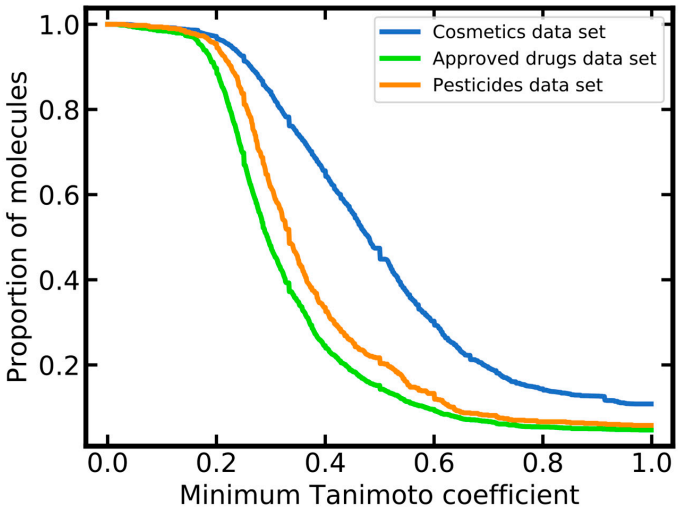


Figure 2. Molecular similarity between each compound of the reference data sets (i.e., cosmetics, approved drugs and pesticides data sets) and its nearest neighbor in the merged local lymph node assay (LLNA) data set (similarity quantified as Tanimoto coefficient based on Morgan2 fingerprints with a length of 2048 bits).

It is important to note that the data set compiled by Di et al. includes many compounds that populate areas in chemical space not (well) covered by the LLNA data set of Alves et al. (Figure 3). It is therefore expected that models trained on the merged data set should be more widely applicable than those based solely on the LLNA data compiled by Alves et al.

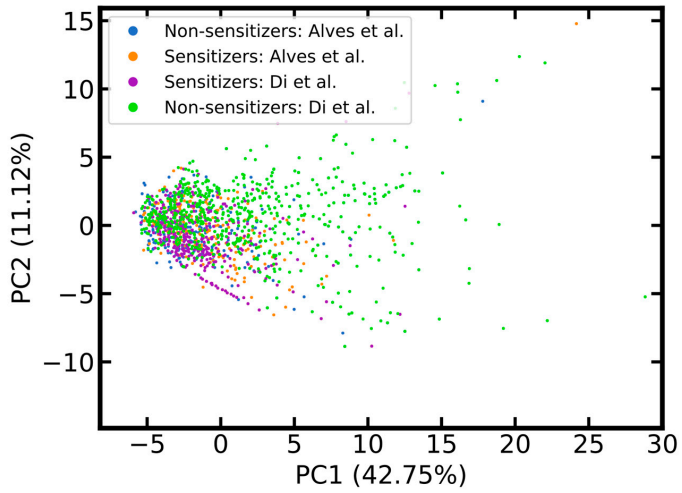


Figure 3. Score plot comparing the chemical space of compounds of the local lymph node assay (LLNA) data sets of Alves et al. and Di et al. The score plot was derived from a PCA based on the identical setup described in the caption of Figure 1. Two data points are located outside the displayed intervals.

2.1.3. Molecular Diversity

The molecular diversity of the merged LLNA data set and the reference data sets was assessed in two different ways: by pairwise comparison of molecular structures and by counting of Murcko scaffolds. Pairwise comparisons were again based on Tanimoto coefficients derived from Morgan2 fingerprints of a length of 2048 bits. The cosmetics data set exhibits a lower diversity compared to the other data sets (Figure 4). This can be attributed, to some extent, to the larger size of the cosmetics data set: 23% of all pairs of compounds in the cosmetic data set have fingerprints with a Tanimoto coefficient of 0.8 or higher, whereas this percentage is 11% or lower for the merged LLNA, approved drugs and pesticides data sets. Of all compounds included in the cosmetics data set, 220 have at least one neighbor with identical molecular fingerprint. These are mostly pairs of molecules with long aliphatic chains, differing only by the length of these chains (note that any duplicate molecules have been removed during data preprocessing).

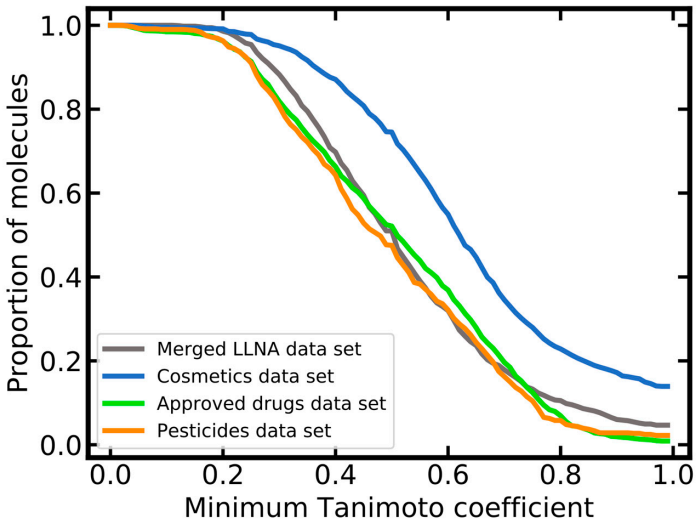


Figure 4. Pairwise molecular similarity within the individual data sets (similarity quantified as Tanimoto coefficient based on Morgan2 fingerprints with a length of 2048 bits).

The merged LLNA data set covers a total of 453 distinct Murcko scaffolds, which is roughly as many as covered by the pesticides data set but only one-third and one-quarter of those covered by the cosmetics and approved drugs data sets, respectively (Table 1). Taking into account the size of the individual data sets, the approved drugs data set clearly is the most diverse data set. In contrast, the cosmetics data set, which counts more molecular structures than all other data sets taken together, is the least diverse data set. This is in part related to the fact that approximately 40% of all cosmetics do not include a ring and, as such, do not have a Murcko scaffold.

Benzene is the most prominent Murcko scaffold across all data sets, with a prevalence of 27%, 28%, 10% and 23% among the merged LLNA, cosmetics, approved drugs and pesticides data sets. Any other scaffolds are represented by only a few instances (Table S2). Note the high percentages of singleton scaffolds (72% or higher) across all data sets, which, particularly in the case of the LLNA data set, illustrate the scarcity of the data available for modeling.

2.2. Molecular Properties of Skin Sensitizers and Non-Sensitizers

The merged LLNA data set contains 572 skin sensitizers and 844 non-sensitizers. As shown in Figure 5a, non-sensitizers cover a broader chemical space than sensitizers. A substantial number of non-sensitizers are of higher molecular weight than sensitizers and have a stronger aromatic character and larger topological polar surface area (Figure 5a,d). A cluster of skin sensitizers and non-sensitizers

with long aliphatic and halogenated chains was identified, observed as a diagonal line in the lower left of the score plot (Figure 5a,c). Interestingly, the compounds of this cluster can only be discriminated in the "MOE 2D" descriptor space but not in the Morgan2 fingerprint space, since molecules with identical halogen substitution but differing chain lengths can result in identical Morgan2 fingerprints.

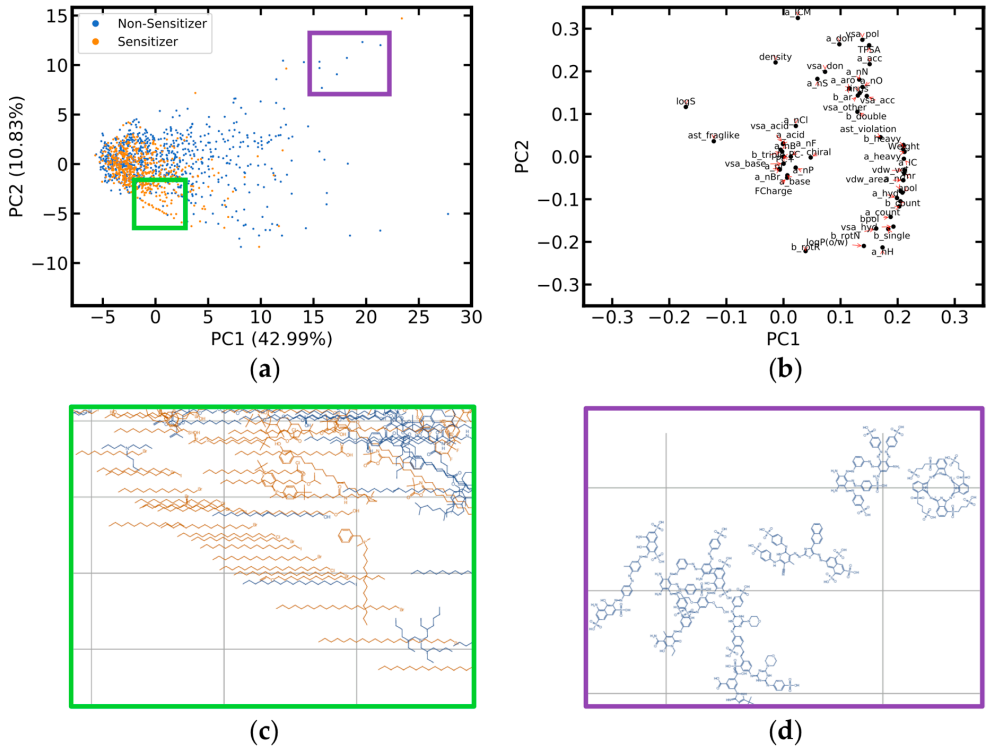


Figure 5. Principal component analysis (PCA) of the physicochemical properties of skin sensitizers and non-sensitizers included in the merged local lymph node assay (LLNA) data set. The PCA is based on the identical setup described in the caption of Figure 1. (**a**) Score plot, with the percentage of variance explained by the individual principal components reported as part of the axis labels. Two data points are located outside the displayed intervals. (**b**) Loadings plot (an enlarged version is provided in Figure S1; the abbreviations of the individual molecular descriptors are explained in Table S1). (**c**) Detailed view of the lower left region of the score plot, where mainly sensitizers are observed to form a line of data points. These sensitizers are aliphatic, monohalogenated hydrocarbons that differ primarily by chain length and halogen atom type. (**d**) Detailed view of the upper right part of the score plot, where mainly non-sensitizing compounds are located, characterized by high molecular weight, aromaticity and a large topological polar surface area.

2.3. Model Development

Prior to model development, the merged LLNA data set was divided into a training (80%) and test (20%) set (Table 3; see Methods for details). All possible combinations of machine learning approaches (random forest (RF) and support vector machine (SVM)) with up to two different sets of molecular descriptors (including molecular fingerprints) were systematically explored (Table 4). One type of descriptors to highlight is a new fingerprint that we derive from the "Protein binding alerts for skin sensitization by OASIS" profiler implemented in the OECD toolbox [29]. This profiler assigns compounds to eleven mechanistic domains associated with skin sensitization, five of which are represented by more than 20 instances in the training set (i.e., Michael addition, S_N2 reaction, Schiff base formation, acylation, and nucleophilic addition). The new fingerprint encodes the presence or absence of alerts matching one or several of these five mechanistic domains.

Table 3. Overview of descriptor sets evaluated in this work.

Descriptor set	Short Name	Number of Descriptors/Length of the Fingerprint	Calculated with	Number of Successfully Processed Molecules[1]	
				Training set	Test set
0D, 1D and 2D descriptors	MOE2D	206	MOE [36]; this set corresponds to all descriptors listed as "2D descriptors" in MOE	1132	284
Selection of 0D, 1D and 2D descriptors	MOE2D_53	53[2]	MOE [36]	1132	284
0D, 1D and 2D descriptors	PaDEL	1444	PaDEL [37,38]; this is the complete set of 0D, 1D and 2D descriptors implemented in PaDEL	1109	279
MACCS keys	MACCS	166	RDKit [39]	1132	284
Morgan2 fingerprints	Morgan2	2048	RDKit [39]	1132	284
OASIS skin sensitization protein binding fingerprint	OASIS	5 bit fingerprint	OECD Toolbox [29]	1128	283
PaDEL estate fingerprint	PaDEL_Est	79	PaDEL [37,38]	1132	284
PaDEL extended fingerprint	PaDEL_Ext	1024	PaDEL [37,38]	1132	284

[1] Descriptor calculation failed for individual compounds depending on the software used. For this reason, there are marginal differences in the composition of the individual data sets used for model development. [2] Fifty-three manually selected, physically meaningful descriptors. A list of the selected descriptors can be found in Table S1. Abbreviations: MOE, Molecular Operating Environment.

Table 4. Overview of models and their performance during cross-validation.

Name	Number of Descriptors	Number of Compounds in Training Data	ACC	ACC STDEV	MCC	MCCSTDEV	AUC	CCR	Se	SP	PPV	NPV
SVM_MOE2D+OASIS	211	1128	0.78	0.054	0.55	0.109	0.83	0.78	0.77	0.78	0.71	0.83
SVM_PaDEL+MACCS	1610	1108	0.76	0.035	0.51	0.069	0.83	0.76	0.75	0.76	0.69	0.82
SVM_PaDEL+Morgan2	3492	1108	0.76	0.036	0.51	0.078	0.82	0.75	0.66	0.83	0.73	0.78
SVM_PaDEL+PaDEL-Ext	2468	1109	0.76	0.039	0.51	0.075	0.84	0.76	0.74	0.78	0.7	0.81
SVM_MOE2D+MACCS	372	1132	0.76	0.047	0.5	0.096	0.81	0.74	0.68	0.81	0.71	0.79
SVM_MOE2D+Morgan2	2254	1132	0.75	0.041	0.5	0.081	0.83	0.75	0.77	0.73	0.66	0.83
SVM_MOE2D+PaDEL	1680	1109	0.76	0.039	0.5	0.079	0.83	0.75	0.74	0.77	0.69	0.81
SVM_MOE2D+PaDEL-Est	285	1132	0.76	0.039	0.5	0.081	0.81	0.75	0.68	0.81	0.71	0.79
SVM_MOE2D+PaDEL-Ext	1230	1132	0.75	0.054	0.5	0.105	0.83	0.75	0.75	0.76	0.68	0.81
SVM_PaDEL	1444	1109	0.75	0.038	0.5	0.075	0.83	0.75	0.75	0.75	0.68	0.81
SVM_PaDEL+OASIS	1449	1109	0.75	0.038	0.5	0.075	0.83	0.75	0.75	0.75	0.68	0.81
SVM_PaDEL+PaDEL-Est	1523	1109	0.75	0.038	0.5	0.075	0.83	0.75	0.75	0.75	0.68	0.81
RF_PaDEL+MACCS	1610	1108	0.76	0.018	0.49	0.037	0.82	0.73	0.62	0.85	0.74	0.77
RF_PaDEL+Morgan2	3492	1108	0.76	0.02	0.49	0.042	0.82	0.74	0.64	0.84	0.73	0.77
RF_PaDEL+OASIS	1449	1109	0.76	0.02	0.49	0.043	0.82	0.74	0.62	0.85	0.74	0.77
RF_PaDEL+PaDEL-Ext	2468	1109	0.76	0.022	0.49	0.048	0.82	0.73	0.61	0.86	0.75	0.76
SVM_PaDEL-Est+MACCS	245	1132	0.75	0.051	0.49	0.106	0.81	0.74	0.69	0.8	0.7	0.79
RF_MOE2D+PaDEL	1680	1109	0.75	0.034	0.48	0.072	0.83	0.73	0.62	0.84	0.73	0.77
RF_Morgan2+PaDEL-Est	2127	1132	0.76	0.033	0.48	0.071	0.82	0.73	0.63	0.84	0.73	0.77
RF_PaDEL	1444	1109	0.75	0.015	0.48	0.033	0.82	0.73	0.62	0.84	0.73	0.76
RF_PaDEL-Est+OASIS	84	1128	0.75	0.043	0.48	0.091	0.8	0.74	0.65	0.82	0.72	0.78
SVM_MACCS+OASIS	171	1128	0.75	0.047	0.48	0.102	0.82	0.74	0.69	0.79	0.69	0.79
SVM_MOE2D	206	1132	0.74	0.037	0.48	0.067	0.82	0.74	0.75	0.74	0.66	0.82
SVM_Morgan2+PaDEL-Ext	3072	1132	0.75	0.044	0.48	0.09	0.82	0.74	0.68	0.8	0.7	0.79
RF_MACCS	166	1132	0.75	0.039	0.47	0.088	0.81	0.73	0.61	0.84	0.73	0.76
RF_MACCS+OASIS	171	1128	0.75	0.034	0.47	0.074	0.8	0.73	0.6	0.85	0.74	0.76
RF_PaDEL+PaDEL-Est	1523	1109	0.75	0.028	0.47	0.06	0.83	0.73	0.61	0.85	0.73	0.76
SVM_MACCS	166	1132	0.74	0.057	0.47	0.12	0.81	0.73	0.69	0.78	0.68	0.79
SVM_PaDEL-Est+OASIS	84	1128	0.74	0.048	0.47	0.099	0.8	0.74	0.71	0.76	0.67	0.8
SVM_PaDEL-Est+PaDEL-Ext	1103	1132	0.74	0.039	0.47	0.08	0.81	0.74	0.7	0.78	0.68	0.79
SVM_PaDEL-Ext	1024	1132	0.74	0.046	0.47	0.093	0.81	0.73	0.7	0.77	0.68	0.79
SVM_PaDEL-Ext+OASIS	1029	1128	0.74	0.036	0.47	0.072	0.82	0.74	0.7	0.77	0.68	0.79
RF_MOE2D+Morgan2	2254	1132	0.74	0.033	0.46	0.071	0.81	0.72	0.62	0.82	0.71	0.76
RF_PaDEL-Est+MACCS	245	1132	0.75	0.045	0.46	0.1	0.81	0.72	0.59	0.85	0.73	0.76

Table 4. *Cont.*

Name	Number of Descriptors	Number of Compounds in Training Data	ACC	ACC STDEV	MCC	MCCSTDEV	AUC	CCR	Se	SP	PPV	NPV
RF_Morgan2	2048	1132	0.74	0.039	0.46	0.081	0.81	0.73	0.64	0.81	0.7	0.77
SVM_Morgan2+MACCS	2214	1132	0.74	0.058	0.46	0.117	0.8	0.73	0.68	0.78	0.68	0.78
SVM_PaDEL-Ext+MACCS	1190	1132	0.74	0.047	0.46	0.097	0.81	0.73	0.68	0.77	0.68	0.78
RF_MOE2D+OASIS	211	1128	0.74	0.041	0.45	0.09	0.81	0.71	0.6	0.83	0.71	0.75
RF_MOE2D+PaDEL-Est	285	1132	0.74	0.032	0.45	0.07	0.81	0.72	0.6	0.84	0.72	0.75
RF_MOE2D+PaDEL-Ext	1230	1132	0.74	0.017	0.45	0.037	0.82	0.72	0.58	0.85	0.73	0.75
RF_MOE2D	206	1132	0.73	0.036	0.44	0.078	0.81	0.71	0.59	0.83	0.71	0.75
RF_MOE2D+MACCS	372	1132	0.73	0.033	0.44	0.072	0.81	0.71	0.58	0.84	0.71	0.75
RF_Morgan2+MACCS	2214	1132	0.73	0.039	0.44	0.086	0.8	0.72	0.63	0.8	0.68	0.76
RF_Morgan2+OASIS	2053	1128	0.74	0.029	0.44	0.063	0.82	0.71	0.59	0.83	0.71	0.75
RF_Morgan2+PaDEL-Ext	3072	1132	0.73	0.036	0.44	0.081	0.81	0.71	0.56	0.85	0.72	0.74
SVM_MOE2D53	53	1132	0.71	0.037	0.44	0.069	0.78	0.72	0.76	0.68	0.62	0.81
SVM_PaDEL-Est	79	1132	0.72	0.037	0.44	0.073	0.77	0.72	0.71	0.73	0.64	0.79
RF_PaDEL-Est	79	1132	0.73	0.022	0.43	0.042	0.77	0.71	0.64	0.79	0.67	0.76
RF_PaDEL-Ext+MACCS	1190	1132	0.73	0.037	0.43	0.081	0.81	0.7	0.55	0.85	0.72	0.74
RF_PaDEL-Ext+OASIS	1029	1128	0.73	0.033	0.43	0.072	0.8	0.7	0.57	0.84	0.71	0.74
RF_PaDEL-Ext+PaDEL-Est	1103	1132	0.73	0.034	0.43	0.074	0.8	0.7	0.56	0.85	0.72	0.74
SVM_Morgan2+OASIS	2053	1128	0.73	0.038	0.43	0.089	0.8	0.69	0.51	0.88	0.75	0.73
SVM_Morgan2+PaDEL-Est	2127	1132	0.72	0.035	0.43	0.064	0.79	0.72	0.69	0.75	0.65	0.78
RF_MOE2D53	53	1132	0.73	0.039	0.42	0.086	0.78	0.7	0.58	0.83	0.69	0.74
RF_PaDEL-Ext	1024	1132	0.72	0.039	0.42	0.088	0.79	0.7	0.55	0.84	0.71	0.73
SVM_Morgan2	2048	1132	0.72	0.031	0.39	0.072	0.8	0.68	0.49	0.87	0.72	0.71
SVM_OASIS	5	1128	0.67	0.064	0.29	0.151	0.63	0.62	0.37	0.87	0.68	0.67
RF_OASIS	5	1128	0.66	0.054	0.27	0.122	0.64	0.63	0.43	0.82	0.62	0.68

Abbreviations: ACC, accuracy; AUC, area under the receiver operating characteristic curve; CCR, correct classification rate; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; Se, sensitivity; Sp, specificity; STDEV, standard deviation.

For any combination of machine learning algorithm and descriptor set(s), optimum hyperparameters were identified via a grid search (Table 5). The grid search was performed within the framework of a 10-fold cross-validation, with Matthews correlation coefficient (MCC) [40] used as the scoring parameter.

Table 5. Overview of hyperparameters optimized by grid search.

Machine Learning Approach	Parameter	Explored Values
RF	*n*_estimators [1]	10, 50, 100, 250, 500, 1000
	max_features [2]	'sqrt', 0.2, 0.4, 0.6, 0.8, None
SVM	C [3]	0.01, 0.1, 1, 10, 100, 1000
	gamma [4]	1, 0.1, 0.01, 0.001, 0.0001, 0.00001

[1] Number of prediction trees. [2] Maximum depth of each tree. [3] Penalty parameter C of the error term. [4] Coefficient for the radial basis function (rbf) kernel. Abbreviations: RF, random forest; SVM, support vector machine.

The outcomes of this grid search are summarized in Table S3. It can be seen that similar hyperparameters tend to be selected by models based on related types and sets of molecular descriptors. No strong preferences for specific hyperparameter values are apparent. This is likely related to the fact that, within a broad value space, the hyperparameters only had a minor impact on model performance.

2.4. Model Performance

2.4.1. Measures for the Evaluation of Model Performance

Eight different measures were applied to describe the performance of the classifiers:

- Matthews correlation coefficient (MCC), which is regarded to be one of the best measures of binary classification performance. It is robust against data imbalance and considers the proportion of all four cases of predictions (i.e., true positive, false positive, true negative and false negative predictions). Note that MCC values range from −1 to +1. A value of +1 indicates perfect prediction, whereas a value of −1 indicates a prediction that is in total disagreement. A value of 0 indicates a performance which is equal to random.
- ACC, which has been most commonly used by others to measure the performance of models for the prediction of the skin sensitization potential. It is defined as the proportion of correct predictions within all predictions made.
- Area under the receiver operating characteristic curve (AUC), which in this case quantifies the ability to correctly rank compounds according to their skin sensitization potential. The AUC does not rely on a decision threshold.
- Sensitivity (Se), which in this case quantifies the proportion of correctly identified skin sensitizers.
- Specificity (Sp), which in this case quantifies the proportion of correctly predicted non-sensitizers.
- Positive predictive value (PPV), which reports the proportion of true positive predictions among all positive predictions.
- Negative predictive value (NPV), which reports the proportion of true negative predictions among all negative predictions.
- CCR, which is the mean of Se and Sp.

2.4.2. Model Performance During Cross-Validation

Depending on the combination of machine learning algorithm (RF or SVM) and descriptor set(s) used, MCC values ranged from 0.27 to 0.55, ACC values from 0.66 to 0.78, and AUC values from 0.63 to 0.84 (Table 4). The machine learning algorithms had only a minor impact on model performance. The average MCC values obtained by RFs and SVMs were 0.45 and 0.48, respectively. Nevertheless, the twelve predictors that obtained the highest MCC values are all based on SVMs. Most of the observed variation in performance stemmed from the use of different descriptor sets.

The best performance during cross-validation was obtained by the SVM_MOE2D+OASIS model. This model yielded an MCC, ACC and AUC of 0.55, 0.78 and 0.83, respectively. The best model based on a single set of descriptors was the SVM_PaDEL model. It reached an MCC, ACC and AUC of 0.50, 0.75 and 0.83, respectively. However, its lead over the corresponding RF model and other models based on a single set of descriptors was small. For example, the best model based on a single type of molecular fingerprint, RF_MACCS, obtained an MCC, ACC and AUC of 0.47, 0.75 and 0.81, respectively. Models based on either machine learning algorithm in combination with "MOE 2D" descriptors or MACCS fingerprints yielded comparable performance. Reduction of the full MOE2D descriptor set to the subset of 53 interpretable MOE descriptors (previously used for analyzing the chemical space coverage) led to a decline in MCC values by a maximum of 0.04. Caution needs to be exercised when interpreting these small differences in performance because of the variance observed during cross-validation. For example, for the SVM_MOE2D_53 model, the standard deviation observed for the MCC during cross-validation was 0.069.

In most cases, the combination of two sets of molecular descriptors was beneficial to model performance. Exceptions include models based on combinations of two sets of descriptors of the same type (e.g., Morgan2 and MACCS fingerprints). These did not outperform the best models based on a single set of descriptors. Also, combinations of 0D/1D/2D molecular descriptors with fingerprints did not consistently outperform models based on a single set of descriptors, albeit nine out of twelve models with MCC values greater than or equal to 0.5 are models combining non-binary molecular descriptors (i.e., MOE2D or PaDEL) with molecular fingerprints. Tables S4 and S5 provide a comprehensive overview of the impact of different combinations of descriptor sets on model performance.

Good performance was also obtained by models generated using non-commercial software only. For example, the SVM_PaDEL+OASIS model obtained MCC, ACC and AUC values of 0.50, 0.75 and 0.83, respectively. With few exceptions, the OASIS fingerprint contributed positively to the performance of models. For instance, adding the OASIS fingerprint to the SVM_MOE2D model led to an increase of the MCC, ACC and AUC by 0.07, 0.04 and 0.01, respectively. Interestingly, with a total of just 84 bits, the RF_PaDEL−Est+OASIS model reached a level of performance that is comparable with that of more complex models (MCC 0.48; ACC 0.75; AUC 0.80). However, when used on its own, the OASIS fingerprint is not sufficient for good classification performance: the RF_OASIS and SVM_OASIS models obtained the lowest MCC values across all models (i.e., 0.27 and 0.29, respectively).

2.4.3. In-Depth Analysis of Selected Models within the Cross-Validation Framework

Based on the cross-validation results, five of the most interesting models were selected for additional studies:

- SVM_MOE2D+OASIS: the model with highest MCC.
- SVM_PaDEL+OASIS: a model performing comparable to the SVM_MOE2D+OASIS and based on freely available software only.
- SVM_PaDEL: the best model based on a single set of molecular descriptors.
- RF_MACCS: the best model based on a single set of molecular fingerprints.
- SVM_PaDEL+MACCS: a model with good performance, combining the descriptor sets used by the above two models.

Within the above-mentioned 10-fold cross-validation framework, we first analyzed how the coverage of the query molecules by the training data affects model performance. For this analysis we calculated the similarity between the individual query molecules and the one, three and five-nearest neighbors in the training set. Two similarity measures were explored: Tanimoto coefficients in the MACCS fingerprint space and negative Euclidean distances in the PaDEL descriptor space. The latter did not correlate well with molecular similarity (likely caused by noise related to the large number of molecular descriptors considered in this approach; Figure S2 and Table S6), for which reason we decided to go ahead with the fingerprint-based distance measure.

For all five models, a direct linear relationship was observed between MCC values and molecular similarity. The relationship was consistent when considering different numbers of nearest neighbors in the training data but tended to be more robust when taking more (i.e., 5) nearest neighbors into account (Pearson correlation coefficient between 0.92 and 0.96 when considering five nearest neighbors). As shown in Figure 6, for compounds dissimilar to those present in the training data (defined by Tanimoto coefficients averaged over the five nearest neighbors of 0.5 or lower), MCC values were below or around 0.4 for all five models. For compounds structurally related to the training data (defined by Tanimoto coefficients of 0.7 or higher), MCC values were at least 0.5 or higher.

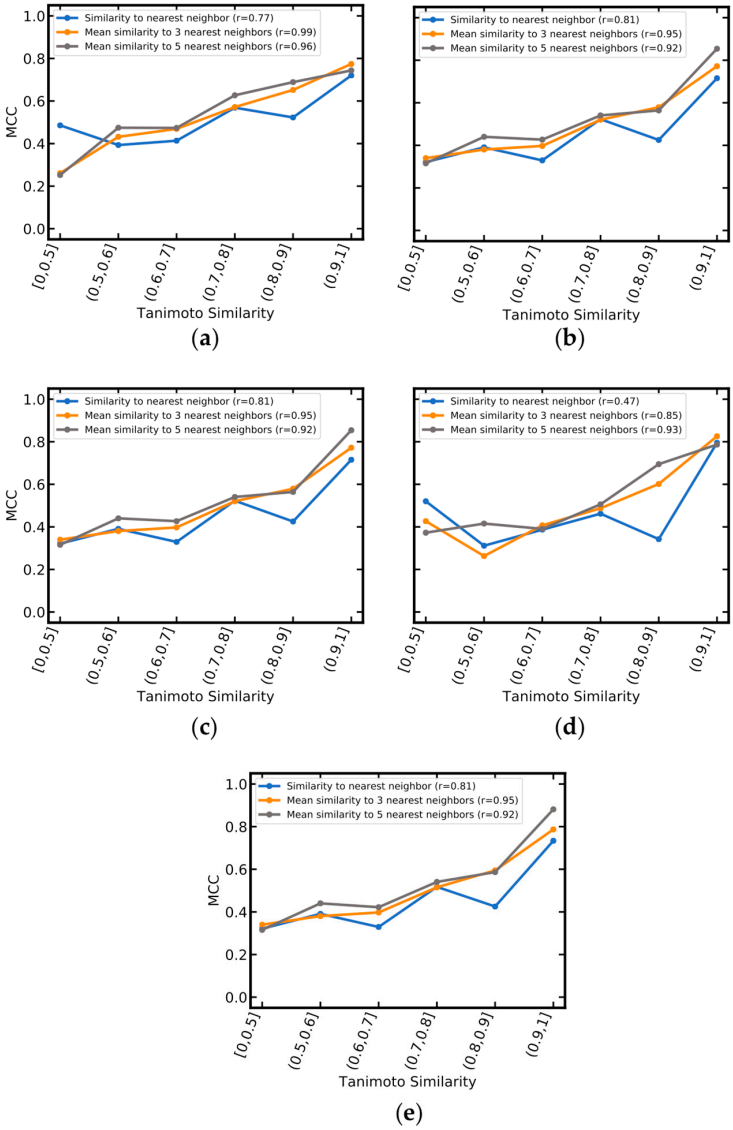


Figure 6. Matthews correlation coefficient (MCC) as a function of molecular similarity between the query compounds and the one, three and five nearest neighbors in the training data (calculated as averaged Tanimoto coefficients based on MACCS fingerprints). (**a**) SVM_MOE2D+OASIS; (**b**) SVM_PaDEL+OASIS; (**c**) SVM_PaDEL; (**d**) RF_MACCS; (**e**) SVM_PaDEL+MACCS. Pearson correlation coefficients are reported in brackets in the figure legends. The number of compounds in each bin is summarized in Table S7.

Secondly, we investigated how changes to the decision threshold of the SVM and RF classifiers (i.e., the value above which a compound is predicted to be a sensitizer) affect the sensitivity and specificity of the models. As shown in Figure 7, both these metrics strongly depend on the selected decision threshold. This allows users to define context-dependent thresholds. For example, in scenarios where for a compound of interest any skin sensitization potential should be ruled out, users may opt for lower decision thresholds to identify any hazard. In the case of the RF_MACCS model, lowering the decision threshold to 0.3 results in a sensitivity of 0.84 and a specificity of 0.61 (Figure 7d).

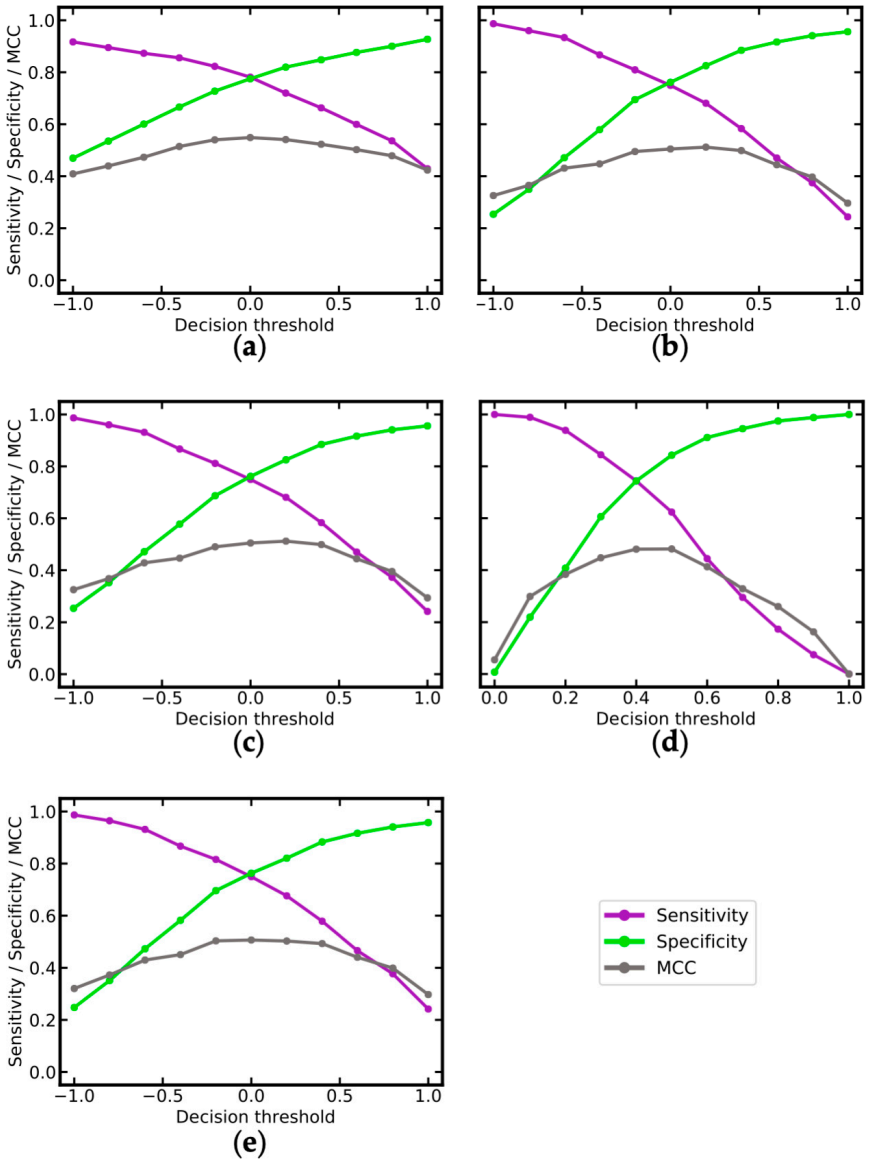


Figure 7. Matthews correlation coefficient (MCC), sensitivity and specificity as a function of the decision threshold, for (**a**) SVM_MOE2D+OASIS; (**b**) SVM_PaDEL+OASIS; (**c**) SVM_PaDEL; (**d**) RF_MACCS; (**e**) SVM_PaDEL+MACCS. Note that different *X*-axis scales are applied to the graphs illustrating the performance of random forest (RF) and support vector machine (SVM) models.

Observing the predicted class probability can be of use for assessing the reliability of a prediction: as shown in Figure 8, the reliability of predictions increases with the absolute distance between the

class probability and the decision threshold. For SVM models, predictions with class probabilities more than 0.5 away from the decision threshold had averaged MCC values between 0.63 and 0.67, whereas predictions with class probabilities less than 0.5 away had averaged MCC values of just 0.20 to 0.29. For the RF_MACCS model, predictions with class probabilities more than 0.35 away from the decision threshold had MCC values above 0.6, whereas predictions with class probabilities closer than 0.15 to the decision threshold had MCCs below 0.4. For the five investigated models, the Pearson correlation coefficients for this relationship were between 0.92 and 0.98.

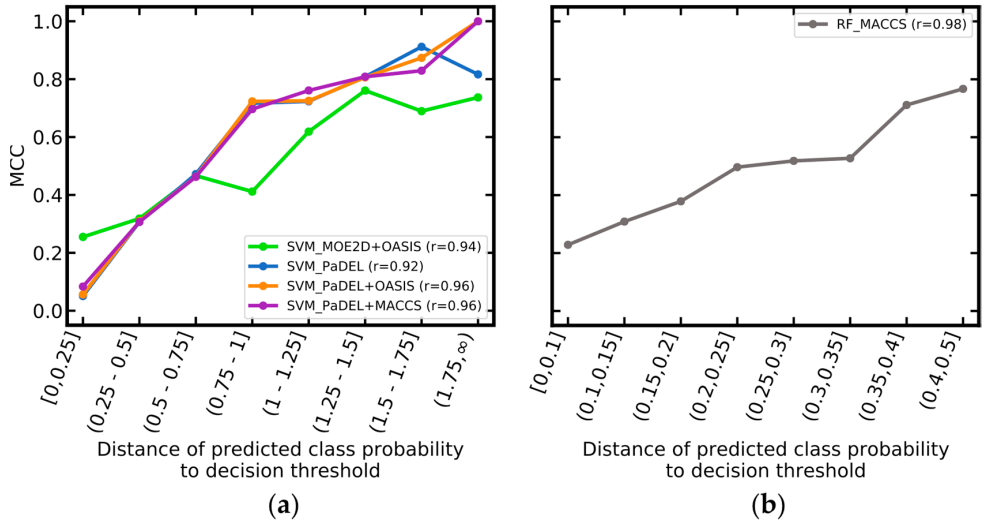


Figure 8. Matthews correlation coefficient (MCC) as a function of the distance between the predicted class probabilities and the decision thresholds, for the (**a**) support vector machine (SVM) models and (**b**) random forest (RF) model. The number of compounds in each bin is summarized in Table S8.

As a further way of analyzing the data, we looked into the reliability of predictions as a function of the number of consecutive nearest neighbors in the training data that are of the same activity class as the one predicted for a compound of interest. From Figure 9, it can be seen that predictions are particularly reliable if the three nearest neighbors in the training data are of the identical class as the class predicted for a compound of interest. The strongest correlation is observed for the RF_MACCS model. For this model the MCC is close to zero for compounds where the predicted class is in conflict with the class assigned to the nearest neighbor. In contrast, the MCC is above 0.6 for compounds where the predicted class and the classes assigned to the three nearest neighbors are identical.

2.4.4. Performance of Selected Models on the Test Set

The performance of the five selected models was tested on holdout data. All models were stable, with only minor losses in MCC, ACC and AUC when compared to the results from cross-validation (Table 6). The largest losses in performance were observed for the RF_MACCS model, with MCC and ACC values decreased by 0.06 and 0.03, respectively (AUC however +0.01).

By defining the applicability domain of the models to include any compounds with a minimum Tanimoto coefficient of 0.75 averaged over the five-nearest neighbors in the training set (based on MACCS fingerprints), MCC values increased, in the case of the RF_MACCS model from 0.41 to 0.59. However, at the same time the coverage of the test set is reduced, in the case of RF_MACCS to 28%.

Defining the applicability domain with a cutoff of 0.50 rather than 0.75 led to only minor performance improvements compared to the model without applicability domain definition. This is related to the fact that only approximately 3% of the compounds of the test set are that dissimilar to the compounds

in the training data. However, predictions for these compounds are unreliable (MCC values 0.2 or lower). Therefore, it is important to observe the applicability domain of the individual models.

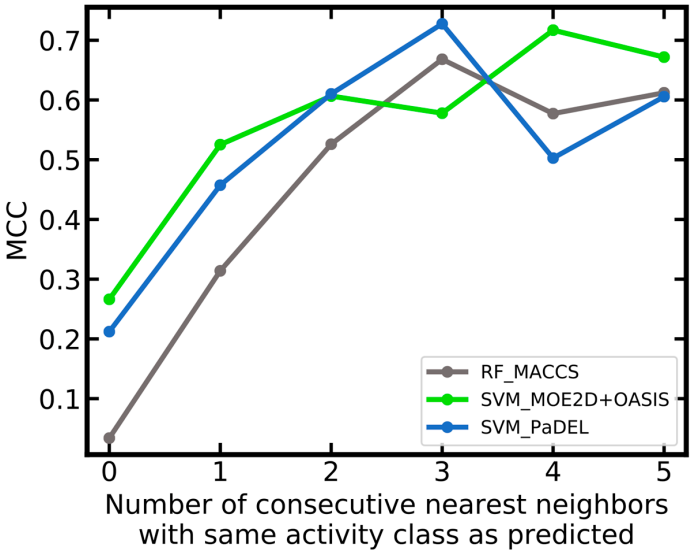


Figure 9. Matthews correlation coefficient (MCC) as a function of the number of consecutive nearest neighbors in the training data that are of the same activity class as the predicted class for a compound of interest (molecular similarity quantified as Tanimoto coefficient based on MACCS fingerprints). The number of compounds in each bin is summarized in Table S9. The graphs for SVM_PaDEL+OASIS and SVM_PaDEL+MACCS are not shown because they are (almost) identical with that of SVM_PaDEL and would overlap.

Table 6. Performance of selected models on the test set.

NAME	Mean Tanimoto Similarity to the Five Nearest Neighbors	Number of Compounds	ACC	MCC	AUC	CCR	Se	Sp	PPV	NPV
RF_MACCS	≥0	284	0.72	0.41	0.82	0.70	0.57	0.82	0.69	0.74
RF_MACCS	≥0.5	273	0.73	0.43	0.82	0.71	0.6	0.82	0.69	0.75
RF_MACCS	≥0.75	79	0.78	0.59	0.91	0.81	0.89	0.73	0.64	0.92
RF_MACCS	<0.5	11	0.45	−0.29	0.60	0.42	0.00	0.83	0.00	0.50
SVM_MOE_2D+OASIS	≥0	283	0.76	0.52	0.83	0.76	0.81	0.72	0.66	0.85
SVM_MOE_2D+OASIS	≥0.5	273	0.76	0.53	0.84	0.77	0.82	0.72	0.67	0.86
SVM_MOE_2D+OASIS	≥0.75	79	0.81	0.64	0.89	0.84	0.93	0.75	0.67	0.95
SVM_MOE2D+OASIS	<0.5	10	0.60	0.20	0.60	0.60	0.60	0.60	0.60	0.60
SVM_PaDEL	≥0	279	0.74	0.47	0.82	0.74	0.76	0.72	0.65	0.82
SVM_PaDEL	≥0.5	269	0.74	0.49	0.83	0.75	0.77	0.73	0.65	0.83
SVM_PaDEL	≥0.75	79	0.80	0.63	0.89	0.83	0.93	0.73	0.65	0.95
SVM_PaDEL	<0.5	10	0.60	0.20	0.56	0.60	0.60	0.60	0.60	0.60
SVM_PaDEL+MACCS	≥0	279	0.75	0.50	0.82	0.75	0.78	0.73	0.66	0.83
SVM_PaDEL+MACCS	≥0.5	269	0.75	0.51	0.83	0.76	0.79	0.73	0.66	0.84
SVM_PaDEL+MACCS	≥0.75	79	0.80	0.63	0.89	0.83	0.93	0.73	0.65	0.95
SVM_PaDEL+MACCS	<0.5	10	0.60	0.20	0.56	0.60	0.60	0.60	0.60	0.60
SVM_PaDEL+OASIS	≥0	279	0.74	0.48	0.82	0.74	0.76	0.73	0.65	0.82
SVM_PaDEL+OASIS	≥0.5	271	0.75	0.49	0.83	0.75	0.77	0.73	0.65	0.83
SVM_PaDEL+OASIS	≥0.75	79	0.80	0.63	0.89	0.83	0.93	0.73	0.65	0.95
SVM_PaDEL+OASIS	<0.5	10	0.60	0.20	0.56	0.60	0.60	0.6	0.60	0.60

Abbreviations: ACC, accuracy; AUC, area under the receiver operating characteristic curve; CCR, correct classification rate; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; Se, sensitivity; Sp, specificity.

Besides the applicability domain definition, users are advised to consider two additional types of information when judging the reliability of a prediction: (i) the distance between the predicted class

probability from the decision threshold and (ii) the number of consecutive nearest neighbors that are of the same activity class than the class predicted for a compound of interest.

Larger distances of the class probability to the decision threshold indicate higher reliability of the prediction. For example, when considering only predictions with class probabilities 0.35 or further away from the decision threshold, the MCC of the RF_MACCS model increases from 0.41 to 0.78 (this covers 23% of the test set; Table 7). Likewise, for the SVM models, MCC values increase from approximately 0.5 to a maximum of 0.78 when considering predictions only if their class probability is 1.25 or further away from the decision threshold (this covers 12% to 37% of the compounds in the test set).

Table 7. Test set performance as a function of the distance of predicted class probabilities from the decision threshold.

Name	Distance to Decision Threshold [1]	Number of Compounds	ACC	MCC	AUC	CCR	Se	Sp	PPV	NPV
RF-MACCS	≥0.15	175	0.85	0.67	0.46	0.84	0.81	0.87	0.76	0.90
RF-MACCS	≥0.35	66	0.91	0.78	0.42	0.89	0.85	0.93	0.85	0.93
RF-MACCS	<0.15	109	0.51	0.04	0.42	0.52	0.32	0.72	0.55	0.50
SVM_MOE2D+OASIS	≥0.5	203	0.82	0.64	0.42	0.83	0.88	0.78	0.73	0.90
SVM_MOE2D+OASIS	≥1.25	106	0.89	0.76	0.41	0.89	0.89	0.88	0.81	0.94
SVM_MOE2D+OASIS	<0.50	80	0.60	0.20	0.52	0.60	0.62	0.58	0.50	0.70
SVM_PaDEL	≥0.5	183	0.80	0.61	0.48	0.81	0.86	0.76	0.71	0.89
SVM_PaDEL	≥1.25	34	0.88	0.78	0.45	0.91	1.00	0.82	0.75	1.00
SVM_PaDEL	<0.50	96	0.61	0.21	0.36	0.60	0.55	0.66	0.51	0.69
SVM_PaDEL+MACCS	≥0.5	183	0.80	0.62	0.49	0.82	0.88	0.75	0.71	0.90
SVM_PaDEL+MACCS	≥1.25	37	0.86	0.75	0.52	0.9	1.00	0.80	0.71	1.00
SVM_PaDEL+MACCS	<0.50	96	0.65	0.27	0.39	0.63	0.58	0.69	0.55	0.71
SVM_PaDEL+OASIS	≥0.5	183	0.80	0.61	0.49	0.81	0.86	0.76	0.71	0.89
SVM_PaDEL+OASIS	≥1.25	34	0.88	0.78	0.45	0.91	1.00	0.82	0.75	1.00
SVM_PaDEL+OASIS	<0.50	96	0.62	0.22	0.37	0.61	0.55	0.67	0.52	0.70

[1] Distance of predicted class probabilities from the decision threshold. Abbreviations: ACC, accuracy; AUC, area under the receiver operating characteristic curve; CCR, correct classification rate; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; Se, sensitivity; Sp, specificity.

Predictions for query molecules that are consistent with the class assigned to the *k*-nearest neighbors in the training data are more reliable than for those that are in conflict. This is also confirmed by the results obtained for the test set (Table 8): Predictions that are in disagreement with the activity class of the nearest neighbor resulted in MCC and ACC values no higher than 0.13 and 0.56, respectively. MCC and ACC values increase to a maximum of 0.98 and 0.99 when considering predictions only if they are consistent with three or more nearest neighbors.

2.4.5. Comparison of Model Performance to that of Existing Models

Major caveats must be considered when attempting to directly compare the performance reported for existing models with those presented in this work. Not only do the underlying training and test sets differ substantially, but also the protocols used for performance evaluation and the definitions of the models' applicability domains. Roughly summarized, Alves et al. reported their predictor of binary LLNA outcomes to yield a CCR of 0.77 during external cross-validation [28]. Di et al. reported their best global model for the binary prediction of LLNA outcomes, a SVM model based on PaDEL-Ext descriptors (Ext-SVM), to have yielded an ACC of 0.84 during cross-validation and an ACC of 0.81 on their test set (when considering the applicability domain according to their definition) [22]. In comparison, our best model (SVM_MOE2D+OASIS) yielded a CCR of 0.78 and identical ACC during cross-validation (MCC 0.55), without consideration of the applicability domain. On the test set, the SVM_MOE2D+OASIS model obtained a CCR of 0.76 and an MCC of 0.52. In this case, the consideration of the applicability domain of the model (defined as including any compound with a mean Tanimoto similarity to the five nearest neighbors in the training set of 0.50 or higher) did not yield a further improvement of performance. The SVM_PaDEL and RF_MACCS models, which are

available via a public web service, yielded comparable CCR values (0.74 and 0.70 without consideration of the applicability domain; 0.75 and 0.71 with consideration of the applicability domain, respectively). The latter model has the additional benefit of being based on a fingerprint with a length of only 166 bits.

Table 8. Test set performance as a function of the number of consecutive nearest neighbors with class assignments consistent with the predicted class.

Name	Number of Concordant Neighbors[1]	Number of Compounds	ACC	MCC	AUC	CCR	Se	Sp	PPV	NPV
RF_MACCS	0	87	0.33	-0.35	0.32	0.33	0.19	0.48	0.26	0.38
RF_MACCS	≥1	197	0.89	0.77	0.97	0.87	0.81	0.94	0.89	0.89
RF_MACCS	≥2	147	0.96	0.90	1.00	0.94	0.89	0.99	0.98	0.95
RF_MACCS	≥3	113	0.99	0.98	1.00	0.98	0.97	1.00	1.00	0.99
SVM_MOE2D+OASIS	0	85	0.56	0.13	0.56	0.57	0.62	0.51	0.55	0.58
SVM_MOE2D+OASIS	≥1	198	0.84	0.69	0.94	0.85	0.92	0.79	0.72	0.94
SVM_MOE2D+OASIS	≥2	146	0.91	0.81	0.99	0.92	0.95	0.89	0.79	0.98
SVM_MOE2D+OASIS	≥3	115	0.91	0.80	0.99	0.92	0.94	0.90	0.79	0.97
SVM_PaDEL	0	86	0.53	0.07	0.52	0.54	0.56	0.51	0.51	0.56
SVM_PaDEL	≥1	193	0.83	0.66	0.92	0.84	0.87	0.8	0.72	0.92
SVM_PaDEL	≥2	147	0.89	0.78	0.96	0.91	0.96	0.86	0.76	0.98
SVM_PaDEL	≥3	113	0.90	0.79	0.97	0.92	0.97	0.88	0.76	0.99
SVM_PaDEL+MACCS	0	86	0.55	0.10	0.53	0.55	0.59	0.51	0.52	0.57
SVM_PaDEL+MACCS	≥1	193	0.84	0.68	0.91	0.85	0.89	0.81	0.73	0.93
SVM_PaDEL+MACCS	≥2	147	0.90	0.80	0.96	0.92	0.96	0.88	0.79	0.98
SVM_PaDEL+MACCS	≥3	113	0.91	0.81	0.97	0.93	0.97	0.89	0.78	0.99
SVM_PaDEL+OASIS	0	86	0.53	0.07	0.52	0.54	0.56	0.51	0.51	0.56
SVM_PaDEL+OASIS	≥1	193	0.83	0.67	0.92	0.84	0.87	0.81	0.73	0.92
SVM_PaDEL+OASIS	≥2	147	0.9	0.79	0.96	0.91	0.96	0.87	0.77	0.98
SVM_PaDEL+OASIS	≥3	113	0.91	0.81	0.97	0.93	0.97	0.89	0.78	0.99

[1] Number of consecutive nearest neighbors in the training data having the same activity class assigned as the one predicted for the test compounds. Abbreviations: ACC, accuracy; AUC, area under the receiver operating characteristic curve; CCR, correct classification rate; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; Se, sensitivity; Sp, specificity.

2.5. Skin Doctor Web Service

The final RF_MACCS and SVM_PaDEL models, trained not on the cross-validation data set but on the complete, preprocessed data set (1416 and 1388 compounds, depending on the number of compounds for which descriptors could be successfully calculated) are provided via the New E-Resource for Drug Discovery (NERDD) [41]. Queries can either be directly drawn or uploaded in different formats. Users may change the default decision threshold to steer the model's sensitivity and specificity. Results are presented in a tabular overview and can be exported as a CSV file. For each query they include information on (i) whether or not the query is within the applicability domain of the model, (ii) the predicted activity classes, (iii) distances from the selected decision threshold, (iv) mean similarity between the query compound and the five-nearest neighbors of the training set and (v) number of consecutive nearest neighbors in the training data of which the activity label is consistent with that of the prediction. The analysis and visualization of the corresponding effects presented in this work may be used as guidance to choose the required confidence in the prediction, being aware of the corresponding effects on the model's applicability domain and the requirements for similarity.

Predictions are flagged with reliability warnings (a) if the mean similarity between the compound of interest and the five nearest neighbors is less than 0.5, or (b) if the predictions are in conflict with the activity of the nearest neighbor in the training data, or (c) if the distance to the decision threshold is small (0.15 for the RF_MACCS model; 0.5 for the SVM_PaDEL model).

3. Materials and Methods

3.1. Data Preparation

The LLNA data set compiled by Alves et al. was downloaded from Chembench. Binary class labels (i.e., "sensitizer", "non-sensitizer") were obtained from the binary property "LLNA result" and not altered. The LLNA data set of Di et al. was obtained from the supporting information associated with their publication [22]. Binary class labels (i.e., "sensitizer", "non-sensitizer") were assigned based on the information provided by the property "class": any compounds with the value "negative" were assigned the label "non-sensitizer"; any compounds with the value "weak", "moderate", "strong" or "extreme" were assigned the label "sensitizer". Reference data sets of cosmetic substances and ingredients (hereafter "cosmetics"), approved drugs and pesticides were obtained from the EU CosIng database, Drugbank and EU pesticides database.

All data sets were processed individually according to the following protocol: Any counterions were removed and the remaining molecular structures neutralized as described in the work of Stork et al. [42]. Tautomers were standardized with the "TautomerCanonicalizer" method implemented in the "tautomer" class of MolVS [43]. This was followed by a deduplication of molecules based on canonicalized SMILES. Stereochemical information was disregarded at this point, leading to conflicting activity labels for one compound (which had different activity labels assigned to the two enantiomers). This compound was removed from the data set.

A merged LLNA data set based on the LLNA data sets of Alves et al. and Di et al. was generated by filtering duplicates based on canonical SMILES and removing any compounds with contradicting class labels.

3.2. Descriptor Calculation

Molecular descriptors were computed with the Molecular Operating Environment (MOE) [36] ("MOE descriptors"), RDKit [39] (Morgan and MACCS fingerprints) and PaDEL [37,38] ("PaDEL descriptors" as well as the molecular fingerprints "PaDEL-Est" and "PaDEL-Ext"). "MOE 2D" descriptors were calculated with default settings. Morgan fingerprints (2048 bits) were calculated with a radius of 2. MACCS fingerprints were calculated with default settings. Also, the PaDEL descriptors were calculated with default settings, with the exception of a maximum allowed runtime of 1000 s per molecule. Structural alerts were computed with the OECD toolbox [29] using the "Protein binding alerts for skin sensitization by OASIS" profiler with default settings. All non-binary descriptors were scaled to unit variance and their mean shifted to zero prior to model building and data analysis using the StandardScaler of scikit-learn [44].

3.3. Data Analysis

PCA was conducted with scikit-learn based on a subset of 53 physically meaningful, scaled "MOE 2D" descriptors (Table S1). RDKit was employed for generating Murcko scaffolds and calculating molecular similarity.

3.4. Compilation of Data Sets for Model Development

The merged LLNA data set was divided into a training set (80%) and a test set (20%) by stratified splitting with the train_test_split function of the model_selection module of scikit-learn (data shuffling prior to data set splitting enabled). This procedure was assigned a random state of 43.

3.5. Model Generation

Models were generated with scikit-learn and a random_state value of 43. Default settings were applied, with the exception of class_weight set to "balanced" for both RF and SVM. SVMs were

used with a radial basis function (RBF) kernel. Optimal settings for n_estimators and max_features (RF models) and C and gamma (SVM models) were derived during grid search.

3.6. Hardware and Software

All calculations were performed on Linux workstations running openSUSE Leap 15.0 and equipped with Intel i5 processors (3.2 GHz) and 16 GB of main memory.

4. Conclusions

Building on the works of Alves et al. and Di et al., we have compiled a collection of 1416 compounds annotated with binary LLNA outcomes. To our knowledge, this is the largest LLNA data set that has been used for the development of models predicting the skin sensitization potential of small organic molecules. As we show by chemical space analysis, those areas most densely populated by cosmetics, approved drugs and pesticides are also well covered by this new LLNA data set. The fraction of compounds covered by structurally related compounds in the new LLNA data set is much higher for cosmetics (30%) than for approved drugs (10%) and pesticides (13%). Therefore, the models are applicable to many compounds typically used in cosmetic products. However, there are chemical classes of drugs and cosmetics that are not adequately represented by the available LLNA data. This emphasizes the importance of considering the applicability domain of models.

An interesting observation to make was that a cluster of skin sensitizers and non-sensitizers with long aliphatic and halogenated chains could only be discriminated in the "MOE 2D" descriptor space but not in the Morgan2 fingerprint space, which should be taken into consideration for model building. The best models derived from the new LLNA data set obtained MCC and ACC values of up to 0.55 and 0.78 during cross-validation and of up to 0.52 and 0.76 on holdout data, respectively. Importantly, some of the models based entirely on free software and/or molecular descriptors of low complexity yielded comparable performance. We identified the RF_MACCS and SVM_PaDEL models as our favorite models, yielding MCC values of 0.41 and 0.47 on the holdout data. Comparison to existing models indicates that our models reach competitive performance. They are trained on a data set consisting of almost 3.5 times as many compounds as the one used by Di et al. The full data set used for modeling and testing is also 42% larger than that of Alves et al. given the fact that the data set compiled by Di et al. holds in particular a diverse set of non-sensitizers not covered by Alves et al. we expect that our models, as they are based on the amalgamated data set, are more widely applicable and more reliable.

A major aspect of this work is the definition of an applicability domain for the individual models and the elaboration of means to estimate the reliability of predictions. The applicability domain was defined based on the mean similarity of a compound of interest to the five-nearest neighbors in the training data (quantified in MACCS fingerprint space). The difference between the predicted class probability and the decision threshold, as well as the number of consecutive nearest neighbors in the training data having the same activity class assigned as the one predicted for the compound of interest proved to be useful indicators of the reliability of predictions. We recommend considering predictions as reliable if all of the following conditions are met:

1. The compound of interest is within the applicability domain of the model.
2. The distance between the predicted class probability and the decision threshold is at least 0.15 for RF models and 0.5 for SVM models.
3. The predicted activity class for a compound of interest is in agreement with the class assigned to the nearest neighbor in the training data.

The public web service, available at https://nerdd.zbh.uni-hamburg.de/, provides access to the final RF_MACCS and SVM_PaDEL models (i.e., models trained on the complete LLNA data set). Users are provided detailed information on whether or not a compound of interest fulfills the three criteria itemized above. A warning is issued in case predictions are determined to be unreliable. Users

may also adjust the decision threshold, allowing them, e.g., to increase the model's sensitivity in scenarios where it is desirable to flag even substances with a low likelihood of being skin sensitizers.

We hope that the models will be well received by the scientific community and will make a contribution to the development and application of non-animal methods for the prediction of the skin sensitization potential of small organic molecules.

Supplementary Materials: Supplementary Materials can be found at http://www.mdpi.com/1422-0067/20/19/4833/s1.

Author Contributions: Conceptualization, A.W., J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); Methodology, A.W., C.S., C.B., J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); Software, A.W. and C.S.; Validation, A.W.; Formal Analysis, A.W.; Investigation, A.W., C.S., C.B., J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); Resources, A.S., J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); Data Curation, A.W.; Writing – Original Draft Preparation, A.W., C.S., C.B., A.S., J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); Writing – Review & Editing, A.W., C.S., C.B., A.S., J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); Visualization, A.W.; Supervision, A.S., J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); Project Administration, J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); Funding Acquisition, A.S., J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair).

Funding: A.W. is supported by Beiersdorf AG through HITeC e.V. C.B. and J.K. (Johannes Kirchmair) are supported by the Trond Mohn Foundation [BFS2017TMT01].

Conflicts of Interest: J.K. (Jochen Kühnl) and A.S. are employed at Beiersdorf AG and A.W. is funded by Beiersdorf AG through HITeC e.V. A.W., A.S. and J.K. (Jochen Kühnl) were involved in the design of the study, the interpretation of the data, the writing of the manuscript, and the decision to publish the results.

Abbreviations

ACC	accuracy
ACD	allergic contact dermatitis
AUC	area under the receiver operating characteristic curve
CCR	correct classification rate
LLNA	local lymph node assay
MCC	Matthews correlation coefficient
MOE	Molecular Operating Environment
NERDD	New E-Resource for Drug Discovery
NPV	negative predictive value
PCA	principal component analysis
RBF	radial basis function
RF	random forest
PPV	positive predictive value
Se	sensitivity
Sp	specificity
SVM	support vector machine

References

1. Kimber, I.; Basketter, D.A.; Gerberick, G.F.; Ryan, C.A.; Dearman, R.J. Chemical allergy: Translating biology into hazard characterization. *Toxicol. Sci.* **2011**, *120* (Suppl. 1), S238–S268. [CrossRef]
2. Thyssen, J.P.; Linneberg, A.; Menné, T.; Johansen, J.D. The epidemiology of contact allergy in the general population—prevalence and main findings. *Contact Dermat.* **2007**, *57*, 287–299. [CrossRef]
3. Lushniak, B.D. Occupational contact dermatitis. *Dermatol. Ther.* **2004**, *17*, 272–277. [CrossRef]
4. Anderson, S.E.; Siegel, P.D.; Meade, B.J. The LLNA: A brief review of recent advances and limitations. *J. Allergy* **2011**, *2011*. [CrossRef]
5. Dent, M.; Amaral, R.T.; Da Silva, P.A.; Ansell, J.; Boisleve, F.; Hatao, M.; Hirose, A.; Kasai, Y.; Kern, P.; Kreiling, R.; et al. Principles underpinning the use of new methodologies in the risk assessment of cosmetic ingredients. *Comput. Toxicol.* **2018**, *7*, 20–26. [CrossRef]
6. Mehling, A.; Eriksson, T.; Eltze, T.; Kolle, S.; Ramirez, T.; Teubner, W.; van Ravenzwaay, B.; Landsiedel, R. Non-animal test methods for predicting skin sensitization potentials. *Arch. Toxicol.* **2012**, *86*, 1273–1295. [CrossRef]

7. Reisinger, K.; Hoffmann, S.; Alépée, N.; Ashikaga, T.; Barroso, J.; Elcombe, C.; Gellatly, N.; Galbiati, V.; Gibbs, S.; Groux, H.; et al. Systematic evaluation of non-animal test methods for skin sensitisation safety assessment. *Toxicol. In Vitro* **2015**, *29*, 259–270. [CrossRef]
8. Ezendam, J.; Braakhuis, H.M.; Vandebriel, R.J. State of the art in non-animal approaches for skin sensitization testing: From individual test methods towards testing strategies. *Arch. Toxicol.* **2016**, *90*, 2861–2883. [CrossRef]
9. Thyssen, J.P.; Giménez-Arnau, E.; Lepoittevin, J.-P.; Menné, T.; Boman, A.; Schnuch, A. The critical review of methodologies and approaches to assess the inherent skin sensitization potential (skin allergies) of chemicals. Part I. *Contact Dermat.* **2012**, *66* (Suppl. 1), 11–24. [CrossRef]
10. Wilm, A.; Kühnl, J.; Kirchmair, J. Computational approaches for skin sensitization prediction. *Crit. Rev. Toxicol.* **2018**, *48*, 738–760. [CrossRef]
11. ECHA (European Chemicals Agency). The Use of Alternatives to Testing on Animals for the REACH Regulation, Third Report under Article 117(3) of the REACH Regulation. Available online: https://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2017_en.pdf (accessed on 10 July 2019).
12. Kleinstreuer, N.C.; Hoffmann, S.; Alépée, N.; Allen, D.; Ashikaga, T.; Casey, W.; Clouet, E.; Cluzel, M.; Desprez, B.; Gellatly, N.; et al. Non-animal methods to predict skin sensitization (II): An assessment of defined approaches. *Crit. Rev. Toxicol.* **2018**, *48*, 359–374. [CrossRef] [PubMed]
13. Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. *Toxicol. Sci.* **2018**, *165*, 198–212. [CrossRef] [PubMed]
14. Luechtefeld, T.; Rowlands, C.; Hartung, T. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. *Toxicol. Res.* **2018**, *7*, 732–744. [CrossRef] [PubMed]
15. Alves, V.M.; Borba, J.; Capuzzi, S.J.; Muratov, E.; Andrade, C.H.; Rusyn, I.; Tropsha, A. Oy vey! A comment on "Machine learning of toxicological big data enables read-across structure activity relationships outperforming animal test reproducibility". *Toxicol. Sci.* **2019**, *167*, 3–4. [CrossRef] [PubMed]
16. Luechtefeld, T.; Marsh, D.; Hartung, T. Missing the difference between big data and artificial intelligence in RASAR versus traditional QSAR. *Toxicol. Sci.* **2019**, *167*, 4–5. [CrossRef] [PubMed]
17. Tung, C.-W.; Lin, Y.-H.; Wang, S.-S. Transfer learning for predicting human skin sensitizers. *Arch. Toxicol.* **2019**, *93*, 931–940. [CrossRef]
18. Chilton, M.L.; Macmillan, D.S.; Steger-Hartmann, T.; Hillegass, J.; Bellion, P.; Vuorinen, A.; Etter, S.; Smith, B.P.C.; White, A.; Sterchele, P.; et al. Making reliable negative predictions of human skin sensitisation using an in silico fragmentation approach. *Regul. Toxicol. Pharm.* **2018**, *95*, 227–235. [CrossRef]
19. Braga, R.C.; Alves, V.M.; Muratov, E.N.; Strickland, J.; Kleinstreuer, N.; Trospsha, A.; Andrade, C.H. Pred-Skin: A fast and reliable web application to assess skin sensitization effect of chemicals. *J. Chem. Inf. Model.* **2017**, *57*, 1013–1017. [CrossRef]
20. Kim, J.Y.; Kim, M.K.; Kim, K.-B.; Kim, H.S.; Lee, B.-M. Quantitative structure–activity and quantitative structure–property relationship approaches as alternative skin sensitization risk assessment methods. *J. Toxicol. Environ. Health* **2019**, *82*, 447–472. [CrossRef]
21. Toropov, A.A.; Toropova, A.P.; Selvestrel, G.; Benfenati, E. Idealization of correlations between optimal simplified molecular input-line entry system-based descriptors and skin sensitization. *SAR QSAR Environ. Res.* **2019**, *30*, 447–455. [CrossRef]
22. Di, P.; Yin, Y.; Jiang, C.; Cai, Y.; Li, W.; Tang, Y.; Liu, G. Prediction of the skin sensitising potential and potency of compounds via mechanism-based binary and ternary classification models. *Toxicol. In Vitro* **2019**, *59*, 204–214. [CrossRef] [PubMed]
23. Alves, V.M.; Muratov, E.; Fourches, D.; Strickland, J.; Kleinstreuer, N.; Andrade, C.H.; Tropsha, A. Predicting chemically-induced skin reactions. Part I: QSAR models of skin sensitization and their application to identify potentially hazardous compounds. *Toxicol. Appl. Pharmacol.* **2015**, *284*, 262–272. [CrossRef] [PubMed]
24. Lu, J.; Zheng, M.; Wang, Y.; Shen, Q.; Luo, X.; Jiang, H.; Chen, K. Fragment-based prediction of skin sensitization using recursive partitioning. *J. Comput. Aided Mol. Des.* **2011**, *25*, 885–893. [CrossRef] [PubMed]
25. Chaudhry, Q.; Piclin, N.; Cotterill, J.; Pintore, M.; Price, N.R.; Chrétien, J.R.; Roncaglioni, A. Global QSAR models of skin sensitisers for regulatory purposes. *Chem. Cent. J.* **2010**, *4*, S5. [CrossRef] [PubMed]
26. Enoch, S.J.; Roberts, D.W. Predicting skin sensitization potency for Michael acceptors in the LLNA using quantum mechanics calculations. *Chem. Res. Toxicol.* **2013**, *26*, 767–774. [CrossRef] [PubMed]

27. Hoffmann, S. LLNA variability: An essential ingredient for a comprehensive assessment of non-animal skin sensitization test methods and strategies. *ALTEX* **2015**, *32*, 379–383. [PubMed]
28. Alves, V.M.; Capuzzi, S.J.; Braga, R.C.; Borba, J.V.B.; Silva, A.C.; Luechtefeld, T.; Hartung, T.; Andrade, C.H.; Muratov, E.N.; Tropsha, A. A perspective and a new integrated computational strategy for skin sensitization assessment. *ACS Sustain. Chem. Eng.* **2018**, *6*, 2845–2859. [CrossRef]
29. Apt Systemst Ltd. Aptsys.net OASIS. QSAR Toolbox 4.3. Available online: http://oasis-lmc.org/products/software/toolbox.aspx (accessed on 10 July 2019).
30. Chembench|Home. Available online: https://chembench.mml.unc.edu (accessed on 26 April 2019).
31. CosIng—Cosmetics—GROWTH—European Commission. Available online: http://ec.europa.eu/growth/tools-databases/cosing/index.cfm?fuseaction=search.simple (accessed on 26 April 2019).
32. DrugBank Version 5.1.2. Available online: https://www.drugbank.ca (accessed on 7 May 2019).
33. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef] [PubMed]
34. EU Pesticides Database—European Commission. Available online: http://ec.europa.eu/food/plant/pesticides/eu-pesticides-database/public/?event=activesubstance.selection&language=EN (accessed on 25 February 2019).
35. Chemical Identifier Resolver. Available online: https://cactus.nci.nih.gov/chemical/structure (accessed on 25 February 2019).
36. Chemical Computing Group Molecular Operating Environment (MOE)|MOEsaic|PSILO. Available online: https://www.chemcomp.com/Products.htm (accessed on 12 June 2019).
37. PaDEL-Descriptor. Available online: http://www.yapcwsoft.com/dd/padeldescriptor/ (accessed on 10 May 2019).
38. Yap, C.W. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef]
39. Landrum, G. RDKit. Available online: http://www.rdkit.org (accessed on 26 April 2019).
40. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* **1975**, *405*, 442–451. [CrossRef]
41. Stork, C.; Embruch, G.; Šícho, M.; de Bruyn Kops, C.; Chen, Y.; Svozil, D.; Kirchmair, J. NERDD: A web portal providing access to in silico tools for drug discovery. *Bioinformatics* **2019**. [CrossRef]
42. Stork, C.; Wagner, J.; Friedrich, N.-O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A machine-learning model for the prediction of frequent hitters. *Chem. Med. Chem.* **2018**, *13*, 564–571. [CrossRef] [PubMed]
43. MolVs. MolVs Version 0.1.1. Available online: https://github.com/mcs07/MolVS (accessed on 26 April 2019).
44. Scikit-Learn: Machine Learning in Python—Scikit-Learn 0.21.0 Documentation. Available online: https://scikit-learn.org/stable/ (accessed on 10 May 2019).

6.2 An aggregated Mondrian conformal prediction workflow to predict binary and ternary skin sensitization potential

Most definitions of the AD of a ML model are based on a more or less arbitrarily selected cutoff value to separate unreliable and reliable predictions. One mathematically sound approach that does not need such a cutoff parameter is wrapping the ML model into a CP framework. The CP framework provides a solid and mathematically proven reliability measure for every single prediction returned by a valid model.

Most common ML models for skin sensitization prediction address binary skin sensitization potential as the final target. Within such a binary setting, all substances labeled as sensitizers have to be treated with care and possibly be removed from final consumer products. In contrast to this, the prediction of skin sensitization potency or at least more than two classes of skin sensitization potential would enable the usage of less severe sensitizers in limited doses.

In the following study, we further increased the quality of our data set for model building and evaluation by applying an additional manual data curation step, which included the manual inspection of each 2D structure in one or several public chemical data sets. We, furthermore, enveloped one of our best performing ML models for the prediction of binary skin sensitization potential (now trained on the slightly smaller but more carefully curated data set) into an aggregated Mondrian CP framework to ensure mathematically defined reliability for all individual predictions. In addition, we presented two different combinations of our binary classifier with a second binary classifier to distinguish weak to moderate from strong to extreme sensitizers. The binary model has been made available as a web service named Skin Doctor CP.

We undertook a careful analysis of the models' validity and efficiency, especially with respect to single activity classes. Furthermore, the ternary model was extensively compared to a ternary model published by Di et al. [16]. Ternary class information retrieved from our data set was also utilized to demonstrate that sensitizers wrongly predicted as non-sensitizers by our binary model are most likely weak to moderate sensitizers and less likely to be strong to extreme sensitizers, which should be beneficial for risk assessment.

P3 Wilm, A., Norinder, U., Agea, M. I., de Bruyn Kops, C., Stork, C., Kühnl, J., Kirchmair, J., Skin doctor CP: Conformal prediction of the skin sensitization potential of small organic molecules, *Chemical Research in Toxicology*, **34**(2) (2020) 330–344

Available at:
https://pubs.acs.org/doi/10.1021/acs.chemrestox.0c00253

A. Wilm, I. M. Agea, U. Norinder, J. Kühnl, and J. Kirchmair conceptualized the work. A. Wilm curated the data set, trained and evaluated the models and analyzed and visualized the research results. A. Wilm developed the underlying software and methodologies, with contributions by I. M. Agea, C. de Bruyn Kops and C. Stork. A. Wilm, C. de Bruyn Kops and C. Stork implemented the Skin Doctor CP web service. A. Wilm prepared the the manuscript, with contributions by C. de Bruyn Kops, C. Stork, I. M. Agea, U. Norinder, J. Kühnl, and J. Kirchmair. J. Kühnl, U. Norinder and J. Kirchmair supervised the work. Resources, project administration and funding were provided or aquired by J. Kühnl and J. Kirchmair. All authors have read and agreed to the published version of the manuscript.

Chemical Research in Toxicology

pubs.acs.org/crt

Article

Skin Doctor CP: Conformal Prediction of the Skin Sensitization Potential of Small Organic Molecules

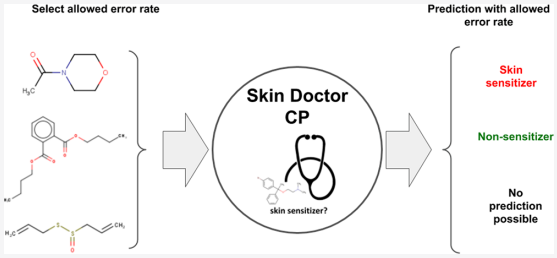
Anke Wilm, Ulf Norinder, M. Isabel Agea, Christina de Bruyn Kops, Conrad Stork, Jochen Kühnl, and Johannes Kirchmair*

Cite This: *Chem. Res. Toxicol.* 2021, 34, 330−344

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Skin sensitization potential or potency is an important end point in the safety assessment of new chemicals and new chemical mixtures. Formerly, animal experiments such as the local lymph node assay (LLNA) were the main form of assessment. Today, however, the focus lies on the development of nonanimal testing approaches (i.e., in vitro and in chemico assays) and computational models. In this work, we investigate, based on publicly available LLNA data, the ability of aggregated, Mondrian conformal prediction classifiers to differentiate between non-sensitizing and sensitizing compounds as well as between two levels of skin sensitization potential (weak to moderate sensitizers, and strong to extreme sensitizers). The advantage of the conformal prediction framework over other modeling approaches is that it assigns compounds to activity classes only if a defined minimum level of confidence is reached for the individual predictions. This eliminates the need for applicability domain criteria that often are arbitrary in their nature and less flexible. Our new binary classifier, named Skin Doctor CP, differentiates nonsensitizers from sensitizers with a higher reliability-to-efficiency ratio than the corresponding nonconformal prediction workflow that we presented earlier. When tested on a set of 257 compounds at the significance levels of 0.10 and 0.30, the model reached an efficiency of 0.49 and 0.92, and an accuracy of 0.83 and 0.75, respectively. In addition, we developed a ternary classification workflow to differentiate nonsensitizers, weak to moderate sensitizers, and strong to extreme sensitizers. Although this model achieved satisfactory overall performance (accuracies of 0.90 and 0.73, and efficiencies of 0.42 and 0.90, at significance levels 0.10 and 0.30, respectively), it did not obtain satisfying class-wise results (at a significance level of 0.30, the validities obtained for nonsensitizers, weak to moderate sensitizers, and strong to extreme sensitizers were 0.70, 0.58, and 0.63, respectively). We argue that the model is, in consequence, unable to reliably identify strong to extreme sensitizers and suggest that other ternary models derived from the currently accessible LLNA data might suffer from the same problem. Skin Doctor CP is available via a public web service at https://nerdd.zbh.uni-hamburg.de/skinDoctorII/.



INTRODUCTION

Skin sensitizers are substances that have the potential to cause allergic contact dermatitis (ACD) during repeated exposure.[1] ACD is a major cause of occupational illnesses[2,3] and can severely diminish the quality of life of affected individuals. Therefore, thorough safety assessment is required prior to market release of new substances to prevent the induction of occupational or product exposure-based ACD. Moreover, in case of a skin sensitization hazard, potency information (i.e., the concentration required to induce skin sensitization) is key to determine safe use concentrations that do not result in the induction of skin sensitization.[4]

Historically, the skin sensitization potential and potency of substances have been mainly assessed by in vivo studies on animals and, rarely, complemented by confirmatory studies using safe doses on humans. The local lymph node assay (LLNA),[5] conducted in mice, is today considered the most advanced animal testing system for skin sensitization potential and potency.[6] In contrast to other animal assays, the LLNA assesses solely the induction phase and delivers potency information in the form of an EC3 value, which is considered to be a quantitative measure of the skin sensitization potency.[7] The EC3 value represents a concentration required to derive a point of departure for quantitative risk assessment. However, the predictive capacity of animal testing for humans is limited (in general[8] and also with regard to skin sensitization prediction[9]), and ethical and practical considerations as well as regulatory constraints have led to the development of alternatives to animal testing. These alternatives comprise in

Special Issue: Computational Toxicology

Received: June 23, 2020
Published: December 9, 2020

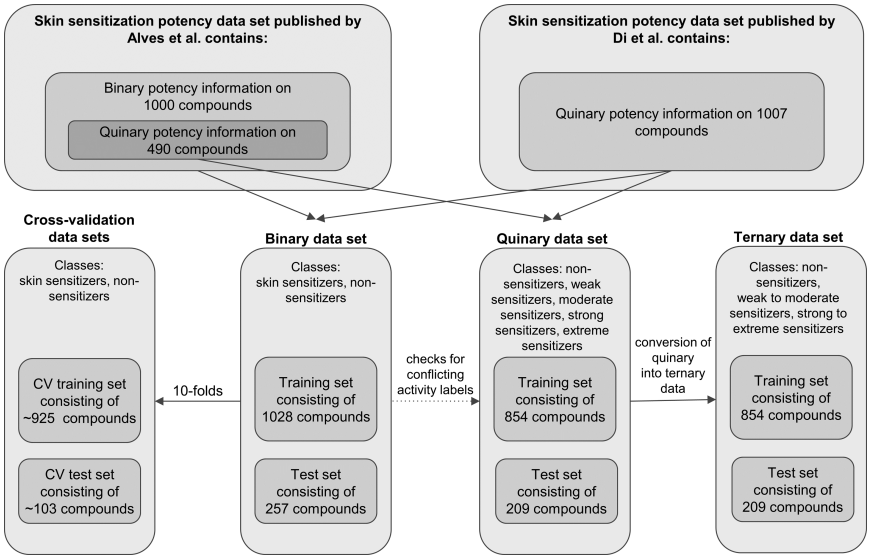


Figure 1. Overview of LLNA data sets and subsets employed in this study.

chemico and in vitro testing methods,[10−13] as well as in silico tools that predict a compound's skin sensitization potential based on its chemical structure or properties calculated therefrom.[12−15] Nevertheless, the reliability and coverage of the individual alternative approaches is still limited, primarily due to the scarcity of available high-quality data for the development and validation of methods. For this reason, researchers have been exploring strategies for the combination of multiple nonredundant assays to achieve or exceed the level of predictive hazard or potency information provided by animal model data.[16] These combined approaches are known as defined approaches (DAs) and as integrated approaches for testing and assessment (IATAs) and have been recently reviewed in ref 9. For the qualification of cosmetic compounds, in silico predictions can contribute to the prioritization of chemicals for efficacy testing and, subsequently, to early phases of (tiered) safety assessment strategies. For the latter, predictions can be used in "weight of evidence" considerations for risk assessment such as the dermal sensitization threshold approach[17] or as input for DAs and IATAs. For a computational model to be accepted within a regulatory context, it should fulfill the five validation principles outlined by the OECD:[18] a defined end point, an unambiguous algorithm, a defined applicability domain (AD), appropriate measures of goodness-of-fit, robustness, and predictivity, and, if possible, a mechanistic interpretation.

In the context of in silico prediction tools, the AD of a method defines the chemical space within which a method produces results with a defined reliability.[19,20] Most AD definitions include a more or less arbitrary or user-defined threshold to differentiate between reliable and unreliable predictions based on similarity to training data or the class probability returned by the modeling algorithm.[21]

An alternative for defining the reliability of a model for a certain compound of interest, without the definition of an AD, is offered by conformal prediction (CP).[22−24] Whereas classical, standalone machine learning models based on support vector machines (SVMs), random forests (RFs), or other methods return a distinct prediction for a compound of interest (or, in the case of RF, a class probability, if desired), a CP model returns statistically justified class membership probabilities for each of the classes. Users may select a desired confidence level, $1-\varepsilon$, and CP will return an observed error equal to, or very close to, the chosen error rate ε, as long as the randomness assumption of the samples (an assumption that is also made for classical machine learning models) holds true. On the basis of the class probabilities and the selected confidence level, the model determines whether a compound is within the AD of the model. If it is within the AD, one or more class labels will be assigned to the compound; if it is outside the AD, no class label will be assigned (or, more precisely, the compound will be assigned to the empty (null) class). As with the AD of classical machine learning models, different measures of the reliability of a prediction (conformity measures) may be selected for the model. However, the CP model offers the advantage that the manual selection of a cutoff value for this measure is not required. Instead, it is deduced in a straightforward mathematical way from the selected confidence level.

Different variants of CP support different needs regarding the characteristics of the modeling data, and the computational effort that should be invested.[25] A CP variant that has been shown to perform favorably on imbalanced data is Mondrian CP, because it treats each class independently of all other(s), thereby ensuring the validity of each individual class without the need for over- or under-sampling.[26−28] An additional type of CP is aggregated CP, which repeats the workflow several times so that each training compound could be used as a proper training and calibration compound.[29] Aggregated CP is therefore favorable for small data sets. The combination of Mondrian CP and aggregated CP works particularly well on small, imbalanced data sets.

In this study, we apply aggregated, Mondrian CP to develop classifiers for the prediction of the skin sensitization potential of small molecules. We start with the development of a binary classifier that distinguishes nonsensitizers from sensitizers and

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

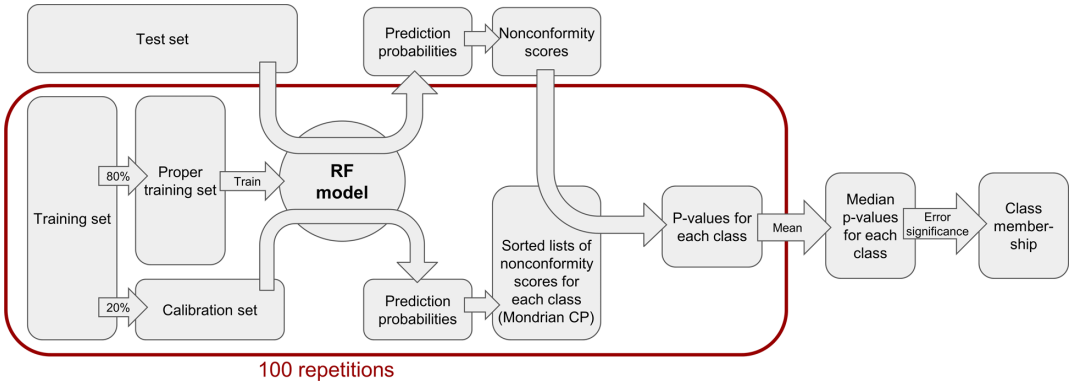


Figure 2. Schematic workflow of the aggregated Mondrian CP model.

then explore strategies to obtain a differentiation of weak to moderate sensitizers from strong to extreme sensitizers. The performance of the models is determined with thorough validation protocols and compared to the performance of existing in silico models. The final classifier, called "Skin Doctor CP", is available as a web service, free of charge for academic use.

METHODS

Data Sets. For the purpose of model development and evaluation, LLNA data sets on the skin sensitization potential of small organic compounds (Figure 1) were derived from the data published by Alves et al.[30] and Di et al.[31] (all data are provided as Supporting Information, SI). The data set was prepared following a protocol described previously,[32] which includes the removal of counterions, neutralization, standardization of tautomers, removal of stereochemical information, and removal of duplicate compounds and compounds with conflicting activity data based on canonical SMILES. For the current work, we refined this protocol by discarding any entries for which, based on the information provided by Alves et al. and Di et al., the exact molecular structure of the compound in question could not be conclusively confirmed. More specifically, we discarded any entries that match at least one of the following criteria:

- the CAS number provided refers to a polymer, an unspecified substance, or an incompletely defined substance (this concerns 49 and 60 entries of the data sets of Alves et al. and Di et al., respectively)
- the CAS number provided refers to a multicomponent substance for which the relevant component could not be unequivocally identified (this concerns 2 and 0 entries, respectively)
- the CAS number provided refers to a metal complex (this concerns 1 and 7 entries, respectively) or a metal salt (this concerns 1 and 1 entries, respectively)
- the CAS number provided refers to a substance with a molecular structure that is not consistent with the SMILES notation provided (this concerns 5 and 5 entries, respectively)
- the CAS number, EC number, compound name, and any further information provided did not allow to confirm the molecular structure of the substance in question (this concerns 2 and 40 entries, respectively)

Further, multicomponent mixtures that have been tested negative and for which the least represented component accounts for at least one-third of the proportion of the major component were split into separate entries, each assigned to the "nonsensitizer" class (this concerns 7 and 15 entries of the data sets of Alves et al. and Di et al., respectively). In the case of two-component mixtures that (i) have been tested positive, (ii) for which one component is listed as a known nonsensitizer in the data sets of Alves et al. or Di et al., and (iii) for which the known nonsensitizer accounts for at least one-third of the mixture, the class label "sensitizer" was assigned to the other component (this concerns 1 entry derived from the data set compiled by Di et al.). The curated data set (Table SI_1) as well as the substances removed by the manual data curation process (Table SI_2) can be found in the SI published with this article.

Binary Data Set. The binary class labels of the data set were retrieved by a protocol identical to the one published in ref 32.

Multiclass Data Sets. All compounds included in the data set of Di et al.[31] and approximately half of the compounds included in the data set of Alves et al.[30] are annotated with quinary LLNA data (Figure 1). The quinary potency information was used to derive a ternary data set (for the development of a ternary classifier) and a quinary data set (for the evaluation of the binary classifier with regard to quinary class memberships) following the identical data processing protocol of Wilm et al.[32]

Compounds originating from the work of Alves et al. were assigned class labels based on the "LLNA class" property, whereas compounds sourced from the work of Di et al. were assigned class labels according to the "Classes" property. Compounds labeled as "Nonsensitizer" (Alves et al.) or "Negative" (Di et al.) were assigned the class label "non-sensitizer". For the compilation of the quinary data set, the class labels "Weak", "Moderate", "Strong", and "Extreme" sensitizers from both sources were preserved. For the compilation of the ternary data set, the quinary data were converted according to the following rules: "Weak" and "Moderate" skin sensitizers from both sources were assigned to the class "weak to moderate sensitizers", whereas "Strong" and "Extreme" skin sensitizers from both sources were assigned to the class "strong to extreme sensitizers". Compounds without data on their skin sensitization potential were removed (220 compounds). Following the conversion of the activity labels, three compounds were removed from the data set because of conflicting class labels.

Determination of Functional Groups for Data Set Analysis. The binary data set was analyzed with respect to the prevalence of the functional groups in organic chemistry encoded by 309 SMARTS patterns.[33] SMARTS pattern matching was performed with RDKit.[34] Any patterns matched by at least 20 out of the investigated compounds (1285 in the case of data set analysis, 275 in the case of performance analysis of the binary classifier) were included in the analysis.

Descriptor Calculation. Skin Doctor CP uses MACCS keys (166 bits), which have been identified as the most suitable descriptors during the development of Skin Doctor.[32] These descriptors are calculated with RDKit.

Model Generation with Aggregated Mondrian Conformal Prediction. *Definition of Training and Test Sets.* The binary data set was divided into a training set (80% of the data) and a test set (20% of the data). To maximize the comparability of the current study with our previous work,[32] we preserved the data set split.

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

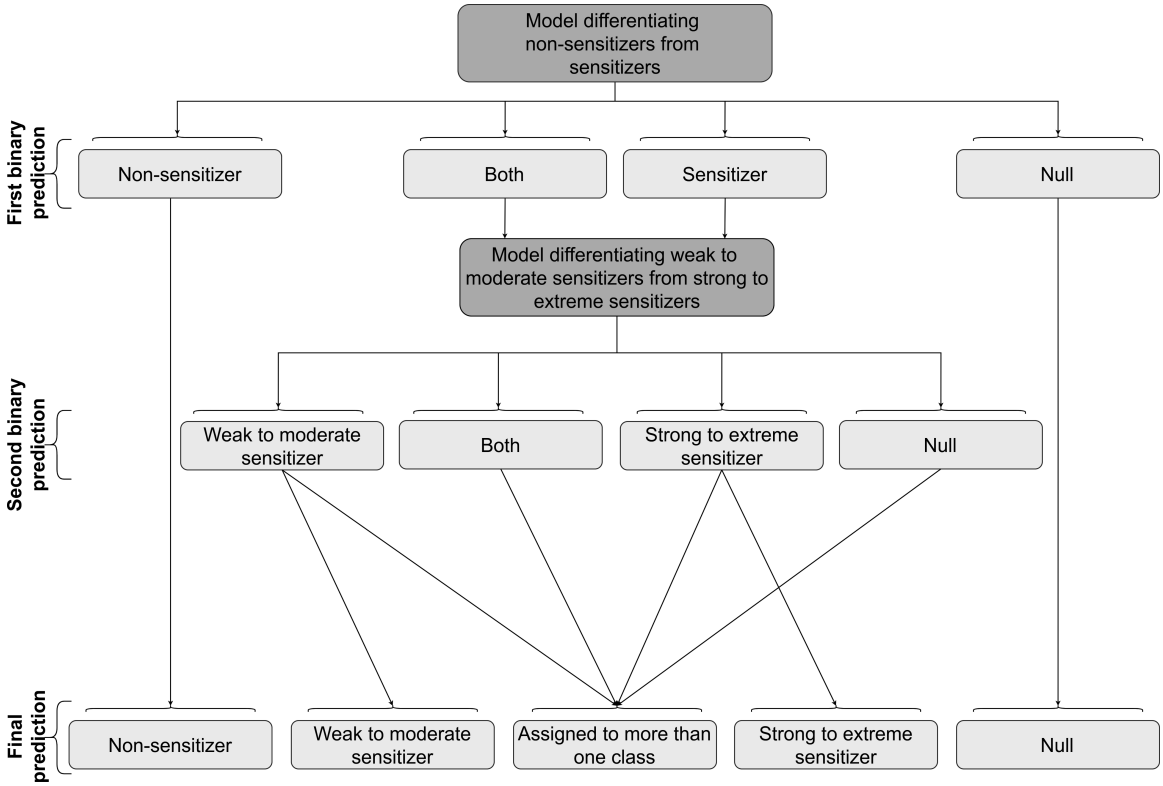


Figure 3. Schematic overview of the workflow underlying the ternary prediction of the skin sensitization potential of compounds. In the first step, the binary model differentiating nonsensitizers from sensitizers (as described in the subsection "Development of Binary Classifier for Predicting Skin Sensitization Potential") is applied to a compound. Depending on the p-values and the selected significance level (a compound is considered to belong to a certain class if the corresponding p-value exceeds the selected error significance), the compound is labeled "sensitizer", "non-sensitizer", "both", or "null". For compounds labeled "non-sensitizer" or "null", these predictions are final. Compounds labeled "sensitizer" or "both" are forwarded to a second model for the discrimination of weak to moderate from strong to extreme sensitizers. Note that compounds labeled by the first binary classifier as "both" and labeled by the second binary classifier as "weak to moderate sensitizer" or "strong to extreme sensitizer" assigned to more than one class. Compounds labeled "sensitizer" by the first model and not assigned to any potency class by the second model are automatically labeled as both weak to moderate sensitizers and as strong to extreme sensitizers. This procedure is to ensure consistent predictions of the binary and the ternary classifiers. Note that this procedure increases the validity and decreases the efficiency of the second model (the performance measures validity and efficiency are explained in the section "Performance metrics").

However, because of the data set refinements described above (first and foremost, the removal of potentially problematic compounds), this means that the test set for the current study is effectively a subset of the previous work (test set present work: 257 compounds; test set previous work: 284 compounds). The 14 additional compounds that resulted from the splitting of two-component mixtures were added to the training set (training set present work: 1028 compounds; training set previous work: 1132 compounds). For both multiclass data sets, the same split into training and test sets was performed as on the binary data set. Thus, the training and test sets of the multiclass data sets are subsets of the training and test sets of the binary data set.

Each training set was divided into a proper training set (80%) and a calibration set (20%) by stratified random splitting with the train_test_split function of the model_selection module of scikit-learn[35] (data shuffling prior to data set splitting enabled), as shown in Figure 2. A random forest model was derived with scikit-learn from each proper training set (hyperparameters adopted from Wilm et al.,[32] with n_estimators = 1000, max_features = "sqrt", random_state = 43) and applied to the calibration set.

Model Development Approach. Two binary aggregated Mondrian CP models based on RF estimators were generated (technical details of the CP approach are provided in the next subsection): one classifier to distinguish nonsensitizers from sensitizers, and one classifier to distinguish weak to moderate sensitizers from strong to extreme sensitizers. The initial version of the classifier distinguishing nonsensitizers from sensitizers was evaluated on the respective training set within a 10-fold cross-validation framework. The second and final version of this classifier was trained on the full training set and evaluated on the corresponding test set. The performance of the final binary classifier was also evaluated on the quinary test set with regard to the quinary class membership. The classifier distinguishing weak to moderate sensitizers from strong to extreme sensitizers was trained and tested on all sensitizers included in the ternary training and test sets, respectively.

Finally, both classifiers were combined in a two-step workflow. First, the model distinguishing nonsensitizers from sensitizers (in its final version) is applied to each compound of interest. Compounds classified by that model as sensitizers (independent of the predicted class membership of the nonsensitizing class) are then subjected to predictions with the second classifier to distinguish weak to moderate sensitizers from strong to extreme sensitizers. The two-step workflow was evaluated by applying it to the ternary test set.

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

Technical Aspects of Conformal Prediction. Nonconformity scores (α-values) for the calibration and test data were calculated based on the following nonconformity function for each class i:

$$\alpha_i = 0.5 - \frac{\hat{P}(y_i|x_i) - max_{y \neq y_i}\hat{P}(y|x_i)}{2}$$

with $\hat{P}(y_i|x_i)$ being the class probability for class i returned by the RF model, and $max_{y \neq y_i}\hat{P}(y|x_i)$ being the maximum class probability for any other class returned by the RF model.

The α-values for each class (nonsensitizers and sensitizers, or weak to moderate sensitizers and strong to extreme sensitizers) from the calibration set were sorted, and p-values for each class were derived for each test compound based on the rank of the corresponding α-value of the test compound.

This procedure to derive p-values for each compound of the test set by developing a RF model on the proper training data and applying it to the calibration and test sets was repeated 100 times with different splits into proper training and calibration data to achieve aggregated CP. This was realized by random states (ranging from 0 to 99) assigned to the function used to split the data into a proper training and a calibration set. All 100 models were applied to the test data, and the median p-values from all 100 runs were used as the final p-values for the test data.

If the p-value of a test compound for a given class exceeded the selected significance level ε, the compound was assigned to that class. A compound may be assigned to a single class, to several classes, or to no class, depending on the p-values and significance level.

Combined Workflow for Prediction of Ternary Skin Sensitization Potential. Finally, the two binary models were integrated into a workflow for the ternary classification of the skin sensitization potential of compounds (Figure 3).

Within the workflow, the binary model is first applied to distinguish nonsensitizers from sensitizers. If this model assigns a compound to the sensitizer class (note that the compound may, in addition, be assigned to the nonsensitizer class), it is forwarded to the second classifier to differentiate weak to moderate from strong to extreme sensitizers. To result in a ternary prediction, the predictions of the two classifiers are combined in an array of three values (Booleans), one for each potency class. The selection rules of this process are illustrated in Figure 3.

Performance Metrics. For all models, the CP-specific measures validity and efficiency were used for evaluation. In the context of CP, validity is defined as the percentage of predictions that include the true class of a compound. For a binary model, this includes distinct predictions (i.e., predictions that predict exactly one class to be true) for the true class as well as predictions that state both classes are true. Analogous to a classical model, which returns correct predictions with a defined reliability only for compounds that are within the AD of the model, predictions made by a CP model can be considered valid as long as the correct label is part of the returned prediction set. The percentage of compounds for which a distinct prediction is obtained is quantified as efficiency. As such, efficiency is equivalent with the definition of coverage found for most non-CP models in the field of toxicity prediction (and also consistent with the definition of coverage used for the non-CP version of Skin Doctor).[32] Analogous to the definition of the AD in classical models, validity and efficiency were calculated based on all predictions. In addition, the values of the general performance measures accuracy (ACC), Matthews correlation coefficient (MCC),[36] correct classification rate (CCR, also known as balanced accuracy), sensitivity (SE), specificity (SP), positive predictive value (PPV) and negative predictive value (NPV) were calculated based on all distinct predictions (i.e., all predictions that assigned a compound to exactly one activity class). For the binary as well as for the ternary model, class-wise validity and efficiency are the validity and efficiency measured on a subset of the tested compounds that have been experimentally determined to belong to the particular potency class.

For the ternary model, we consider both overall and class-wise performance, whereby overall performance refers to the mean values for each of the performance measures from the three potency classes. Class-wise performance measures are calculated individually for each potency class. In the cases of the non-CP performance measures (ACC, MCC, CCR, SE, SP, NPV, and PPV), class-wise performance values are calculated by combining all experimental and predicted class labels not belonging to the class of interest so that the performance measure can be calculated as if defined for two classes.

RESULTS

Development of Binary Classifier for Predicting Skin Sensitization Potential. The processed and refined data sets of Alves et al. and Di et al. comprise binary activity data for a total of 946 and 909 substances, respectively. Among those, 562 substances are listed in both data sets. After duplicate removal (during which 7 unique substances, distributed over 15 entries, were removed because of conflicting class labels), the (final) binary data set comprises 760 nonsensitizers and 525 sensitizers. The prepared data set was divided into a training set (610 nonsensitizers and 418 sensitizers) and a test set (150 nonsensitizers and 107 sensitizers) for model development and evaluation, respectively (Table 1).

Table 1. Composition of Binary Training and Test Data Sets

	training set	test set
nonsensitizers	610	150
sensitizers	418	107
total no. compounds	1028	257

Generation of Initial Binary Classifier and Its Performance during Cross-validation. An initial binary classifier was trained on a set of 610 nonsensitizers and 418 sensitizers and tested within a 10-fold cross-validation framework. The model was valid at all of the four tested significance levels (ε = 0.05, 0.10, 0.20 and 0.30), meaning that the validity was equal or close to $1-\varepsilon$. The standard deviations of the model validity and efficiency were all below 0.04 and 0.05 (Table 2). The highest standard deviation for each value was generally observed for ε = 0.05. This observed trend is related to the comparably small number of compounds for which the model returns unambiguous results at this significance level.

Some of the models were overconservative (i.e., the validity was higher than $1-\varepsilon$), which is a known phenomenon of aggregated CP classifiers at low significance levels ($\varepsilon \leq 0.40$) and is caused by an insufficient ability to properly rank the compounds of interest based on the selected nonconformity measure or one of the factors (modeling algorithm, type of descriptors, etc.) contributing to it. Overconservativeness of the model does not call into question the validity of the model and might, on the contrary, be favorable with respect to the reliability of predictions. Nevertheless, due to the trade-off between error rate and efficiency with regard to choice of significance level, overconservativeness coincides with an unnecessarily low efficiency for the selected significance level.[37]

At a significance level of 0.05, the model obtained an ACC of 0.88 and an MCC of 0.73 during cross-validation, with an efficiency of 0.28. At a significance level of 0.30, predictions could be made for almost all test compounds (96%), at the cost of a reduced ACC and MCC (0.76 and 0.51, respectively). Predictions of compounds being nonsensitizers were very reliable. For significance levels from 0.05 to 0.30, the NPVs were between 0.93 and 0.82, indicating that the model could be particularly valuable in a regulatory context where harmful

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

Table 2. Overall Performance during 10-Fold Cross-validation of Binary Aggregated Mondrian CP Classifier Differentiating Nonsensitizers from Sensitizers[1]

ε	validity	efficiency	ACC	MCC	CCR	SE	SP	NPV	PPV
0.05	0.96 (0.02)	0.28 (0.05)	0.88 (0.07)	0.73 (0.15)	0.87 (0.08)	0.86 (0.13)	0.89 (0.09)	0.93 (0.06)	0.80 (0.14)
0.10	0.91 (0.02)	0.51 (0.05)	0.83 (0.03)	0.66 (0.06)	0.84 (0.03)	0.84 (0.07)	0.83 (0.06)	0.89 (0.05)	0.76 (0.05)
0.20	0.82 (0.03)	0.83 (0.04)	0.78 (0.03)	0.56 (0.07)	0.78 (0.04)	0.78 (0.09)	0.78 (0.05)	0.84 (0.05)	0.71 (0.04)
0.30	0.73 (0.04)	0.96 (0.02)	0.76 (0.03)	0.51 (0.06)	0.76 (0.03)	0.76 (0.06)	0.76 (0.05)	0.82 (0.04)	0.69 (0.05)

[1]Standard deviation in brackets next to the values.

properties of substances in question should be ruled out with high reliability.[38] While for the four investigated significance levels only minor differences were observed for SE (between 0.76 and 0.86) and SP (between 0.76 and 0.89), the PPV (between 0.69 and 0.80) was lower than the NPV (between 0.82 and 0.93). Therefore, a negative prediction (non-sensitizer) made by the model seems to be more reliable than a positive prediction (sensitizer).

We also investigated model efficiency as a function of the selected significance level (Figure 4). Efficiency is found to increase steeply with low significance levels, reflecting the ability of the model to make distinct, single label predictions for an increasing amount of compounds (if we allow an increasing amount of erroneous predictions). Maximum efficiency is reached at a significance level of 0.28. Beyond this significance level, efficiency again decreases. This reflects the fact that the CP model will always guarantee an error rate close to the significance level. If, for example, a significance of 0.5 is desired (which in the binary case corresponds to a random model), predictions must be assigned to the empty class to fulfill this criterion (since the underlying model would have a better predictivity than 0.5).

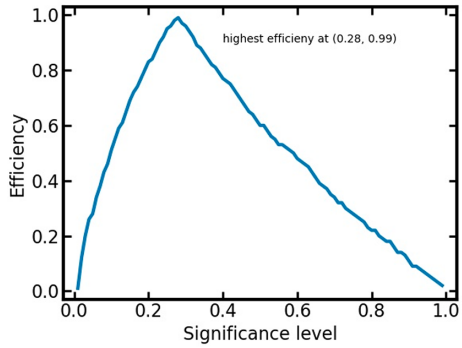


Figure 4. Efficiency of the binary classifier differentiating non-sensitizers from sensitizers within 10-fold CV in dependence of the significance level.

Generation of Final Binary Classifier and Its Performance on the Test Set. Following the CV studies, a final binary classification model, which we call "Skin Doctor CP", was trained on the full training set and evaluated on a test set of 150 nonsensitizers and 107 sensitizers (Figure 1). The final p-values of the test set compounds can be found in Table SI_4.

Overall Performance on the Test Set. The model was valid for all four significance levels (Table 3). Although the validity at the significance level of 0.3 was only 0.69, which is 0.01 lower than the expected validity of $1-\varepsilon$, this value is within the standard deviation observed for validities within CV. Therefore we assume that this slight under-predictivity is caused by statistical fluctuations and consider the model to be valid. The validity and efficiency of the final model were comparable to the values for the initial model (Table 3). The NPV (0.94 to 0.84) and SE (0.91 to 0.81) were higher than the PPV (0.83 to 0.65) and SP (0.88 to 0.70) for all of the four significance levels. While SE and NPV only slowly decreased with increasing error significance (ΔSE = 0.10 and ΔNPV = 0.10 between significance levels 0.05 and 0.30), SP and PPV decreased more drastically (ΔSP = 0.18 and ΔPPV = 0.18 over the range of significance levels). Therefore, negative predictions produced by this model can be considered reliable at all significance levels investigated, while positive predictions should be considered less reliable at high significance levels.

The confusion matrices of the model (Figure 5) reveal that the decrease in PPV observed with increasing error significance originates from an increasing tendency of the model to predict a compound to be a sensitizer (42%, 48%, 49%, and 51% of the molecules were predicted to be sensitizers at an significance level of 0.05, 0.10, 0.20, and 0.30, respectively), while the percentage of experimentally determined sensitizers remained comparably stable, between 38% and 41%.

Class-Wise Performance on the Test Set. To better understand the performance of the model within the CP setting, the class-wise validity and efficiency (i.e., the model's validity and efficiency calculated separately for each class of compounds, nonsensitizers and sensitizers, in the test set) of the binary classifier were analyzed for the selected significance levels (Table 4).

The validity of the model was higher for sensitizers than for nonsensitizers at all significance levels. A slight preference of the model to produce positive predictions was observed that increased proportionally with the significance level. Nevertheless, the difference in model validity between nonsensitizers

Table 3. Overall Performance of Binary Aggregated Mondrian CP Classifier, Differentiating Nonsensitizers from Sensitizers, on the Test Set

ε	validity	efficiency	ACC	MCC	CCR	SE	SP	NPV	PPV
0.05	0.96	0.32	0.89	0.78	0.89	0.91	0.88	0.94	0.83
0.10	0.91	0.49	0.83	0.66	0.84	0.90	0.78	0.92	0.72
0.20	0.82	0.79	0.77	0.55	0.78	0.84	0.72	0.88	0.65
0.30	0.69	0.92	0.75	0.51	0.76	0.81	0.70	0.84	0.65

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

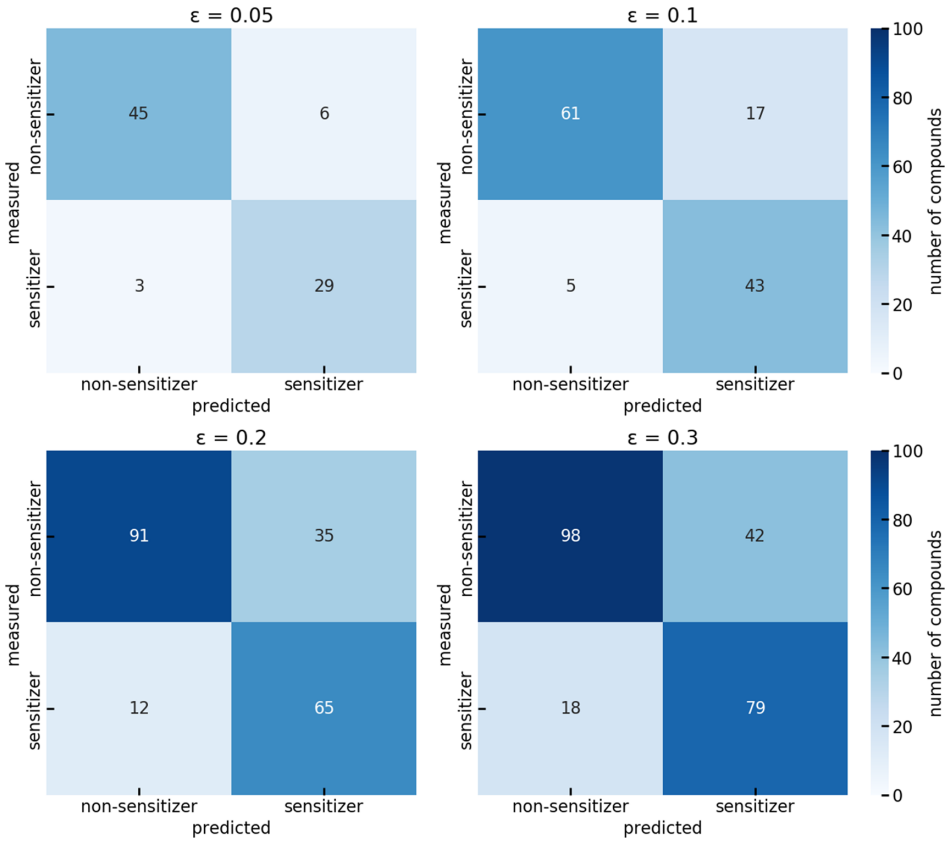


Figure 5. Confusion matrices reporting the classification results for the final binary classifier on the test set.

Table 4. Class-Wise Performance of Binary Classifier Differentiating Nonsensitizers from Sensitizers on the Test Set

ε	class	validity	efficiency
0.05	nonsensitizer	0.96	0.34
	sensitizer	0.97	0.30
0.10	nonsensitizer	0.89	0.52
	sensitizer	0.95	0.45
0.20	nonsensitizer	0.77	0.84
	sensitizer	0.89	0.72
0.30	nonsensitizer	0.65	0.93
	sensitizer	0.74	0.91

and sensitizers was relatively small and was highest (0.12) at the significance level of 0.20.

The model was valid for the sensitizer class at all four significance levels. For the nonsensitizer class, the model was valid at the significance level of 0.05 and only slightly underpredictive at the significance levels of 0.10 and 0.20. Since the deviation from the expected validity is only 0.01 and 0.03, which is within the standard deviations observed for the validity of the models during cross-validation, we nevertheless consider the model as valid for both classes at the significance levels of 0.10 and 0.20. At the significance level of 0.30, the validity of the nonsensitizing class was only 0.65. Because the deviation from the expected validity of 0.70 is not within the standard deviation observed during cross-validation (0.04), we assume that this might not only be caused by statistical fluctuations but might also originate from an underlying systemic problem of the model. We therefore suggest that predictions of sensitizer at this significance level be handled with care.

Differences in efficiency between both classes were similar to the differences observed for validity. The maximum difference in efficiency (0.12) was found at the significance level of 0.20.

Analysis of Performance of Final Binary Classifier Based on Quinary LLNA Data. False predictions are of varying degrees of concern, depending on the specific application scenario. In the regulatory context, false negative predictions will be of primary concern, whereas false positive predictions during the discovery phase may lead to a costly false deprioritization of compounds. Moreover, there is a distinction to be made between the false prediction of a weak skin sensitizer as nonsensitizer, and the false prediction of an extreme sensitizer as nonsensitizer. These types of distinction were examined using the quinary LLNA data (Figure 6).

Quinary LLNA data are available for 124 nonsensitizers, 37 weak sensitizers, 29 moderate sensitizers, 10 strong sensitizers, and 9 extreme sensitizers in the test set. At the significance levels of 0.05, 0.10, 0.20, and 0.30, a distinct prediction could be made for 22%, 53%, 90%, and 82% of compounds in this subset of the binary test set, respectively.

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

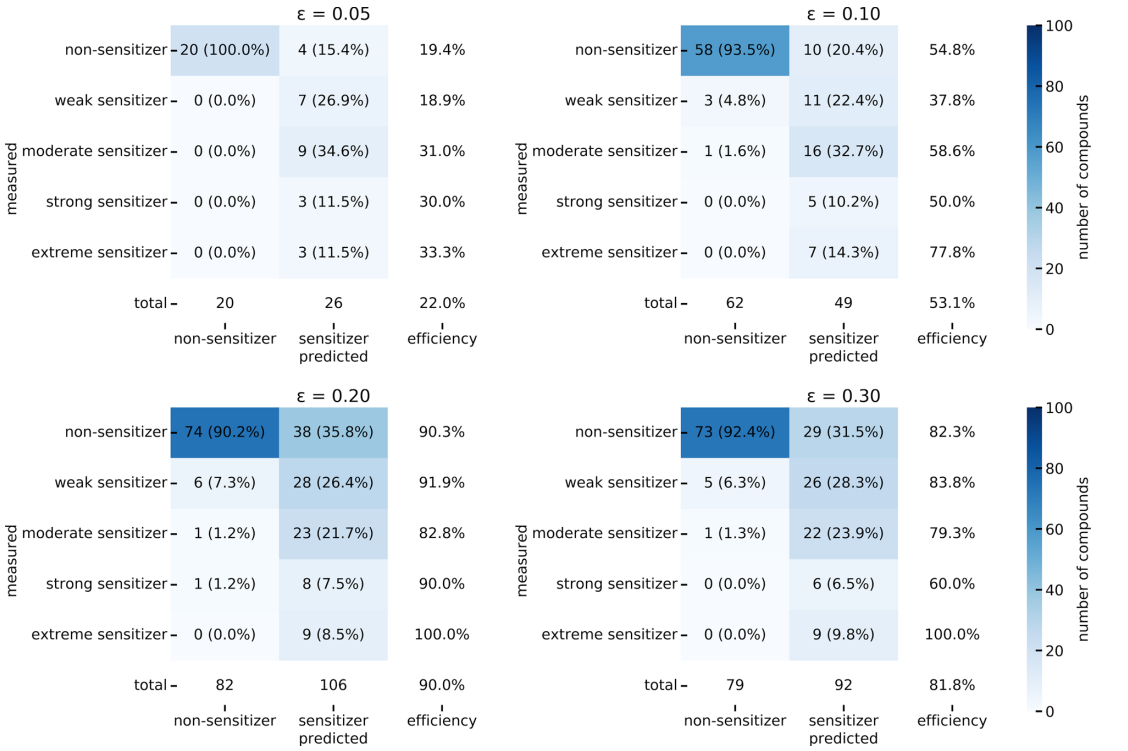


Figure 6. Distribution of the five potency classes among compounds predicted as nonsensitizers or sensitizers by the final binary classifier differentiating nonsensitizers from sensitizers. The percentages reported in parentheses refer to the total number of compounds reported in each column.

The PPV of the quinary subset ranges from 85% at the significance level of 0.05 to 64% and 68% at the significance levels of 0.20 and 0.30. Compounds predicted as nonsensitizers are correctly classified in 90% to 100% of the cases (NPV). At all significance levels investigated, the majority of sensitizers falsely predicted to belong to the nonsensitizing class belong to the class of weak sensitizers. One moderate sensitizer (CAS No. 5205−93−6, an amino functional methacrylamide monomer that is a known skin irritant) was falsely predicted as nonsensitizers at the significance levels of 0.10 or higher. In addition, a strong sensitizer (CAS No. 106359−91−5, a complex naphthalenetrisulfonic acid dye and known skin irritant) has been misclassified as a nonsensitizer at the significance level of 0.20. No extreme sensitizers have been misclassified. Thus, there seems to be an inverse trend between the potency of a sensitizer and the likelihood of it being falsely predicted as a nonsensitizer, which is an encouraging result.

Analysis of Performance of Final Binary Classifier with Respect to Functional Groups Present in the Test Compounds. Using a collection of 309 SMARTS patterns representing functional groups in organic chemistry, we identified 35 such groups that are presented in at least 20 compounds of the test set (Table SI_5). At the significance level 0.3, the binary classifier was found to perform particularly well (ACC values between 0.83 and 0.90) on compounds that contained at least one of the following functional groups: 1,5-tautomerizable moiety, amide, phenol, ketone, primary alcohol, secondary amide, sulfonic acid (derivative), or carboxylic acid (derivative). Among those, phenols are a particularly interesting case as the number of nonsensitizers and sensitizers among this group is nearly balanced (59% vs 41%). The model correctly identified 10 nonsensitizers and 9 sensitizers while only assigning three nonsensitizers and no sensitizer to the wrong activity class. Note that the model assigns 19% of the phenols to the empty class, which is the highest percentage of empty predictions among the 35 selected functional groups.

In contrast, we found low rates of prediction accuracies (between 0.56 and 0.67) for compounds comprising a heteroaromatic ring system with a nonbasic nitrogen atom, carboxylic esters, and dialkylethers (for the individual groups of compounds the ratio between nonsensitizers and sensitizers is well balanced).

The tendencies observed for the significance level of 0.3 could also be recognized for the other significance levels that we investigated but are based on weaker statistics.

Comparison of Model Performance with Skin Doctor. The binary classifier enveloped in the CP framework presented in this work was developed using the identical machine learning method and hyperparameters as in one of the previously reported "Skin Doctor" models.[32] However, Skin Doctor CP is trained on a modified training set and tested on a subset of the test set compared to the original Skin Doctor models. This limits direct comparability between the two approaches. Nevertheless, a qualitative comparison was performed here to estimate the main differences between the two approaches. Whereas the CP model allows the definition of the error significance level, the Skin Doctor model ("non-CP model")

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

Table 5. Overall Performance of Corresponding Non-CP Model "Skin Doctor", Differentiating Nonsensitizers from Sensitizers, on the Test Set

AD cutoff[1]	coverage[2]	ACC	MCC	CCR	SE	SP	NPV	PPV
0	1.0	0.72	0.41	0.70	0.57	0.82	0.74	0.69
≥0.5	0.96	0.73	0.43	0.71	0.60	0.82	0.75	0.69
≥0.75	0.28	0.78	0.59	0.81	0.89	0.73	0.92	0.64

[1]Defined as the mean Tanimoto similarity to the five nearest neighbors. [2]Coverage of the classical Skin Doctor is defined as the percentage of compounds in the test set that lie within the AD (i.e., for which a reliable prediction can be made by the model). This can be considered comparable to the definition of efficiency applied in this work, which is defined as the percentage of distinct predictions.

features an AD definition that is based on the Tanimoto coefficient, calculated using Morgan2 fingerprints and averaged over the five nearest neighbors in the training set. Any compound with a Tanimoto coefficient below a threshold (usually 0.5) is considered to be outside of the AD.

When a Tanimoto coefficient of 0.5 is applied as the threshold for the AD, the classical Skin Doctor model yields a coverage of 0.96 for the test set (Table 5), which is comparable to the efficiency of the CP model at a significance level of 0.3 (efficiency 0.92). In this setting, the classical Skin Doctor model obtained an ACC of 0.73 and an MCC of 0.43, which is comparable to the performance of the CP model (ACC = 0.75, MCC = 0.51). When increasing the threshold of the AD to 0.75, the classical Skin Doctor model yielded a coverage of 0.28. This is comparable to the efficiency of the CP model at a significance level of 0.05 (efficiency 0.32). In this setup, the CP model clearly outperformed the non-CP model by obtaining an ACC of 0.89 (vs 0.78) and an MCC of 0.78 (vs 0.59). At a significance level of 0.2, the performance of the CP model is comparable to that of the non-CP model with the strict definition of the AD (ACC 0.77 vs 0.78 and MCC 0.55 vs 0.59), despite superior efficiency/coverage (0.79 vs 0.28).

In-depth analysis of model performance showed that for the non-CP model the NPV increases with a stricter definition of the AD, whereas the PPV does not. This means that a stricter definition of the AD improves the reliability of the negative predictions but not of the positive ones. Within Skin Doctor CP, an increase in NPV from 0.84 to 0.94 and in PPV from 0.65 to 0.83 with decreasing error significance from 0.3 to 0.05 was found. Therefore, the use of Skin Doctor CP should in general be advantageous over the use of the non-CP models of Skin Doctor.

Development of Ternary Classifier for Predicting Skin Sensitization Potential. In an attempt to extend the capabilities of the machine learning approach to distinguish between three potency classes (nonsensitizer, weak to moderate sensitizer, and strong to extreme sensitizer), the feasibility of a two-step ternary model was explored, in which the (final) binary classifier forwards all compounds predicted as sensitizers to a downstream binary classifier to discriminate weak to moderate sensitizers from strong to extreme sensitizers. To ensure the validity of the two-step approach, the downstream binary model as well as the combined workflow was evaluated separately using (a subset of) the ternary data set. The composition of the full ternary training and test sets is shown in Table 6.

The binary classifier distinguishing weak to moderate sensitizers from strong to extreme sensitizers was developed following the same protocol and identical hyperparameters as described for the binary model distinguishing nonsensitizers from sensitizers (RF with 1000 estimators, enveloped by aggregate Mondrian CP; see Methods for details). This second

Table 6. Composition of Ternary Training and Test Data Sets

	training set[1]	test set[2]
nonsensitizer	510	124
weak to moderate sensitizer	279	66
strong to extreme sensitizer	65	19
total no. compounds	854	209

[1]Compared to the binary training set, 173 compounds have been removed because of missing multiclass labels and one compound has been rejected because of conflicting ternary class labels. [2]Compared to the binary test set, 47 compounds have been removed because of missing multiclass labels and one compound has been rejected because of conflicting ternary class labels.

model was trained and evaluated on subsets of the ternary training and test sets that comprise only sensitizing compounds. Within these subsets, 81% and 78% of the compounds in the training and test set belong to the class of weak to moderate sensitizers, respectively, while 19% and 22% of the compounds belong to the class of strong to extreme sensitizers, respectively. Unfortunately, the number of compounds in the training set (344) and test set (85) was relatively small and not sufficient to produce statistically solid evidence. The exact numbers in the following section should therefore not be considered reliable results. Rather, they should be considered as a proof of concept and an indication of a route that could be followed in the future with a larger database when more data become available.

Binary Classifier Distinguishing Weak to Moderate Sensitizers from Strong to Extreme Sensitizers. The binary model differentiating between weak to moderate sensitizers and strong to extreme sensitizers (for p-values see Table SI_4) was overconservative at all significance levels investigated (Table 7; validity = 0.94, 0.88, and 0.75 at significance levels of 0.10, 0.20, and 0.30; note that the significance level of 0.05 was not investigated since the efficiency of the model on the test set was 8%). As expected for an overconservative model, the efficiency of the model was comparably low (0.45, 0.71, and 0.98). At the three significance levels investigated, reasonably high values for SE (between 0.79 and 1.00) and SP (between 0.73 and 0.84) were found. The prediction that a compound is a weak to moderate sensitizer was highly reliable (NPV between 0.92 and 1.00) for all significance levels investigated, while a compound predicted to be a strong or extreme sensitizer could belong with almost equal probability to each of the two classes (PPV between 0.47 and 0.58). This strongly limits the model's applicability in any use case, but the model could be improved by a larger data set that includes a higher number of strong to extreme sensitizers when such data become available.

The observation of low PPV was also supported by the confusion matrices shown in Figure 7. The confusion matrices

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

Table 7. Overall Performance of Binary Model Distinguishing Weak to Moderate Sensitizers from Strong to Extreme Sensitizers on the Test Set

ε	validity	efficiency	ACC	MCC	CCR	SE	SP	NPV	PPV
0.10	0.94	0.45	0.87	0.70	0.92	1.00	0.84	1.00	0.58
0.20	0.88	0.71	0.83	0.63	0.87	0.92	0.81	0.97	0.57
0.30	0.75	0.98	0.75	0.45	0.76	0.79	0.73	0.92	0.47

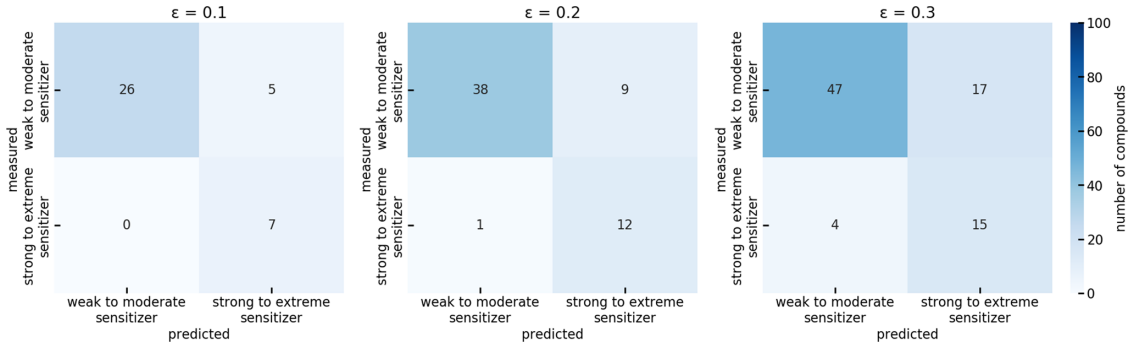


Figure 7. Confusion matrix of the binary model distinguishing weak to moderate sensitizers from strong to extreme sensitizers on the test set.

revealed that only 18% to 23% of the distinct predictions were made on strong to extreme sensitizers, which is the minority class.

The classifier is overall overconservative, which is also reflected in the class-wise validities, all of which are higher than $1-\varepsilon$ (Table 8). Class-wise validities and efficiencies are almost balanced between both classes, with a maximum difference of 0.09 and 0.10 in validity and efficiency, respectively.

Table 8. Class-Wise Performance of Binary Model Distinguishing Weak to Moderate Sensitizers from Strong to Extreme Sensitizers on the Test Set

ε	class	validity	efficiency
0.10	weak to moderate sensitizers	0.92	0.47
	strong to extreme sensitizers	1.00	0.37
0.20	weak to moderate sensitizers	0.86	0.71
	strong to extreme sensitizers	0.95	0.68
0.30	weak to moderate sensitizers	0.74	0.97
	strong to extreme sensitizers	0.79	1.00

Combined Workflow for Ternary Classification of Skin Sensitization Potential. Finally we combined, as a proof of concept, the two binary models in one workflow for the prediction of ternary skin sensitization potential and passed the resulting boolean array (storing the class membership of each compound to the three potency classes investigated) to our evaluation workflow. Within our test set, there was no case observed in which the first binary model predicted a compound to be a sensitizer but the second binary model predicted the compound to be neither a weak to moderate nor a strong to extreme sensitizer. We therefore believe there is no risk of artificially increasing the validity on this test set by reporting the validity and efficiency of the combined workflow.

Overall Performance on the Test Set. The combined workflow was valid overall, that is, in terms of the mean values among the three potency classes (overall validity = 0.92 and 0.80), at the significance levels of 0.10 and 0.20. At the error significance level of 0.30, the overall validity was only 0.66, which is 0.04 below the expected validity of 0.70. Although this value is still within the standard deviation observed for the significance level of 0.30 during 10-fold CV, it is larger than the deviations observed for other models and error significances within this work. We therefore cannot be sure that this underpredictiveness is only caused by statistical fluctuations and consider the validity of the model at the significance level of 0.30 as questionable.

The efficiency of the combined workflow (values between 0.42 and 0.90) was lower than or equal to the efficiency of the binary classifier differentiating between nonsensitizers and sensitizers (values between 0.49 and 0.92) at the three investigated significance levels (comparability of the two models is limited since the combined workflow is evaluated on only a subset of the data used for evaluation of the binary classifier) and lower than the efficiency of the binary classifier differentiating between weak to moderate and strong to extreme sensitizers (values between 0.45 and 0.98).

Satisfactory ACC values (from 0.90 to 0.73 for the significance levels investigated) and MCC values (from 0.78

Table 9. Overall Performance of Combined Workflow for Ternary Prediction of Skin Sensitization Potential on the Test Set[1]

ε	validity	efficiency	ACC	MCC	CCR	SE	SP	NPV	PPV
0.10	0.92	0.42	0.90	0.78	0.91	0.91	0.93	0.92	0.84
0.20	0.80	0.71	0.80	0.63	0.79	0.79	0.89	0.87	0.71
0.30	0.66	0.90	0.73	0.54	0.70	0.70	0.86	0.84	0.64

[1]All performance measures are reported as the mean of the corresponding performance measure over all classes investigated.

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

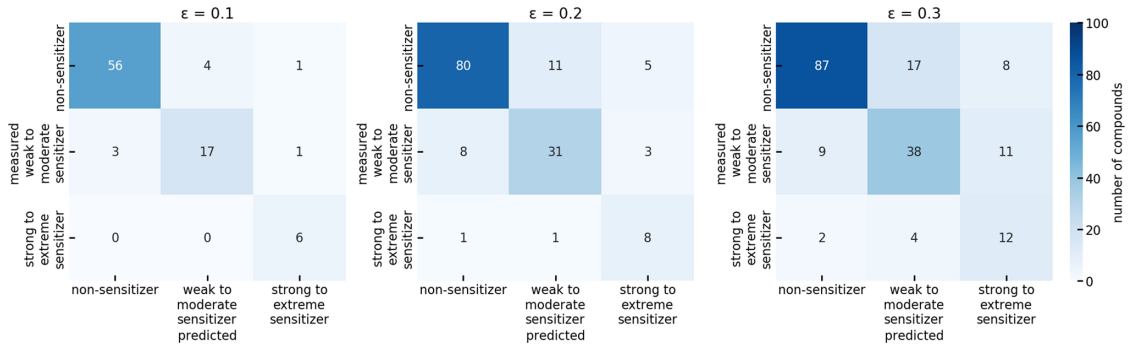


Figure 8. Confusion matrix obtained with the combined workflow for the ternary prediction of the skin sensitization potential of all compounds of the ternary test set.

Table 10. Class-Wise Performance of Combined Workflow for Ternary Prediction of Skin Sensitization Potential on the Test Set

ε	class	validity	efficiency	SE	SP	PPV	NPV
0.10	nonsensitizer	0.93	0.49	0.92	0.89	0.95	0.83
	weak to moderate sensitizer	0.88	0.32	0.81	0.94	0.81	0.94
	strong to extreme sensitizer	1.00	0.32	1.00	0.98	0.75	1.00
0.20	nonsensitizer	0.81	0.77	0.83	0.83	0.90	0.73
	weak to moderate sensitizer	0.74	0.64	0.74	0.89	0.72	0.90
	strong to extreme sensitizer	0.89	0.53	0.80	0.94	0.50	0.98
0.30	nonsensitizer	0.70	0.90	0.78	0.86	0.89	0.72
	weak to moderate sensitizer	0.58	0.88	0.66	0.84	0.64	0.84
	strong to extreme sensitizer	0.63	0.95	0.67	0.89	0.39	0.96

to 0.54 for the significance levels investigated) were achieved on the ternary test set (Table 9).

Analysis of the confusion matrices of the combined workflow on the test set (Figure 8) revealed that, at a significance level of 0.10 and 0.20, only 7% (6 out of 88 and 10 out of 148, respectively) of the compounds with distinct predictions were experimentally assigned as strong or extreme sensitizers. Thus, we expect the model to have limited impact on the prediction of strong to extreme sensitizers.

At a significance level of 0.30, which covers 90% of the test data, only 10% (18 out of 188) of the compounds were experimentally labeled as strong or extreme sensitizers. At the same time, 31 compounds were predicted to belong to this potency class. The likelihood of a compound predicted as being a strong or extreme sensitizer to belong to any of the three potency classes under investigation is almost equal for all three classes. A prediction with such a high false positive rate is not generally useful.

Class-Wise Performance on the Test Set. Since the low efficiency and the high false positive rate of strong to extreme sensitizers was not reflected by the overall performance measures, class-wise performance measures for each class of compounds were evaluated and summarized in Table 10.

At the significance levels of 0.10 and 0.20, the model was class-wise valid to over-predictive for nonsensitizers and strong to extreme sensitizers. With validities of 0.88 and 0.74, the model was slightly under-predictive for weak to moderate sensitizers at the significance levels of 0.10 and 0.20, respectively. We assume that the model can nevertheless be considered valid within the expected fluctuations on such a small data set. At a significance level of 0.30, the model was under-predictive for all classes investigated except non-sensitizers. With the validities for weak to moderate sensitizers and strong to extreme sensitizers being 0.58 and 0.63, the model must be considered invalid for these classes at the significance level of 0.30.

At all three significance levels, we observed a decrease in the PPV and an increase in the NPV from nonsensitizers to extreme sensitizers. These trends are related to the number of samples of each class in the training and test sets. The more samples of one class are present in a training set, the more reliable positive predictions and the less reliable negative predictions for that particular class become. While the PPV becomes unacceptably low (0.50 and 0.39) for strong to extreme sensitizers at significance levels of 0.2 and 0.3, respectively, the NPV stays reasonably high for all classes investigated (0.83 to 1.00 at ε = 0.10; 0.73 to 0.98 at ε = 0.20; 0.72 to 0.96 at ε = 0.30). Thus, a compound predicted to be a strong to extreme sensitizer most likely does not belong to that class, while the prediction that a compound is not a strong to extreme sensitizer can be considered reliable at all significance levels. This finding is supported by the reasonably high SE of strong to extreme sensitizers, indicating that 98%, 94%, and 89% of the strong and extreme sensitizers are correctly identified at the significance level of 0.10, 0.20, and 0.30, respectively. These tendencies also reflect the prevalence of the potency classes within the test set.

Within CP, a compound is assigned to a certain potency class if the corresponding p-value exceeds the selected significance level. Therefore, compounds with p-values in between the significance levels investigated will alter class membership when the significance level is altered. A prediction will be constant throughout all significance levels investigated, as long as the corresponding p-values are smaller than 0.10

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

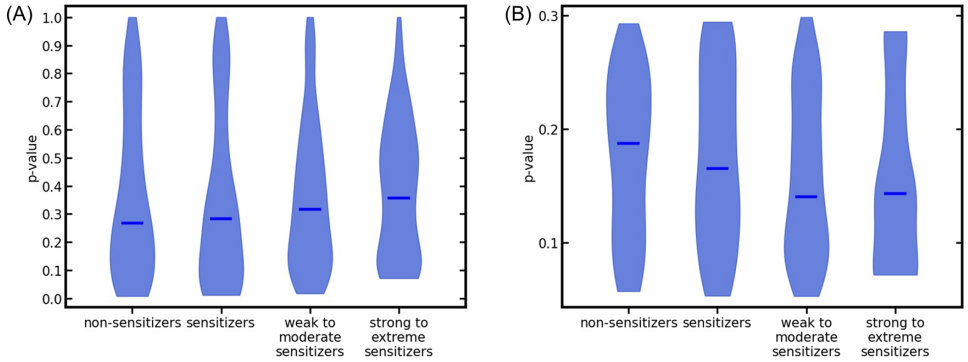


Figure 9. Violin plots of the distribution of p-values obtained for the ternary test set for the different classes of compounds as returned by the binary classification models: (A) complete test set; (B) detailed view of the p-value distributions close to the investigated significance levels, only considering p-values equal to or between 0.05 and 0.30. The median of the p-values for each potency class is indicated by a blue horizontal line.

(the lowest significance level investigated for the combined workflow) or larger than 0.30 (the highest significance level investigated in this work). The violin plots of the p-values returned by the two binary classifiers (Figure 9) visualize the distribution of p-values for each of the predicted classes within the ternary test set. All four distributions of p-values investigated show highest densities below 0.5. Compared with the two p-value distributions returned by the classifier that differentiates between nonsensitizers and sensitizers, the two distributions returned by the classifier differentiating between weak to moderate sensitizers and strong to extreme sensitizers comprise a lower percentage of compounds with p-values in extreme regions (below 0.05 or above 0.8). Thus, predictions are more likely to change depending on the significance level. The low-populated class of strong to extreme sensitizers intensifies this tendency compared to the weak to moderate sensitizing class.

Comparison of Ternary Classifier with Recently Published Model by Di et al. The data set of Di et al.[31] is one of the two data resources employed for the testing and development of Skin Doctor and Skin Doctor CP. Di et al. derived ternary in silico models for the prediction of the skin sensitization potential of compounds from their data. The model that they selected as their best model uses MACCS keys just like ours, but their modeling algorithm differs (CP+RF vs SVM), and although similar, the data sets used for training and testing by Di et al. and by us are not identical. This makes a direct comparison of both models difficult. Indicators suggest that the overall performance of both models is comparable. With a coverage of 98% of the compounds of the test set, the model of Di et al. was reported to obtain an ACC of 0.71, whereas our model, at a significance level of 0.30, obtained an ACC of 0.73 on our test set (see Table 11 for details). At this significance level, the efficiency of our model (90%) is lower than the coverage of the Di et al. model (98%; recall that we consider the efficiency of a CP classifier to represent a similar concept to the coverage of a non-CP model). The efficiency of our model decreases further at lower significance levels, to 42% and 71% at the significance levels of 0.10 and 0.20, respectively. However, at the significance levels of 0.10 and 0.20, our combined workflow exhibits higher overall performance (ACC = 0.90 and 0.80, respectively) than the Di et al. model (ACC = 0.71).

Table 11. Comparison of Overall Performance Measures of Best Ternary Model Reported by Di et al. and Our Combined CP Workflow for Ternary Classification Applied

	ACC	SE	SP	NPV	PPV	coverage/ efficiency
Di et al.	0.71	0.61	0.83	0.84	0.68	98%
combined CP workflow at significance level of 0.3	0.73	0.70	0.86	0.84	0.64	90%
our reconstruction of the model reported by Di et al. (without AD applied)	0.70	0.60	0.83	0.83	0.67	100%

From our investigations of the class-wise performance of our own ternary classifier, we know that its capacity to discriminate weak to moderate from strong to extreme sensitizers is insufficient. Since this limitation is mainly caused by a lack of LLNA data, we found it surprising that the ternary classifier of Di et al. seems to not suffer from this problem. Therefore, we reconstructed the ternary model published by Di et al. using the identical training and testing data, the identical type of descriptors (MACCS keys fingerprint) and the same modeling algorithm (SVM, probability = True, gamma = 0.125). For this reconstructed model, we found similar overall performances as reported by Di et al., who did not publish any values pertaining to the class-wise performance of their model. Like the original model of Di et al., the reconstructed model achieved an ACC of 0.80 on the external test set. On the test set, the reconstructed model achieved an ACC of 0.70 (further indicators: SE = 0.60, SP = 0.83, NPV = 0.83, and PPV = 0.67). Since we did not apply any AD, the reconstructed model has a coverage of 100%. Di et al. report a coverage of 98% on the test set and similar but slightly better performance measures (see above). Differences in performance might originate from our not applying any AD definition (in contrast to Di et al.) and the usage of different modeling software with perhaps different default values.

Of particular interest, however, is how the class-wise performance of the reconstructed model compares to that of our ternary classifier. This experiment reveals that the reconstructed Di et al. model suffers from class-wise unreliability just as our own ternary classifier does (Tables 12 and 13). The SE of the reconstructed Di et al. model is unsatisfyingly low for strong to extreme sensitizers

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

Table 12. Class-Wise Performance of Reconstructed Non-CP SVM MACCS Model on the Di et al. Test Set

class	SE	SP	PPV	NPV	number of compounds
nonsensitizer	0.79	0.71	0.65	0.83	33
weak to moderate sensitizer	0.72	0.79	0.76	0.76	39
strong to extreme sensitizer	0.30	0.97	0.60	0.91	10

Table 13. Class-Wise Performance of Reconstructed Non-CP SVM MACCS Model on the Di et al. External Test Set

class	SE	SP	PPV	NPV	number of compounds
nonsensitizer	0.88	0.6	0.88	0.60	461
weak to moderate sensitizer	0.59	0.88	0.54	0.90	115
strong to extreme sensitizer	0.05	0.98	0.10	0.96	22

(0.30 on the test set and 0.05 on the external test set). The confusion matrices (Figure 10) show that the model only very rarely predicts that a compound belongs to the class of strong to extreme sensitizers. This is a similar finding to what we observed with our own CP-based ternary classifier (see above; Table 10). These results indicate that also our reconstructed Di et al. model is unable to properly differentiate between the two classes of skin sensitizers.

CONCLUSION

In this work, we explored the scope and limitations of aggregated Mondrian CP in the development of approaches for the binary and ternary classification of compounds with respect to their skin sensitization potential. First, we developed and evaluated a binary classifier to differentiate nonsensitizers from sensitizers. The CP model was found to be valid for all classes at nearly all significance levels investigated and revealed to be favorable in terms of the portion of compounds for which a distinct or reliable prediction could be made compared to our previously published non-CP RF model that was trained and tested on the identical descriptors and a similar but slightly larger data set.

Second, we developed and tested a binary classifier that differentiates weak to moderate sensitizers from strong to extreme sensitizers based on a data set containing all sensitizing compounds with ternary class information from our ternary data set. Although the model was valid both overall and class-wise, and resulted in reasonable efficiencies, the model must be taken with caution due to the low quantity of data available for development and testing. The model was found to be not sufficiently reliable when being applied to strong to extreme sensitizers.

Finally, we integrated both binary classifiers within a combined workflow to result in a ternary prediction of the skin sensitization potential. We showed that the combined workflow, which was overall valid at the significance levels of 0.10 and 0.20, suffered from poor PPV for strong and extreme sensitizers at the significance levels of 0.20 and 0.30. This limits the ability of the model to correctly identify compounds belonging to that class. Investigation of a recent ternary model published by others[31] indicated that a low class-wise performance despite satisfying overall performance might also be a problem elsewhere and should be further investigated when publishing models developed using the currently available LLNA data.

From our studies, we conclude that aggregated Mondrian CP is a favorable approach for small and imbalanced data sets such as the LLNA data used in this work. This CP approach seems to be capable of improving the reliability and efficiency/coverage of binary classifiers for skin sensitization potential compared to non-CP approaches. In addition, CP offers the advantage of defined error rates that differentiate reliable from unreliable predictions without the need for a manually set threshold for a possible AD cutoff.

The ternary prediction of sensitizing potential would be highly relevant in a real-world setting. Our analysis has indicated that aggregated Mondrian CP provides benefits in efficiency and performance compared to the non-CP approach in this case as well. However, the amount of data currently available is unfortunately too small to properly distinguish different classes of sensitizing compounds, which strongly limits the applicability and reliability of the model. For better modeling, as well as for a statistically more solid evaluation of the model, more data (especially on strong and extreme sensitizers) are urgently needed.

Skin Doctor CP is available via a public web service at https://nerdd.zbh.uni-hamburg.de/skinDoctorII.

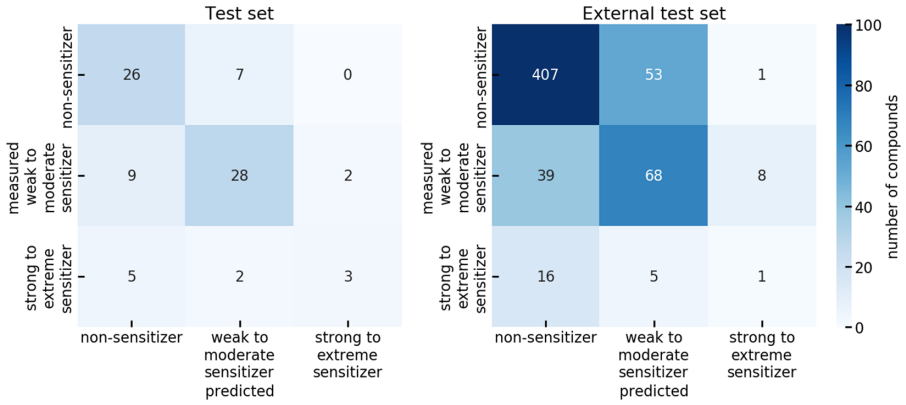


Figure 10. Confusion matrices of the reconstructed model of Di et al.

https://dx.doi.org/10.1021/acs.chemrestox.0c00253
Chem. Res. Toxicol. 2021, 34, 330−344

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrestox.0c00253.

Full Skin Doctor CP data set, including class labels and declaration of which substances were used for model training and for model testing; set of substances removed from original data set during manual data curation process; list of most common functional groups in data set and distribution of nonsensitizers and sensitizers within molecules containing this functional group; final p-values returned by two binary classifiers on binary and ternary test set; analysis of prediction accuracy for substances containing most common functional groups among test set (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Johannes Kirchmair − *Center for Bioinformatics (ZBH), Department of Informatics, Universität Hamburg, 20146 Hamburg, Germany; Department of Pharmaceutical Chemistry, University of Vienna, 1090 Vienna, Austria;* orcid.org/0000-0003-2667-5877; Phone: +43 1-4277-55104; Email: johannes.kirchmair@univie.ac.at

Authors

Anke Wilm − *Center for Bioinformatics (ZBH), Department of Informatics, Universität Hamburg, 20146 Hamburg, Germany; HITeC e.V., 22527 Hamburg, Germany;* orcid.org/0000-0003-2891-1407

Ulf Norinder − *Department of Computer and Systems Sciences, Stockholm University, SE-16407 Kista, Sweden; Department of Pharmaceutical Biosciences, Uppsala University, SE-75124 Uppsala, Sweden; MTM Research Centre, School of Science and Technology, Örebro University, SE-70182 Örebro, Sweden*

M. Isabel Agea − *Department of Informatics and Chemistry, University of Chemistry and Technology Prague, 16628 Prague, Czech Republic;* orcid.org/0000-0002-3017-7742

Christina de Bruyn Kops − *Center for Bioinformatics (ZBH), Department of Informatics, Universität Hamburg, 20146 Hamburg, Germany;* orcid.org/0000-0001-8890-2137

Conrad Stork − *Center for Bioinformatics (ZBH), Department of Informatics, Universität Hamburg, 20146 Hamburg, Germany;* orcid.org/0000-0002-5499-742X

Jochen Kühnl − *Front End Innovation, Beiersdorf AG, 22529 Hamburg, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrestox.0c00253

Funding

A.W. is supported by Beiersdorf AG through HITeC e.V. C.d.B.K. is funded by a PhD completion scholarship from the Universität Hamburg. C.S. and J.Ki. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number KI 2085/1−1. M.I.A. is supported by the Ministry of Education, Youth and Sports of the Czech Republic - project number LM2018130.

Notes

The authors declare the following competing financial interest(s): A.W. is funded by Beiersdorf AG through HITeC e.V. and J.K. is employed at Beiersdorf AG. A.W. and J.K. were involved in the design of the study, the interpretation of the data, the writing of the manuscript, and the decision to publish the results.

■ ACKNOWLEDGMENTS

The authors thank the Editor and the three anonymous expert reviewers for their valuable feedback.

■ ABBREVIATIONS

ACC, accuracy; AD, applicability domain; CCR, correct classification rate; CP, conformal prediction; MCC, Matthews correlation coefficient; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SE, sensitivity; SP, specificity

■ REFERENCES

(1) Kimber, I., Basketter, D. A., Gerberick, G. F., Ryan, C. A., and Dearman, R. J. (2011) Chemical Allergy: Translating Biology into Hazard Characterization. *Toxicol. Sci. 120* (Suppl 1), S238−S268.

(2) Lushniak, B. D. (2004) Occupational Contact Dermatitis. *Dermatol. Ther. 17*, 272−277.

(3) Thyssen, J. P., Linneberg, A., Menné, T., and Johansen, J. D. (2007) The Epidemiology of Contact Allergy in the General Population − Prevalence and Main Findings. *Contact Dermatitis 57*, 287−299.

(4) Felter, S., Kern, P., and Ryan, C. (2018) Allergic Contact Dermatitis: Adequacy of the Default 10X Assessment Factor for Human Variability to Protect Infants and Children. *Regul. Toxicol. Pharmacol. 99*, 116−121.

(5) OECD. (2010) *OECD Guidelines for the Testing of Chemicals, Section 4 Test No. 429: Skin Sensitisation Local Lymph Node Assay: Local Lymph Node Assay*, OECD Publishing.

(6) Anderson, S. E., Siegel, P. D., and Meade, B. J. (2011) The LLNA: A Brief Review of Recent Advances and Limitations. *J. Allergy 2011*, 424203−424213.

(7) Gerberick, G. F., House, R. V., Fletcher, E. R., and Ryan, C. A. (1992) Examination of the Local Lymph Node Assay for Use in Contact Sensitization Risk Assessment. *Fundam. Appl. Toxicol. 19*, 438−445.

(8) Leenaars, C. H. C., Kouwenaar, C., Stafleu, F. R., Bleich, A., Ritskes-Hoitinga, M., De Vries, R. B. M., and Meijboom, F. L. B. (2019) Animal to Human Translation: A Systematic Scoping Review of Reported Concordance Rates. *J. Transl. Med. 17*, 223.

(9) Hoffmann, S., Kleinstreuer, N., Alépée, N., Allen, D., Api, A. M., Ashikaga, T., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Goebel, C., Kern, P. S., Klaric, M., Kühnl, J., Lalko, J. F., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Parakhia, R., van Vliet, E., Zang, Q., and Petersohn, D. (2018) Non-Animal Methods to Predict Skin Sensitization (I): The Cosmetics Europe Database. *Crit. Rev. Toxicol. 48*, 344−358.

(10) Mehling, A., Eriksson, T., Eltze, T., Kolle, S., Ramirez, T., Teubner, W., van Ravenzwaay, B., and Landsiedel, R. (2012) Non-Animal Test Methods for Predicting Skin Sensitization Potentials. *Arch. Toxicol. 86*, 1273−1295.

(11) Reisinger, K., Hoffmann, S., Alépée, N., Ashikaga, T., Barroso, J., Elcombe, C., Gellatly, N., Galbiati, V., Gibbs, S., Groux, H., Hibatallah, J., Keller, D., Kern, P., Klaric, M., Kolle, S., Kuehnl, J., Lambrechts, N., Lindstedt, M., Millet, M., Martinozzi-Teissier, S., Natsch, A., Petersohn, D., Pike, I., Sakaguchi, H., Schepky, A., Tailhardat, M., Templier, M., van Vliet, E., and Maxwell, G. (2015) Systematic Evaluation of Non-Animal Test Methods for Skin Sensitisation Safety Assessment. *Toxicol. In Vitro 29*, 259−270.

(12) Ezendam, J., Braakhuis, H. M., and Vandebriel, R. J. (2016) State of the Art in Non-Animal Approaches for Skin Sensitization Testing: From Individual Test Methods towards Testing Strategies. *Arch. Toxicol. 90*, 2861−2883.

(13) Thyssen, J. P., Giménez-Arnau, E., Lepoittevin, J.-P., Menné, T., Boman, A., and Schnuch, A. (2012) The Critical Review of Methodologies and Approaches to Assess the Inherent Skin Sensitization Potential (skin Allergies) of Chemicals. *Contact Dermatitis 66* (Suppl 1), 11−24.
(14) Wilm, A., Kühnl, J., and Kirchmair, J. (2018) Computational Approaches for Skin Sensitization Prediction. *Crit. Rev. Toxicol. 48*, 738−760.
(15) ECHA (European Chemicals Agency). (2017) *The use of alternatives to testing on animals for the REACH regulation, third report under article 117(3) of the REACH regulation*, ECHA. https://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2017_en.pdf (accessed Jul 10, 2019).
(16) Jowsey, I. R., Basketter, D. A., Westmoreland, C., and Kimber, I. (2006) A Future Approach to Measuring Relative Skin Sensitising Potency: A Proposal. *J. Appl. Toxicol. 26*, 341−350.
(17) Safford, R. J., Api, A. M., Roberts, D. W., and Lalko, J. F. (2015) Extension of the Dermal Sensitisation Threshold (DST) Approach to Incorporate Chemicals Classified as Reactive. *Regul. Toxicol. Pharmacol. 72*, 694−701.
(18) OECD. (2004) *OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models*, OECD. https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf.
(19) Netzeva, T. I., Worth, A., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D., Schultz, T., Stanton, D. W., van de Sandt, J. J. M., Tong, W., Veith, G., and Yang, C. (2005) Current Status of Methods for Defining the Applicability Domain of (quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *ATLA, Altern. Lab. Anim. 33*, 155−173.
(20) Carrió, P., Pinto, M., Ecker, G., Sanz, F., and Pastor, M. (2014) Applicability Domain ANalysis (ADAN): A Robust Method for Assessing the Reliability of Drug Property Predictions. *J. Chem. Inf. Model. 54*, 1500−1511.
(21) Klingspohn, W., Mathea, M., Ter Laak, A., Heinrich, N., and Baumann, K. (2017) Efficiency of Different Measures for Defining the Applicability Domain of Classification Models. *J. Cheminf. 9*, 44−61.
(22) Vovk, V., Gammerman, A., and Shafer, G. (2005) *Algorithmic Learning in a Random World*, Springer Science & Business Media.
(23) Norinder, U., Carlsson, L., Boyer, S., and Eklund, M. (2015) Introducing Conformal Prediction in Predictive Modeling for Regulatory Purposes. A Transparent and Flexible Alternative to Applicability Domain Determination. *Regul. Toxicol. Pharmacol. 71*, 279−284.
(24) Norinder, U., Rybacka, A., and Andersson, P. L. (2016) Conformal Prediction to Define Applicability Domain − A Case Study on Predicting ER and AR Binding. *SAR and QSAR in Environmental Research 27*, 303−316.
(25) Cortés-Ciriano, I., and Bender, A. (2020) Concepts and applications of conformal prediction in computational drug discovery. *ArXiv*. https://arxiv.org/pdf/1908.03569.pdf (accessed 03-17-2020).
(26) Svensson, F., Afzal, A. M., Norinder, U., and Bender, A. (2018) Maximizing Gain in High-Throughput Screening Using Conformal Prediction. *J. Cheminf. 10*, 7.
(27) Norinder, U., and Svensson, F. (2019) Multitask Modeling with Confidence Using Matrix Factorization and Conformal Prediction. *J. Chem. Inf. Model. 59*, 1598−1604.
(28) Norinder, U., Ahlberg, E., and Carlsson, L. (2019) Predicting Ames Mutagenicity Using Conformal Prediction in the Ames/QSAR International Challenge Project. *Mutagenesis 34*, 33−40.
(29) Carlsson, L., Eklund, M., and Norinder, U. (2014) Aggregated Conformal Prediction. In *Artificial Intelligence Applications and Innovations*, Springer, pp 231−240.
(30) Alves, V. M., Capuzzi, S. J., Braga, R. C., Borba, J. V. B., Silva, A. C., Luechtefeld, T., Hartung, T., Andrade, C. H., Muratov, E. N., and Tropsha, A. (2018) A Perspective and a New Integrated Computational Strategy for Skin Sensitization Assessment. *ACS Sustainable Chem. Eng. 6*, 2845−2859.
(31) Di, P., Yin, Y., Jiang, C., Cai, Y., Li, W., Tang, Y., and Liu, G. (2019) Prediction of the Skin Sensitising Potential and Potency of Compounds via Mechanism-Based Binary and Ternary Classification Models. *Toxicol. In Vitro 59*, 204−214.
(32) Wilm, A., Stork, C., Bauer, C., Schepky, A., Kühnl, J., and Kirchmair, J. (2019) Skin Doctor: Machine Learning Models for Skin Sensitization Prediction That Provide Estimates and Indicators of Prediction Reliability. *Int. J. Mol. Sci. 20*, 4833−4856.
(33) Laggner, C. (2005) *SMARTS Patterns for Functional Group Classification*, Inte:Ligand Software-Entwicklungs und Consulting GmbH. https://github.com/openbabel/openbabel/blob/master/data/SMARTS_InteLigand.txt (accessed 10-02-2020).
(34) Landrum, G. (2019) *RDKit*, GitHub. http://www.rdkit.org (accessed 04-26-2019).
(35) (2019) *scikit-learn: machine learning in Python — scikit-learn 0.21.0 documentation*, scikit. https://scikit-learn.org/stable/ (accessed 05-10-2019).
(36) Matthews, B. W. (1975) Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta, Protein Struct. 405*, 442−451.
(37) Linusson, H., Norinder, U., Boström, H., Johansson, U., and Löfström, T. (2017) On the calibration of aggregated conformal predictors. *Proc. Mach Learn Res. 60*, 1−20.
(38) Williams, R. V., Amberg, A., Brigo, A., Coquin, L., Giddings, A., Glowienke, S., Greene, N., Jolly, R., Kemper, R., O'Leary-Steele, C., Parenty, A., Spirkl, H.-P., Stalford, S. A., Weiner, S. K., and Wichard, J. (2016) It's Difficult, but Important, to Make Negative Predictions. *Regul. Toxicol. Pharmacol. 76*, 79−86.

6.3 Maximising interpretability with a small selection of biologically meaningful descriptors

The acceptance of a computational model for risk assessment also depends on its transparency and interpretability [110]. For a ML model to be interpretable by human beings, a comprehensible ML algorithm as well as a manageable number of meaningful descriptors is needed.

In the following study, we utilized a set of 750 biologically meaningful descriptors which encodes the predicted probabilities of a compound being active or inactive in 375 different bioactivity assays. In order to decrease the number of descriptors drastically, we established a strict and iterative feature selection process utilizing a LASSO algorithm during a 10-fold CV protocol on our training data. A qualitative analysis of the ten final descriptors selected by the algorithm was undertaken with respect to the possible biological relationships between the descriptors (bioactivity assay activity) and the skin sensitization AOP. The relevance of the descriptors selected was demonstrated by application of PCA on our LLNA data set as well as on three reference data sets containing cosmetics, approved drugs, and pesticides within the chemical space spanned by the ten bioactivity descriptors. Furthermore, a PCA plot conducted of the LLNA data set clearly shows the ability of the descriptor set to distinguish between sensitizing and non-sensitizing substances.

As a proof of concept, the ten bioactivity descriptors, were applied in an aggregated Mondrian CP framework and revealed performance measures similar to the ones achieved in earlier studies, while increasing interpretability of the model.

P4 **Wilm, A.**, Garcia de Lomana, M., Stork, C., Mathai, N., Hirte, S., Norinder, U., Kühnl, J., Kirchmair, J., Predicting the skin sensitization potential of small molecules with machine learning models trained on biologically meaningful descriptors, *Pharmaceuticals*, **14**(8) (2021) 790

Available at:
https://www.mdpi.com/1424-8247/14/8/790

Author Contributions: A. Wilm, U. Norinder, J. Kühnl, and J. Kirchmair conceptualized the work. A. Wilm developed the methodologies, with contributions by M. Garcia de Lomana, N. Mathai, J. Kühnl, and J. Kirchmair. A. Wilm developed the underlying software, with contributions by M. Garcia de Lomana, N. Mathai, C. Stork, and S. Hirte. A. Wilm performed the validation of the models and findings, with contributions by J. Kühnl, and J. Kirchmair. A. Wilm, N. Mathai, and M. Garcia de Lomana performed the data curation. A. Wilm visualized most of the research results. S. Hirte visualized the learning curves. A. Wilm wrote

the manuscript, with contributions by M. Garcia de Lomana, C. Stork, N. Mathai, S. Hirte, U. Norinder, J. Kühnl, and J. Kirchmair. J. Kühnl, J. Kirchmair, and U. Norinder supervised the project. Resources, project administration and funding were provided or aquired by J. Kühnl and J. Kirchmair. All authors have read and agreed to the published version of the manuscript.

pharmaceuticals

Article

Predicting the Skin Sensitization Potential of Small Molecules with Machine Learning Models Trained on Biologically Meaningful Descriptors

Anke Wilm [1,2], Marina Garcia de Lomana [3], Conrad Stork [1], Neann Mathai [4], Steffen Hirte [3], Ulf Norinder [5,6,7], Jochen Kühnl [8] and Johannes Kirchmair [1,3,*]

[1] Center for Bioinformatics (ZBH), Department of Informatics, Universität Hamburg, 20146 Hamburg, Germany; wilm@zbh.uni-hamburg.de (A.W.); stork@zbh.uni-hamburg.de (C.S.)
[2] HITeC e.V., 22527 Hamburg, Germany
[3] Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria; a11853333@unet.univie.ac.at (M.G.d.L.); steffen.hirte@univie.ac.at (S.H.)
[4] Computational Biology Unit (CBU), Department of Chemistry, University of Bergen, N-5020 Bergen, Norway; neann.mathai@uib.no
[5] MTM Research Centre, School of Science and Technology, Örebro University, SE-70182 Örebro, Sweden; ulf.norinder@farmbio.uu.se
[6] Department of Computer and Systems Sciences, Stockholm University, SE-16407 Kista, Sweden
[7] Department of Pharmaceutical Biosciences, Uppsala University, SE-75124 Uppsala, Sweden
[8] Front End Innovation, Beiersdorf AG, 22529 Hamburg, Germany; Jochen.Kuehnl@Beiersdorf.com
* Correspondence: johannes.kirchmair@univie.ac.at; Tel.: +43-1-4277-55104

Citation: Wilm, A.; Garcia de Lomana, M.; Stork, C.; Mathai, N.; Hirte, S.; Norinder, U.; Kühnl, J.; Kirchmair, J. Predicting the Skin Sensitization Potential of Small Molecules with Machine Learning Models Trained on Biologically Meaningful Descriptors. *Pharmaceuticals* **2021**, *14*, 790. https://doi.org/10.3390/ph14080790

Academic Editor: Osvaldo Andrade Santos-Filho

Received: 14 July 2021
Accepted: 6 August 2021
Published: 11 August 2021

Abstract: In recent years, a number of machine learning models for the prediction of the skin sensitization potential of small organic molecules have been reported and become available. These models generally perform well within their applicability domains but, as a result of the use of molecular fingerprints and other non-intuitive descriptors, the interpretability of the existing models is limited. The aim of this work is to develop a strategy to replace the non-intuitive features by predicted outcomes of bioassays. We show that such replacement is indeed possible and that as few as ten interpretable, predicted bioactivities are sufficient to reach competitive performance. On a holdout data set of 257 compounds, the best model ("Skin Doctor CP:Bio") obtained an efficiency of 0.82 and an MCC of 0.52 (at the significance level of 0.20). Skin Doctor CP:Bio is available free of charge for academic research. The modeling strategies explored in this work are easily transferable and could be adopted for the development of more interpretable machine learning models for the prediction of the bioactivity and toxicity of small organic compounds.

Keywords: skin sensitization; toxicity prediction; in silico prediction; machine learning; random forest; conformal prediction; bioactivity descriptors

1. Introduction

Substances that can induce allergic contact dermatitis after repeated contact to the skin are called skin sensitizers [1,2]. In order to prevent the induction of skin sensitization, exposure to skin sensitizers must be minimized [3–8]. The ability to detect and predict skin sensitizers is therefore of significant importance for several sectors of industry to develop safe and efficacious functional small molecules [9].

Until recent years, strategies to assess the risk of small molecules to induce skin sensitization relied on animal experiments. Historically, an important animal experiment to address skin sensitization potential is the guinea pig maximization test (GPMT), which was used to determine the percentage of test animals that develop contact allergy symptoms after repeated exposure to the test substance. Typically, a substance was classified

as a sensitizer if at least 15% of the guinea pigs developed allergic symptoms. The GPMT was later replaced by the murine local lymph node assay (LLNA) [10], an animal model measuring the proliferation rate of cells in the draining lymph node in mice. The LLNA is still regarded as the gold standard among the animal experiments to assess skin sensitization potential as it provides advantages concerning animal welfare (compared to other animal models) and additional information to quantify the skin sensitization potency of compounds (based on the EC3 value, defined as the substance concentration that induces a 3-fold stimulation of proliferation) [11,12].

Ambitious efforts are ongoing to fully replace animal experiments, and a diverse set of alternative experimental and theoretical methods have been developed [13,14] to assess skin sensitization potential and, to a limited degree, skin sensitization potency [15]. Among others, these approaches include non-animal testing methods (i.e., in vitro and in chemico assays) [16–19] and in silico methods [18–21].

Several OECD-validated non-animal testing methods address three out of four key events of the adverse outcome pathway of skin sensitization induction: The first key event, or molecular initiating event, describes the so-called haptenization, which is the covalent binding of the substance to skin proteins or peptides. This is experimentally assessed by the direct peptide reactivity assay (DPRA) [22]. The second key event, which is the activation of keratinocytes [23], is covered by the KeratinoSens and LuSens assays, while the third key event, which is the activation of the skin's dendritic cells [24], is addressed, among others, by the U937 cell line activation test (U-SENS) and the human cell line activation test (h-CLAT). As all of these assays cover certain aspects of the adverse outcome pathway; none of them is suitable as a standalone methodology for the prediction of the skin sensitization potential of small molecules.

Computational methods that predict skin sensitization can be classified into expert systems, similarity-based approaches, and (quantitative) structure–activity relationship (QSAR) approaches [20]. These approaches offer fast predictions at low cost, enabling their use also in early stages of research and development, where a large number of candidate compounds may be under investigation. To be accepted as a component of regulatory risk assessment, computational methods have to fulfill certain quality criteria. For example, according to the OECD [25], a model should have a defined endpoint, an unambiguous algorithm, a defined applicability domain, appropriate measures of goodness-of–fit, robustness, and predictivity, and, if possible, a mechanistic interpretation.

No particular non-animal testing method or individual computational model has so far yielded a level of performance, robustness, interpretability, and coverage to be accepted as a standalone approach for skin sensitization prediction in the regulatory context. The most promising strategy to advance alternative testing methods is the combination of experimental and computational tools [26] within defined approaches, integrated approaches for testing and assessment (IATAs; for a review of IATAs and defined approaches see ref. [27]), or in "weight of evidence" considerations [28].

In our previous work [29], we presented Skin Doctor CP, a random forest (RF) model for the prediction of LLNA outcomes for small molecules that complies with the abovementioned OECD principles to the furthest possible extent. The Skin Doctor CP model is trained on a set of 1278 compounds annotated with binary LLNA outcomes (i.e., skin sensitizer and skin non-sensitizer). To the best knowledge of the authors, this data set represents the largest collection of high-quality LLNA data in the public domain at present. The data set has been characterized regarding its composition and chemical space coverage [29]. The RF model derived from this data set is wrapped into an aggregated Mondrian conformal prediction (CP) framework, which ensures predictivity and robustness by a mathematically founded measure of reliability [30–32]. More specifically, the CP framework guarantees an observed prediction error of the model close to the error rate ε set by the user (this is as long as the randomness assumption of the samples holds true; an assumption that is also made for any classical machine learning model). The CP framework will only return a predicted class membership for a substance if the prediction lies within

the desired confidence level 1-ε. The measure of reliability offered by the CP approach can guide the use of Safety Assessment Factors of different levels and serve as a powerful, mathematically founded alternative to applicability domain definitions [33].

Depending on the available data and computational capacities, different variants of CP may be developed [34]. In the case of LLNA prediction, the data available for model development are limited and imbalanced; hence, the use of an aggregated CP framework is advised. The aggregated CP framework repeats the framework several times with different proper training and calibration sets [35]. This reduces the variance in the model predictions and allows every datapoint of the training set to be used for model development. It is therefore best suited for modeling small data sets.

To address data imbalance in addition to data scarcity (such as in the case of the LLNA data modeled in our previous study), the combination of the aggregated CP framework with Mondrian CP is advised. Mondrian CP is tailored to describe imbalanced data as it treats each of the classes independently and ensures the validity of their predictions [36–38]. This is especially beneficial in toxicity prediction, where the toxic class is usually the minority class and therefore more difficult to predict [39].

In addition to the OECD requirement for a model to produce results with defined reliability (which we address by using a CP framework), model interpretability is a further key factor to consider. Model interpretability depends on the types of descriptors employed in model building. Most of the existing models for the prediction of the skin sensitization potential of compounds, including our Skin Doctor CP models, rely on molecular fingerprints [29,40–42]. Interpreting these fingerprints can prove challenging, but in general, some links between chemical patterns and the biological outcomes can be identified [43].

In an attempt to generate predictive models from physically meaningful (and hence more intuitive) descriptors, we previously investigated the capacity of physicochemical property descriptors to produce predictive models for the prediction of the skin sensitization potential [44]. However, the models trained on physicochemical property descriptors do not perform as well as those trained on molecular fingerprints, and their interpretation is still challenging due to the high number of descriptors required to obtain models with an acceptable performance.

Recent studies have shown that in silico models for the prediction of complex in vivo endpoints can benefit from the inclusion of measured or predicted biological data (i.e., in vivo and/or in vitro data) into the feature set. More specifically, descriptive models have been built on small sets of hand-picked biological descriptors relevant to the endpoint of interest [45], as well as on large sets of screening data that may or may not be directly related to the endpoint of interest [46–50]. There are several examples of in silico models, nearest neighbor approaches in particular, that are trained on predicted bioactivities [51,52]. For example, the RASAR models [53] are RF models that predict nine health hazard endpoints (including the skin sensitization potential) based on the distances of a compound of interest to its nearest active and inactive neighbors in reference data sets for 19 toxicological outcomes. Another computational approach utilizes a reasoning framework to build an information-rich network based on assay knowledge, assay data, and predicted bioactivities [54]. The visualization of this network can provide guidance to researchers for the assessment of the safety profile of small molecules.

Recently, Norinder et al. [55] presented a CP framework that utilizes predicted bioactivities as input for in silico models for bioactivity and cytotoxicity prediction. This approach has the advantage of improving a model's predictivity by the use of bioactivity data without the need to perform additional experimental testing for a compound of interest. A similar methodological framework was successfully applied to three in vivo toxicological endpoints (i.e., genotoxicity, drug-induced liver injury, and cardiological complications) by some of us [56].

The aim of this work is to investigate the capacity of predicted bioactivities to produce simple, interpretable machine learning models for the prediction of the skin sensitization potential of small organic compounds without compromising on performance. In order to reach this goal, we explored strategies to replace the molecular fingerprints (MACCS keys) used in Skin Doctor CP by a small set of predicted bioactivities. We selected these predicted bioactivities using Lasso regression from a panel of 372 published CP models for compound toxicity prediction [56] plus three new, additional models for assays of direct relevance to skin sensitization (i.e., DPRA, KeratinoSens assay, and h-CLAT). The final classifiers for the prediction of the skin sensitization potential of compounds were trained on 1021 compounds. They utilize only 10 predicted bioactivity descriptors but perform comparably to the Skin Doctor CP models. The best model ("Skin Doctor CP:Bio") is available free of charge for academic research purposes.

2. Materials and Methods

2.1. Data Sets and Data Processing

2.1.1. Binary LLNA Data

This work is based on the identical LLNA data set that was used for the development of Skin Doctor CP [29]. The random split into a training set (80%) and a test set (20%) was also preserved. The chemical structures were processed with a refined preprocessing protocol that was developed by Garcia de Lomana et al. [56]. This protocol includes the removal of solvents and salts, annotation of aromaticity, neutralization of charges, and mesomerization. Substances containing (i) different components with non-identical SMILES or (ii) fewer than four heavy atoms or (iii) elements other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I were removed from the data set.

The use of the new structure preprocessing protocol led to the rejection of 7 compounds of the training set (and none of the test set) because they do not fulfill the requirements for molecules to be composed of at least one carbon atom and to consist of at least four heavy atoms. The processed training set consists of 1021 compounds and the test set of 257 compounds.

2.1.2. Non-Animal Data on Skin Sensitization

For the calculation of additional bioactivity descriptors, chemical information, and binary assay data for 194 compounds measured in the DPRA, 190 compounds measured in the KeratinoSens assay and 160 compounds measured in the h-CLAT were collected from Alves et al. [57]. The chemical structures were preprocessed following the protocol described above. Preprocessing resulted in the removal of one particular substance (formaldehyde) that is present in all three data sets. The final KeratinoSens assay, h-CLAT, and DPRA data sets comprised 189, 159, and 193 compounds, respectively.

2.1.3. Data for Chemical Space Analysis

In preparation for chemical space comparison, the 7030 cosmetics and 4036 agrochemicals included in the CompTox Chemicals Dashboard [58] and the 2509 approved drugs included in DrugBank [59] were downloaded and processed following the protocol described above. This resulted in a data set of 4488 cosmetics, 2433 agrochemicals, and 2227 approved drugs (the significant reductions are related to the fact that many of the listed cosmetics and agrochemicals are either inorganic salts or without a defined molecular structure).

2.2. Descriptor Calculation and Normalisation

A set of 750 bioactivity descriptors related to 375 predicted binary assay outcomes was calculated for all compounds of the LLNA data set and the three reference data sets (the number of bioactivity descriptors is double that of the predicted binary assay outcomes because the predicted class probabilities of the active and the inactive class were

included in the descriptor set independently from each other). More specifically, class probabilities for 372 bioactivity assays were calculated with aggregated Mondrian CP models that we trained on bioactivity assay data collected from ToxCast [60], eMolTox [61], the eChemPortal [62], and literature, following the identical protocol published by Garcia de Lomana et al. [56]. In addition, predicted class probabilities for three assays relevant to skin sensitization prediction (i.e., DPRA, KeratinoSens assay, h-CLAT) were computed using Mondrian CP models generated by applying the identical model generation framework as described for the other assays [56] to the three corresponding data sets retrieved from Alves et al. Prior to modeling, the standard scaler of the preprocessing module of scikit-learn [63] was used (with default settings) to normalize all bioactivity descriptors. The standard scaler was trained on the LLNA training set only and applied to the full LLNA data set (training and test set). In addition, MACCS keys were calculated with RDKit version 2020.09.1 [64] for all compounds in the LLNA data set.

2.3. Model Development

2.3.1. Aggregated Mondrian Conformal Prediction Modeling

In preparation for model generation, each training set was divided into a proper training set (80%) and a calibration set (20%) by stratified random splitting utilizing the train_test_split function of the Model_selection module of scikit-learn (data shuffling was enabled prior to data set splitting). Then, a RF model was generated (with the RandomForestClassifier function of scikit-learn; all parameters kept default, except for n_estimators = 500 and random_state = 43) and applied to the corresponding calibration and test set.

From the prediction probabilities obtained for the calibration set and the test set, non-conformity scores (α-values) were calculated following Equation (1):

$$\alpha_i = 0.5 - \frac{\hat{P}(y_i|x_i) - max_{y \neq y_i}\hat{P}(y|x_i)}{2} \quad (1)$$

where $\hat{P}(y_i|x_i)$ is the class probability for class i returned by the model, and $max_{y \neq y_i}\hat{P}(y|x_i)$ is the maximum class probability for any other class returned by the model.

The non-conformity scores of the calibration set were sorted class-wise (following the Mondrian conformal prediction protocol), and the relative ranks of the non-conformity scores of each compound of the test set in relation to these lists were retrieved as so-called p-values.

Within the aggregated CP framework, the procedure was repeated for 20 times with different stratified random splits into a proper training and calibration set, altering the random state of the train_test_split function from 0 to 19. For every compound in the test set, a p-value was derived during each run. The median over the p-values obtained during all 20 runs was processed as the final p-value of the compound. The p-values denote the probability of a compound belonging to the corresponding activity class. The model assigns a compound to a specific activity class if the corresponding p-value exceeds the selected error significance level ε.

2.3.2. Measurement of Model Performance

In this work, the classical performance measures (i.e., accuracy (ACC), Matthews correlation coefficient (MCC) [65], correct classification rate (CCR), sensitivity (Sens), specificity (Spec), negative predictive rate (NPV), and positive predictive rate (PPV)) are calculated based exclusively on compounds that were assigned by the CP models to exactly one activity class, i.e., "sensitizer" or "non-sensitizer". This is to enable the application of classic performance measures to CP and, at the same time, to ensure the comparability of the classical performance measures and the results reported for classical non-CP models elsewhere.

In contrast, the CP-specific performance measures (i.e., validity and efficiency) are calculated for all models based on the full sets of compounds to fulfill the common definition of these measures and enable the comparison with other CP models. Validity is defined as the percentage of predictions that include the true class, independently of the prediction of the other class (i.e., it includes "true" predictions as well as "both" predictions). A model is deemed to be valid if the validity is close or equal to the expected value of 1-ε. Efficiency can be understood as an equivalent to the term coverage for non-CP models. It is defined as the percentage of distinct predictions (i.e., predictions that predict exactly one class to be true).

2.3.3. Feature Selection and Parameter Optimization

For feature selection (Figure 1), 10-fold cross-validation (CV) was performed on the training set using the scikit-learn StratifiedKFold function (Model_selection module; n_splits = 10, shuffle = True, random_state = 43).

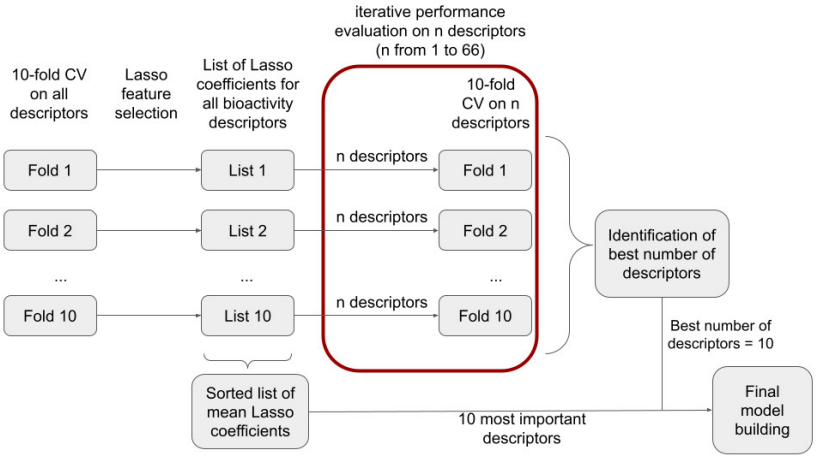


Figure 1. Schematic representation of the workflow for feature selection.

First, the relative importance of each feature within each fold of the CV was investigated. Therefore, hyperparameters for a Lasso classifier were optimized by a 10-fold CV within each fold of the outer CV. This was achieved with the scikit-learn LassoCV function (Linear_model module; random_state = 43, cv = 10, max_iter = 3000, n_alphas = 200). The optimized Lasso classifier was then used to obtain the Lasso coefficients of all bioactivity descriptors within the corresponding fold. The relative importance of each descriptor was calculated as the absolute value of the mean Lasso coefficient calculated over all folds of the CV run.

Second, the optimum number of bioactivity descriptors for model generation was determined. To do so, the 10-fold CV on the training data was repeated, this time without feature selection with Lasso. Instead, a varying number of the most important bioactivity descriptors (i.e., 1 to 66 descriptors; selected based on their coefficients obtained with Lasso) were selected for model building. The mean performance during 10-fold CV in dependence of the number of descriptors was used to select the number of features for the final model.

3. Results and Discussion

3.1. Identification of the Optimum Number of Bioactivity Descriptors for Model Building

In order to identify the most suitable number of bioactivity descriptors *n* for model building, we investigated, within a 10-fold CV framework, the performance of models as

a function of the number of descriptors used (reflecting model interpretability/complexity). Within each CV fold, we performed Lasso regression to rank the descriptors by their corresponding Lasso coefficients (Table S1) and selected the *n* most important descriptors for model building. In Figure 2, we show the improvement of model performance as more bioactivity descriptors are added. In particular, for the first 10 descriptors, a steep increase in MCC and efficiency is observed (see section "Measurement of model performance" of the Methods for important information on how, and in particular on what data, the individual performance measures are calculated). Beyond 10 descriptors, the improvements in model performance are minor and reach a plateau at approximately 25 descriptors. This led us to the conclusion that models based on the 10 most relevant bioactivity descriptors offer the best balance between model performance and complexity (Table 1). Validity is close to the expected value of 1-ε for all the significance levels (i.e., ε = 0.05, 0.10, 0.20, and 0.30) and numbers of descriptors (in this experiment, 1 to 66) investigated.

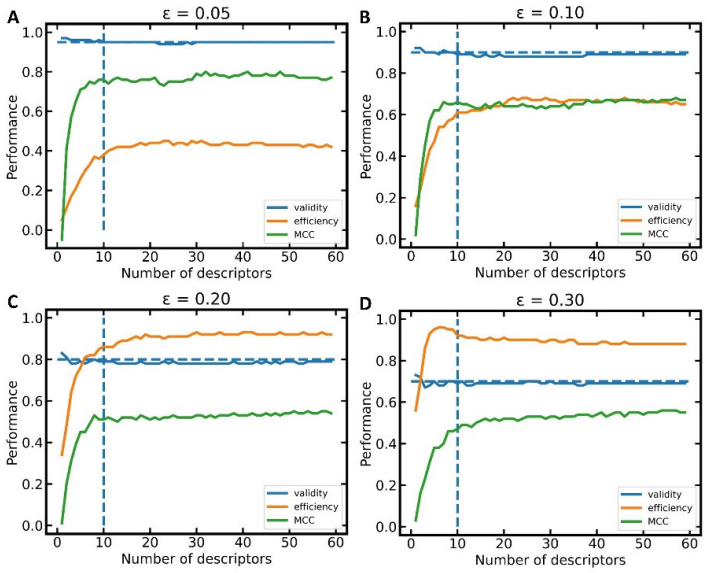


Figure 2. Mean performance of 10-fold CV as a function of the number of bioactivity descriptors selected for model building at the significance level (**A**) $\varepsilon = 0.05$, (**B**) $\varepsilon = 0.10$, (**C**) $\varepsilon = 0.20$ and (**D**) $\varepsilon = 0.30$. The horizontal, dashed line indicates the validity expected from the selected significance level ε; the vertical, dashed line marks the performance of models trained on 10 descriptors.

Table 1. Ten-fold CV Performance of Models Based on 10 Bioactivity Descriptors [1].

Error Significance ε	Validity	Efficiency	ACC	MCC	CCR	Sens	Spec	NPV	PPV
0.05	0.95 (0.03)	0.38 (0.06)	0.88 (0.06)	0.76 (0.12)	0.88 (0.05)	0.87 (0.08)	0.89 (0.08)	0.92 (0.06)	0.85 (0.11)
0.10	0.89 (0.03)	0.61 (0.06)	0.83 (0.05)	0.66 (0.10)	0.83 (0.05)	0.83 (0.10)	0.83 (0.08)	0.88 (0.07)	0.77 (0.09)
0.20	0.79 (0.05)	0.86 (0.04)	0.76 (0.06)	0.51 (0.12)	0.76 (0.06)	0.75 (0.09)	0.77 (0.08)	0.82 (0.07)	0.69 (0.07)
0.30	0.69 (0.07)	0.92 (0.03)	0.74 (0.06)	0.47 (0.11)	0.74 (0.06)	0.72 (0.08)	0.75 (0.07)	0.79 (0.06)	0.67 (0.06)

[1] Standard deviation in parentheses.

3.2. Investigation of the Ten Most Relevant Bioactivity Descriptors

With 10 identified as the optimum number of bioactivity descriptors for model building, we reiterated the above-mentioned descriptor selection process on the full training set and analyzed the relevance and biological meaning of the 10 descriptors with the highest absolute Lasso coefficients averaged over the 10 folds of the CV (Table 2).

The bioactivity descriptor ranked first by the Lasso model is the ToxCast assay "BSK KF3CT ICAM1 down" (Lasso coefficient 0.074). This feature describes the expression of ICAM1 in human keratinocytes. This ToxCast assay is observed to correlate with predictions for other keratinocytes and foreskin assays from the ToxCast BSK family (Kendall τ correlation coefficients between 0.77 and 0.79). The ICAM1 readout is also known as CD54, which is a readout of the skin sensitization-related h-CLAT. The underlying model shows good predictivity (validity = 0.80, efficiency = 0.83, MCC = 0.41 at the significance level of 0.20) The nine further bioactivity descriptors all have similar Lasso coefficients, between 0.036 and 0.051 (validities between 0.74 and 0.87; efficiencies between 0.51 and 0.87; MCCs between 0.30 and 0.98, respectively). Among these are the three assays that we added to the descriptor set because of their direct relevance to skin sensitization: DPRA, KeratinoSens assay, and h-CLAT. As expected, a direct correlation between a positive outcome in any of these three assays and the probability of a compound being a skin sensitizer is identified by the Lasso model. The fact that these assays do not show a high correlation with any other bioactivity descriptors within our full set of descriptors underlines the fact that these descriptors may add important additional information on the skin sensitization potential of compounds. The models predicting these bioactivity descriptors are built on comparably small data sets (<200 compounds). This is reflected by a higher deviation of the significance of these models from the expected value of 0.80 at the investigated significance level of 0.20, compared to the other models. The MCCs of these models are between 0.30 and 0.54.

The ToxCast assay "ATG NRF2 ARE CIS up" describes the activation of NRF2 in human liver cells. Being the fundamental concept of keratinocyte activation analysis via KeratinoSens and LuSens assay, Nrf2 activation is known to play a vital role in the regulation of cellular cytoprotective responses, metabolism, and immune regulation. Included in the top-10 features are also the ToxCast assays "BSK 3C E-selectin down" and "BSK 4H uPAR down", both of which describe inflammation-related biological processes in the endothelium environment. As such, these assays might encode aspects of the immunological response of the human body. "BSK 3C E-selectin down" correlates with other assays associated with inflammation and immune reaction and which are often located in the endothelium. While it shows a positive correlation with the skin sensitization potential (which might indicate an activation of compounds or increased bioavailability), "BSK 4H uPAR down" is one out of only two bioactivity descriptors (among the top-10 features) that show negative correlation with the skin sensitization potential. This assay may therefore report processes involving the deactivation of a compound or the reduction of its bioavailability.

The chromosome aberration assay may not be directly linked to skin sensitization, but it may be relevant to the detection of reactive compounds. The feature is weakly correlated with other assays that are linked to the detection of reactive molecules (e.g., mammalian cell gene mutation assay or AMES mutagenicity assay). Chromosome aberration predictions show no strong correlation with any other descriptors in the set of models.

Table 2. Overview of the Top-10 Bioactivity Descriptors.

Descriptor Name	Assay Title	Mean Lasso Coefficient [1]	σ (Lasso Coefficient)	Correlation to Positive LLNA outcome [2]	5-Fold CV Performance at Significance Level of 0.20			Most Correlating Assays [3]
					Validity	Efficiency	MCC	
p0 BSK KF3CT ICAM1 down	Bioseek human keratinocytes and foreskin fibroblasts intercellular adhesion molecule 1 assay	0.074	0.009	positive	0.80	0.83	0.41	BSK KF3CT SRB down (0.79) BSK KF3CT TGFb1 down (0.78) BSK KF3CT MCP1 down (0.78) BSK KF3CT uPA down (0.78) BSK hDFCGF TIMP1 down (0.77)
p1 BSK 4H uPAR down	Bioseek human umbilical vein endothelium plasminogen activator, urokinase receptor assay	0.051	0.045	negative	0.81	0.82	0.46	BSK 3C uPAR down (0.83) BSK LPS SRB down (0.81) BSK 3C MCP1 down (0.81) BSK 4H SRB down (0.8) BSK SAg MCP1 down (0.8)
p0 Chromosome aberration	Chromosome aberration assay	0.049	0.010	positive	0.79	0.70	0.30	Mammalian cell gene mutation (0.47) AMES (0.41) Inhibitors of Hepatocyte nuclear factor 4 (HNF4) dimerization (0.35) Modulator of Muscarinic acetylcholine receptor M4 (−0.33) Modulator of Bradykinin B2 receptor (−0.33)
p1 DPRA	Direct peptide reactivity assay	0.047	0.013	positive	0.74	0.71	0.30	h-CLAT (0.42) Inhibitors of Hepatocyte nuclear factor 4 (HNF4) dimerization (0.31) KeratinoSens (0.31) Inhibit CYP2C19 Activity (−0.29) Modulator of Peroxisome proliferator-activated receptor gamma (−0.29)
p1 Modulator of Dopamine D1 receptor	Modulator of Dopamine D1 receptor assay	0.045	0.006	positive	0.81	0.81	0.98	Modulator of Alpha-2b adrenergic receptor (0.37) Modulator of Serotonin 1a (5-HT1a) receptor (0.32) Modulator of Alpha-2a adrenergic receptor (0.31) Modulator of Serotonin 2a (5-HT2a) receptor (0.31) Modulators of myocardial damage (0.3)
p1 h-CLAT	Human cell line activation test	0.043	0.013	positive	0.87	0.56	0.54	PGPinhibition (−0.48)

								Caco2 (0.46) LTEA HepaRG DDIT3 up (−0.46) ATG TA CIS up (−0.46) Modulator of P2X purinoceptor 3 (−0.45)
p1 BSK 3C E-selectin down	Bioseek human umbilical vein endothelium selectin E assay	0.043	0.021	positive	0.79	0.77	0.41	BSK 3C VCAM1 down (0.82) BSK 4H Pselectin down (0.81) BSK 4H VCAM1 down (0.81) BSK 3C MCP1 down (0.81) BSK 4H SRB down (0.79)
p1 LTEA HepaRG APOA5 dn	LifeTech/Expression Analysis human HepaRG apolipoprotein A-V assay	0.040	0.012	negative	0.82	0.77	0.51	LTEA HepaRG CYP4A22 dn (0.78) LTEA HepaRG CYP4A11 dn (0.77) LTEA HepaRG FMO3 dn (0.76) LTEA HepaRG HMGCS2 dn (0.76) LTEA HepaRG GSTA2 dn (0.75)
p1 KeratinoSens	ARE-Nrf2 Luciferase test method	0.039	0.004	positive	0.82	0.51	0.31	DPRA (0.31) h-CLAT (0.31) Inhibitors of Hepatocyte nuclear factor 4 (HNF4) dimerization (0.29) Inhibit CYP1A2 Activity (0.27) Modulator of Monoamine oxidase A (0.27)
p0 ATG NRF2 ARE CIS up	Attagene human HepG2 nuclear factor, erythroid 2-like 2 assay	0.036	0.014	positive	0.81	0.87	0.55	ATG PPARg TRANS up (0.67) ATG VDRE CIS up (0.66) ATG MRE CIS up (0.65) ATG PXR TRANS up (0.64) ATG AP 1 CIS up (0.64)

[1]Mean over the 10 folds of the CV. Note that the feature importance rankings of the Lasso model and the RF model may differ. [2]Correlation of the positive assay outcomes and the skin sensitization potentials measured in the LLNA. Since the probability of a compound to belong to the inactive class (p0) or the active class (p1) in a given assay are strongly correlated, either p0 or p1 is selected as an important descriptor by the Lasso model for that assay. Depending on whether p0 or p1 has been selected, and depending on the algebraic sign of the mean Lasso coefficient, a positive predicted assay outcome can either be associated with a positive or a negative LLNA result (i.e., if p0 has a positive correlation with the LLNA result this describes anticorrelation between the positive outcome of both endpoints). [3]Numbers in parentheses report the Kendall τ correlation coefficients between the descriptor and the (most) correlated assay. The full names of the assays are provided in Table S2.

3.3. Coverage of the Chemical Space Relevant to the Development of Cosmetics, Drugs and Agrochemicals

In order to develop an understanding of to what extent the LLNA data set, which we will use to develop the in silico models, represents drugs, cosmetics, and agrochemicals in the feature space defined by the ten selected bioactivity descriptors, a principal component analysis (PCA) was performed on the LLNA data set and the reference sets. As shown in the PCA scatter plot in Figure 3 (PCA loadings plot provided in Figure S1), the LLNA data set covers well the areas in feature space populated by cosmetics, approved drugs, and agrochemicals.

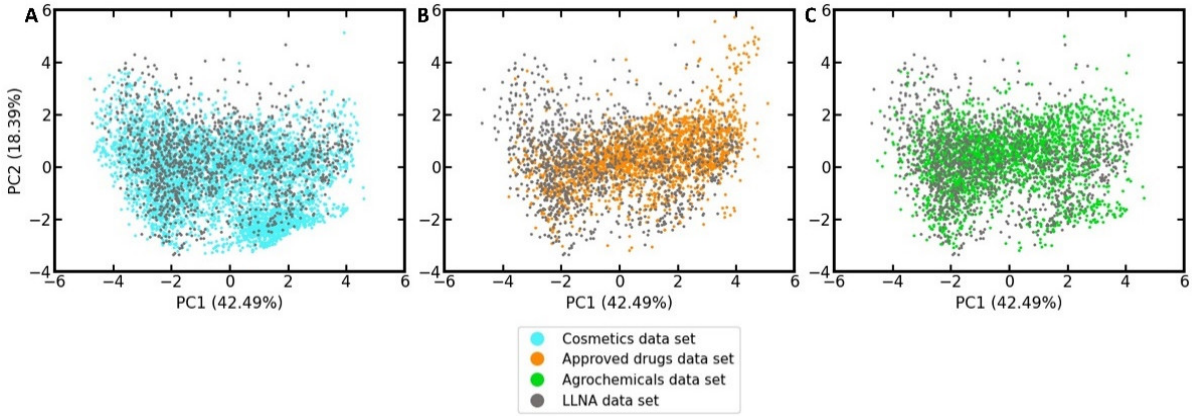


Figure 3. PCA quantifying the coverage of the LLNA data by the reference sets of (**A**) cosmetics, (**B**) approved drugs, and (**C**) agrochemicals in the feature space of the 10 selected bioactivity descriptors. The percentages in parentheses report the variance explained by the respective principal component (PC).

3.4. Analysis of the Distribution of Sensitizers and Non-Sensitizers in the Feature Space of the Ten Selected Bioactivity Descriptors

To investigate the distribution of sensitizers and non-sensitizers within the feature space of the ten selected bioactivity descriptors, another PCA was performed, this time exclusively on the compounds of the LLNA data set (Figure 4). Three characteristic areas can be identified in the scatter plot resulting from this PCA (Figure 4A): Area 1, covering mainly sensitizers; Area 2, covering mainly non-sensitizers; and Area 3, showing intense mixing of sensitizers and non-sensitizers.

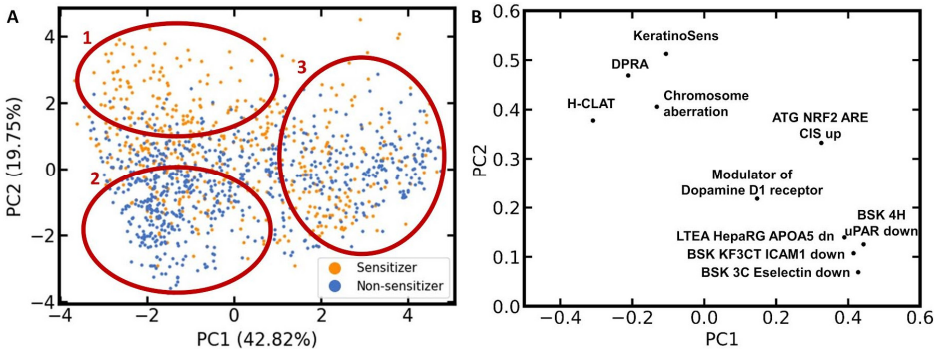


Figure 4. LLNA data set analyzed by PCA in the feature space of the ten selected bioactivity descriptors. (**A**) Scatter plot colored by the binary skin sensitization potential; (**B**) loadings plot of the ten descriptors. The percentages in parentheses report the variance explained by the respective principal component (PC). Note that the axis sections differ for panels (**A**,**B**).

The corresponding loadings plot (Figure 4B) places the bioactivity descriptors for the three skin sensitization assays (h-CLAT, DPRA, and KeratinoSens assay) and the chromosome aberration assay in quadrant 2 (upper left). All four of these assays contribute positively to PC2 and, to a lower degree, negatively to PC1. Since a positive outcome in one or several of the skin sensitization assays should be correlated with a positive skin sensitization potential, this is in agreement with the PCA scatter plot showing a high accumulation of sensitizers in the upper left region. Since a positive outcome in the chromosome aberration assay is likely correlated with a reactive compound, it is also within the expectations that it will shift a compound towards this Area 1 in the PCA scatter plot.

For the remaining six bioactivity descriptors, higher PC1 and PC2 values are expected for compounds that are active in the corresponding assay. Thus, all ten bioactivity descriptors contribute positively to PC2. This means that every compound predicted to be positive in those bioactivity assays is moved towards Area 1 or 3 in the scatter plot. This comes along with the increased probability of a compound to be a skin sensitizer (i.e., to be located in Area 1). At the same time, every negative predicted assay outcome moves the compound towards Area 2, where we mainly expect non-sensitizers to be located, or Area 3, where no prevalence in activity is detected. This positive contribution to PC2 is higher for KeratinoSens, DPRA, chromosome aberration, h-CLAT, and ATG NRF2 than for the other five bioactivity descriptors. In Area 3, we observe intense mixing of skin sensitizers and non-sensitizers, hence posing a significant challenge to classification.

3.5. Model Based on Ten Selected Bioactivity Descriptors

Following the identification of the optimum model setup, a final, aggregated Mondrian CP model based on the ten selected bioactivity descriptors was derived from the full training set and evaluated on the holdout data set. From here on, we refer to this model as the SkinDoctor CP:Bio model.

3.5.1. Performance on the Test Set

Within the standard deviation expected from CV, the SkinDoctor CP:Bio model was valid at all four significance levels investigated (Table 3). The efficiencies of the model ranged from 0.39 to 0.95 and the MCCs ranged from 0.72 to 0.49, depending on the significance level.

Class-wise performance analysis (Table 4) showed that the SkinDoctor CP:Bio model was valid for sensitizers and non-sensitizers at all significance levels investigated. The largest difference in validity between the two classes (0.08) was observed at the significance level of 0.30. Efficiency was in general similar for both classes (largest difference 0.04).

Table 3. Performance of the model based on ten selected bioactivity descriptors on the test set.

Error Significance ε	Validity	Efficiency	ACC	MCC	CCR	Sens	Spec	NPV	PPV
0.05	0.95	0.39	0.86	0.72	0.86	0.84	0.88	0.88	0.84
0.10	0.89	0.56	0.81	0.62	0.81	0.85	0.77	0.88	0.74
0.20	0.81	0.82	0.76	0.53	0.77	0.80	0.74	0.83	0.69
0.30	0.70	0.95	0.74	0.49	0.75	0.78	0.72	0.82	0.67

Table 4. Class-wise performance of the model based on ten selected bioactivity descriptors on the test set.

Error Significance ε	Class	Validity	Efficiency
0.05	Non-sensitizer	0.95	0.38
	Sensitizer	0.93	0.41
0.10	Non-sensitizer	0.87	0.55
	Sensitizer	0.92	0.58
0.20	Non-sensitizer	0.79	0.82
	Sensitizer	0.83	0.82
0.30	Non-sensitizer	0.67	0.94
	Sensitizer	0.75	0.95

3.5.2. Comparison of the New Model with the Skin Doctor CP Model

The previously developed Skin Doctor CP model [29] is trained on MACCS keys (166 features), whereas the Skin Doctor CP:Bio model is trained on ten selected bioactivity descriptors. All other differences in the data and protocols used for model building and testing are minor (Table S3), thus enabling a direct, comparative assessment of the two feature types and their impact on model performance and behavior.

On the holdout data set of 257 compounds measured in the LLNA (none of these compounds is part of the training set of either model), both the Skin Doctor CP model and the Skin Doctor CP:Bio model were valid at all significance levels investigated. For the sake of clarity, we focus our discussion here on the commonly applied significance level of 0.20; performance data on all significance levels are provided in Table S4. At the significance level of 0.20, the Skin Doctor CP and Skin Doctor CP:Bio models yielded validities of 0.82 and 0.81, respectively. The efficiencies (0.78 vs. 0.82) and MCCs (0.55 vs. 0.53) obtained for the Skin Doctor CP and Skin Doctor CP:Bio models were also comparable. The differences in performance between the two models are slightly above the standard deviation observed for the 10-fold CV experiments but small enough to consider the performance of the two models similar.

3.6. Combination of Bioactivity Descriptors with MACCS Keys in an Attempt to Improve Model Performance

MACCS keys encode structural patterns of molecules and thus information that is very different from that encoded by the bioactivity descriptors. The use of MACCS keys in combination with the ten selected bioactivity descriptors could hence yield better models. However, a RF model derived from the combined set of MACCS keys and the ten selected bioactivity descriptors (n_estimators = 500; all other parameters default) did not yield better performance on the test set.

Therefore, we generated a model trained exclusively on MACCS keys plus a model trained exclusively on the ten selected bioactivity descriptors (both models with n_estimators = 500; all other parameters default), and, based on a simple set of rules (see Figure 5), combined both models to form a consensus model. This set of rules follows the idea that only unambiguous predictions by the single models (i.e., predictions assigning a compound to exactly one class) are considered. If one model returns an unambiguous prediction or if both models return an unambiguous prediction and are in agreement, the unambiguous prediction is reported as the final result. In all other cases, the consensus model does not return a prediction.

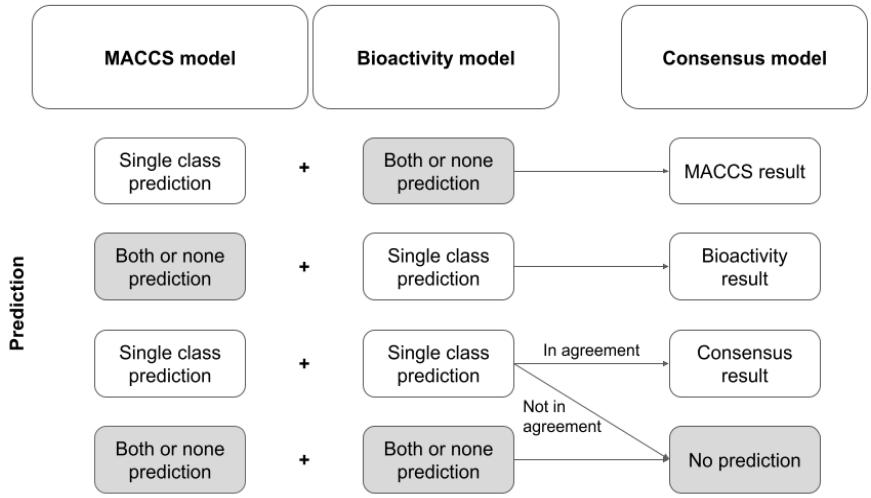


Figure 5. Architecture of the consensus model.

Table 5 reports on the performance of this consensus model at different error significance levels. Note that because the consensus model does not fulfill the definitions of a pure CP model, validity and efficiency cannot be calculated for this model.

When running the two CP models underlying the consensus approach at a significance level of 0.20, the consensus approach reached a coverage of 0.89 and an MMC of 0.54. Hence, compared to the Skin Doctor CP:Bio model (efficiency 0.82 and MCC 0.53 at a significance level of 0.20), the consensus model obtained only slightly better coverage while maintaining the MCC.

A second, combined, model was constructed by averaging the *p*-values returned for each class by the model based on MACCS keys and the model based on bioactivity descriptors. The model was valid to over-predictive at the four significance levels investigated. At the significance level of 0.20, the validity was 0.82. The efficiency at this significance level was 0.79 (vs. 0.82 for the Skin Doctor CP:Bio model) and the MCC was 0.56 (vs. 0.53 for the Skin Doctor CP:Bio model). Hence, compared to the Skin Doctor CP:Bio model, this combined model obtains a slightly higher MCC, at the cost of efficiency.

Table 5. Performance of the consensus and the combined models on the test set.

Consensus Model Based on a Set of Rules								
Error significance ε [1]	**Coverage**	**ACC**	**MCC**	**CCR**	**Sens**	**Spec**	**NPV**	**PPV**
0.05	0.51	0.86	0.72	0.86	0.84	0.88	0.88	0.84
0.10	0.71	0.79	0.59	0.80	0.83	0.77	0.88	0.70
0.20	0.89	0.77	0.54	0.78	0.82	0.73	0.85	0.68
0.30	0.83	0.78	0.56	0.79	0.85	0.72	0.88	0.68
Combined Model Based on Mean *p*-Values								
Error significance ε	**Validity**	**Efficiency**	**ACC**	**MCC**	**CCR**	**Sens**	**Spec**	**NPV**	**PPV**
0.05	0.97	0.24	0.89	0.77	0.89	0.92	0.86	0.94	0.82
0.10	0.93	0.46	0.86	0.72	0.87	0.94	0.80	0.95	0.76
0.20	0.82	0.79	0.77	0.56	0.78	0.85	0.72	0.87	0.69
0.30	0.71	0.95	0.75	0.50	0.76	0.80	0.72	0.84	0.66

[1] Error significance of the underlying model, not of the combined model itself.

In order to obtain a better understanding of the advantages and disadvantages of the two combined models over the single models, we investigated the relationship between classification performance (MCC) and coverage. From Figure 6, it can be seen that the combined models tend to obtain better MCC values at a given coverage than the single models. At higher coverages, the combined model based on averaged *p*-values has slightly better MCCs than the combined model based on the set of rules. A further advantage of the combined model based on *p*-value averaging is that users can select a confidence level; this is not possible with the combined model based on the set of rules.

Overall, the *p*-value averaging approach seems to be preferable over the rule-based approach. Compared to the single model (i.e., the Skin Doctor CP:Bio model), the advantages of the combined approach with respect to performance are outweighed by the fact that the single model has much lower complexity and, hence, better interpretability.

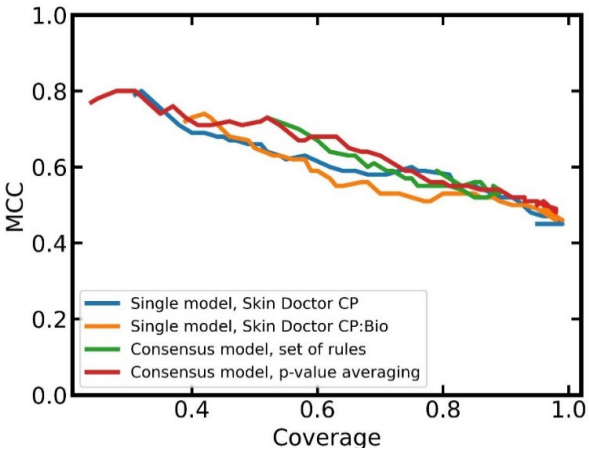


Figure 6. Relationship between MCC and coverage for the individual and the combined models.

3.7. Investigation of the Influence of Experimental Skin Sensitization Assay Results on Predictivity

Feature selection with Lasso and the RF algorithm identified the three bioactivity descriptors derived from the three skin sensitization-specific assays (i.e., DPRA, KeratinoSens assay, h-CLAT) as important for modeling the LLNA. In order to obtain a better understanding of the role and significance of these three bioactivity descriptors, we investigated them from different perspectives.

First, we determined the (5-fold) CV performance of the CP models for the DPRA, KeratinoSens assay, and h-CLAT descriptors on the (i) 194 compounds measured in the DPRA, (ii) 190 compounds measured in the KeratinoSens assay, and (iii) 160 compounds measured in the h-CLAT. The KeratinoSens and h-CLAT models (Table 6) were valid at a significance level of 0.2 (validities of 0.82 and 0.87, respectively) while the DPRA model showed a slight underperformance (validity 0.74). The efficiencies of the models were fairly low (0.51 to 0.71) in comparison to most of the other CP models for bioactivity prediction. We assume that the low efficiency is related to the fact that the training sets for these CP models are small (<200 compounds). The other evaluated performance measures are within expectations (e.g., MCC between 0.30 and 0.54). Overall, we conclude from these results that the predicted assay outcomes from these three models could make a substantial contribution to models predicting the skin sensitization potential.

Second, we investigated (by 10-fold CV on the full LLNA data set) whether the high importance attributed by Lasso to the skin sensitization-specific assays could be a result of overlaps in the training or test data of the LLNA model (SkinDoctor CP:Bio model) and the training data of the DPRA/KeratinoSens assay/h-CLAT models. For the overlapping

compounds, the *p*-values used as bioactivity descriptors should be accurate (since the experimental value of the in vitro assays is known) and therefore more informative. In order to investigate this, we determined the performance of the SkinDoctor CP:Bio model in dependence of the number of compounds overlapping between the LLNA data set (i.e., the test data within each fold) and the training data of the DPRA/KeratinoSens assay/h-CLAT models. We found that six compounds of the LLNA data set were present also in exactly one of the DPRA/KeratinoSens assay/h-CLAT training sets, 45 compounds were present in exactly two of these assays, and 132 compounds in each of these three assays. Note that the number of compounds present in the LLNA data set and in exactly one of the three non-animal assay data sets is too low to make any meaningful observations, for which reason this case was not further pursued. For the remaining two subsets of compounds, the performances of the models were comparable to each other as well as to the subset containing the compounds that are not present in any of three assay data sets (Table 7). For this reason, we are confident that the importance attributed to the predicted DPRA, KeratinoSens assay, and h-CLAT outcomes is genuine and not a result of a bias in the data.

Table 6. Five-fold CV performance of the CP models for DPRA, KeratinoSens assay, and h-CLAT at the significance level of 0.20 [1].

Assay to be Predicted	No. Compounds in Data Set	Validity	Efficiency	ACC	ACC (Sensitizers)	ACC (Non-Sensitizers)	F1 Score	MCC
DPRA	194	0.74 (0.09)	0.71 (0.14)	0.64 (0.07)	0.60 (0.06)	0.69 (0.20)	0.64 (0.07)	0.30 (0.18)
KeratinoSens	190	0.82 (0.11)	0.51 (0.08)	0.67 (0.19)	0.66 (0.24)	0.68 (0.23)	0.64 (0.19)	0.31 (0.35)
h-CLAT	160	0.87 (0.03)	0.56 (0.56)	0.78 (0.05)	0.76 (0.15)	0.75 (0.29)	0.74 (0.06)	0.54 (0.08)

[1] Standard deviation in parentheses.

Table 7. Performance of the SkinDoctor CP:Bio model during 10-fold CV on the full LLNA data set in dependence of the number of skin sensitization assays for which experimental data are available.

Error Significance ε	Validity	Efficiency	MCC	Validity	Efficiency	MCC	Validity	Efficiency	MCC
	For Compounds Exclusive to the LLNA Data Set			For Compounds Present in the LLNA Data Set Plus Exactly Two of the DPRA/KeratinoSens Assay/h-CLAT Training Sets			For Compounds Present in the LLNA Data Set Plus Exactly Three of the DPRA/KeratinoSens Assay/h-CLAT Training Sets		
0.05	0.97	0.33	0.77	0.98	0.44	0.79	0.98	0.45	0.77
0.10	0.91	0.56	0.64	0.96	0.62	0.74	0.93	0.65	0.67
0.20	0.81	0.83	0.52	0.91	0.84	0.66	0.86	0.81	0.51
0.30	0.69	0.94	0.45	0.82	0.93	0.68	0.73	0.93	0.42

Third, we tested the capacity of a model trained only on DPRA, KeratinoSens assay, and h-CLAT assay data to predict the outcomes of the LLNA. This experiment is particularly interesting because a number of existing in silico models for the prediction of the skin sensitization potential are trained exclusively on data from these three assays [66–68].

In five-fold CV, our CP model trained exclusively on DPRA, KeratinoSens assay, and h-CLAT assay data descriptors (n_estimators = 500; all other parameters default) was valid at all error significance levels investigated (Table 8), but its efficiency (0.21 at $\varepsilon = 0.05$; 0.88 at $\varepsilon = 0.30$) and MCC (0.48 at $\varepsilon = 0.05$; 0.37 at $\varepsilon = 0.30$) were substantially lower than those of the CP model derived from the ten selected bioactivity descriptors. These results indicate that the bioactivity descriptors derived from other assays add relevant, additional information to the models that is needed to obtain good classifiers.

Table 8. Test set performance of the classifier trained exclusively on predicted values of the DPRA, KeratinoSens and h-CLAT assays.

Error Significance ε	Validity	Efficiency	ACC	MCC	CCR	Sens	Spec	NPV	PPV
0.05	0.94	0.21	0.71	0.48	0.72	0.92	0.52	0.88	0.63
0.10	0.90	0.40	0.75	0.51	0.76	0.82	0.69	0.84	0.67
0.20	0.80	0.70	0.72	0.44	0.72	0.76	0.69	0.81	0.61
0.30	0.72	0.88	0.68	0.37	0.69	0.71	0.66	0.78	0.59

3.8. Impact of the Limitation of the Available Experimental Data on Model Performance

Most of the freely available models for the prediction of the skin sensitization potential of small molecules are trained on LLNA data, and the evaluation reports for many of these models indicate that their performance is comparable [29,40,44,57]. It is plausible that the observed plateauing of model performance is related to the limited quantity and quality of the data available for model development. In order to investigate whether our classifiers could benefit from additional LLNA data, we investigated the relationship between model performance and the size of the training data.

As expected, and shown in Figure 7, the performance of models increases with the number of training instances, regardless of the type of descriptors used. The MCCs of the models based on bioactivity descriptors improve from an average of 0.41 to an average of 0.50, respectively. Consistent with our initial CV experiments, the use of more than ten bioactivity descriptors yields minor improvements in model performance that we believe are outweighed by higher model complexity.

The MCC of the model based on MACCS keys improves from 0.28 (when trained on 115 compounds) to 0.47 (when trained on 1150 compounds), indicating that the models trained on MACCS keys require substantially more training data than the models trained on bioactivity descriptors to obtain good performance. In this particular case, the MACCS keys model reaches a comparable performance to the model based on bioactivity descriptors only when all the available LLNA data are used for modeling. This leaves the MACCS keys model clearly more data-hungry than the models based on predicted bioactivities, with the benefit of showing the potential to surpass the model based on predicted bioactivities given the availability of sufficient amounts of data.

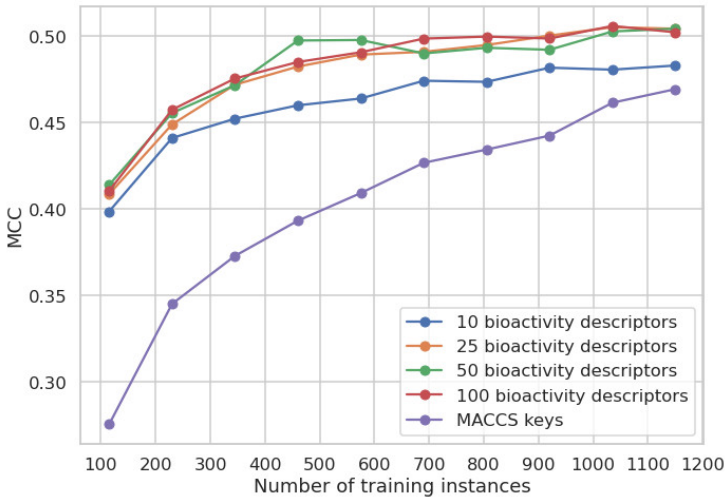


Figure 7. Performance of the RF classifier (n_estimators = 500; all other parameters default) underlying the CP model as a function of the number of instances the model was trained on.

4. Conclusions

In this work, we report on the development and validation of a new machine learning model for the prediction of the skin sensitization potential of small organic molecules: Skin Doctor CP:Bio. Whereas the previously reported models are mostly based on molecular fingerprints (which in general are difficult to interpret), Skin Doctor CP:Bio utilizes just ten bioactivity descriptors to reach competitive performance. Most of these bioactivity descriptors are known to be directly or indirectly linked to skin sensitization, which adds to the interpretability of the model and supports its meaningfulness.

At the significance level of 0.20, Skin Doctor CP:Bio obtained an efficiency of 0.82 and an MCC of 0.53 on the holdout data set of 257 compounds. These results demonstrate the good performance of the model and, hence, the relevance of the selected bioactivity descriptors. Analysis of the LLNA training data projected into the new feature space proves that cosmetics, drugs, and agrochemicals are well embedded in the data, hence corroborating the relevance of the model to different industries.

In an attempt to further improve model performance and coverage, we explored different strategies to exploit the information contained in molecular fingerprints (MACCS keys) and biological descriptors. The models obtained from these experiments showed minor improvements in performance that are outweighed by the costs of higher model complexity and limited interpretability.

An important observation to make was that models based on MACCS keys are clearly more data-hungry than models based on predicted bioactivities. Only when using all of the available LLNA data, the model based on MACCs keys was able to catch up with the model based on predicted bioactivities. This highlights the relevance of the presented approach to the development of strategies to address the many questions in biology, pharmacology, and toxicology where measured data are scarce. We believe that the modeling strategies presented in this work could be easily adopted to address many of these research questions. The Skin Doctor CP:Bio model is available free of charge for academic research purposes.

Supplementary Materials: The following are available online at www.mdpi.com/article/10.3390/ph14080790/s1. Figure S1. Loadings plot for the PCA on the LLNA and the three reference data sets, based on the ten selected bioactivity descriptors; Table S1: Mean absolute lasso coefficients and standard deviation σ retrieved from the 10-fold CV; Table S2: Full name of the assays with high correlation to the ten selected bioactivity descriptors; Table S3: Comparison of the Skin Doctor CP and Skin Doctor CP:Bio approaches; Table S4: Results of Skin Doctor CP on the test set.

Author Contributions: Conceptualization, A.W., U.N., J.K. (Jochen Kühnl), and J.K. (Johannes Kirchmair); methodology, A.W., M.G.d.L., N.M., J.K. (Jochen Kühnl), and J.K. (Johannes Kirchmair); software, A.W., M.G.d.L., N.M., C.S., and S.H.; validation, A.W., J.K. (Jochen Kühnl), and J.K. (Johannes Kirchmair); resources, J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); data curation, A.W., N.M., and M.G.d.L.; writing—original draft preparation, A.W., M.G.d.L., C.S., N.M., S.H., U.N., J.K. (Jochen Kühnl), and J.K. (Johannes Kirchmair); visualization, A.W. and S.H.; supervision, J.K. (Jochen Kühnl), J.K. (Johannes Kirchmair), and U.N.; project administration, J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair); funding acquisition, J.K. (Jochen Kühnl) and J.K. (Johannes Kirchmair). All authors have read and agreed to the published version of the manuscript.

Funding: N.M. and J.K. (Johannes Kirchmair) are supported by the Trond Mohn Foundation (BFS2017TMT01). C.S. and J.K. (Johannes Kirchmair) are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)-project number KI 2085/1-1. A.W. is supported by Beiersdorf AG through HITeC e.V.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The models for generating the ten bioactivity descriptors as well as the final LLNA model are available free of charge for academic research from https://doi.org/10.5281/zenodo.5101594, accessed on 7 August 2021 [69]. The LLNA data set used for model development and evaluation was published earlier [29].

Conflicts of Interest: A.W. is funded by Beiersdorf AG through HITeC e.V and J.K. (Jochen Kühnl) is employed at Beiersdorf AG.

Abbreviations

ACC	accuracy
CCR	correct classification rate
CP	conformal prediction
CV	cross validation
DPRA	direct peptide reactivity assay
GPMT	guinea pig maximization test
h-CLAT	human cell line activation test
IATA	integrated approach for testing and assessment
LLNA	local lymph node assay
MCC	Matthews correlation coefficient
NPV	negative predictive rate
PC	principal component
PCA	principal component analysis
PPV	positive predictive rate
(Q)SAR	(quantitative) structure activity relationship
RF	random forest
Sens	sensitivity
Spec	specificity

References

1. Kimber, I.; Basketter, D.A.; Gerberick, G.F.; Ryan, C.A.; Dearman, R.J. Chemical Allergy: Translating Biology into Hazard Characterization. *Toxicol. Sci.* **2011**, *120* (Suppl. 1), S238–S268, doi:10.1093/toxsci/kfq346.
2. Olusegun, O.A.; Martincigh, B.S. Allergic Contact Dermatitis: A Significant Environmental and Occupational Skin Disease. *Int. J. Dermatol.* **2021**, doi:10.1111/ijd.15502.
3. Lushniak, B.D. Occupational Contact Dermatitis. *Dermatol. Ther.* **2004**, *17*, 272–277, doi:10.1111/j.1396-0296.2004.04032.x.
4. Thyssen, J.P.; Linneberg, A.; Menné, T.; Johansen, J.D. The Epidemiology of Contact Allergy in the General Population—Prevalence and Main Findings. *Contact Dermat.* **2007**, *57*, 287–299, doi:10.1111/j.1600-0536.2007.01220.x.
5. van Amerongen, C.C.A.; Ofenloch, R.F.; Cazzaniga, S.; Elsner, P.; Gonçalo, M.; Naldi, L.; Svensson, Å.; Bruze, M.; Schuttelaar, M.L.A. Skin Exposure to Scented Products Used in Daily Life and Fragrance Contact Allergy in the European General Population—The EDEN Fragrance Study. *Contact Dermat.* **2021**, *84*, 385–394, doi:10.1111/cod.13807.
6. Aalto-Korte, K.; Suuronen, K. Ten Years of Contact Allergy from Acrylic Compounds in an Occupational Dermatology Clinic. *Contact Dermat.* **2021**, *84*, 240–246, doi:10.1111/cod.13739.
7. Nedorost, S.; Hammond, M. Art of Prevention: Allergic Sensitization through Damaged Skin: Atopic, Occupational, and Stasis Dermatitis. *Int. J. Women's Dermatol.* **2020**, *6*, 381–383, doi:10.1016/j.ijwd.2020.08.004.
8. Salah, S.; Taieb, C.; Demessant, A.L.; Haftek, M. Prevalence of Skin Reactions and Self-Reported Allergies in 5 Countries with Their Social Impact Measured through Quality of Life Impairment. *Int. J. Environ. Res. Public Health* **2021**, *18*, 4501, doi:10.3390/ijerph18094501.
9. Felter, S.; Kern, P.; Ryan, C. Allergic Contact Dermatitis: Adequacy of the Default 10X Assessment Factor for Human Variability to Protect Infants and Children. *Regul. Toxicol. Pharmacol.* **2018**, *99*, 116–121, doi:10.1016/j.yrtph.2018.09.011.
10. OECD. *OECD Guidelines for the Testing of Chemicals, Section 4 Test No. 429: Skin Sensitisation Local Lymph Node Assay: Local Lymph Node Assay*; OECD Publishing: Paris, France, 2010; ISBN 9789264071100
11. Anderson, S.E.; Siegel, P.D.; Meade, B.J. The LLNA: A Brief Review of Recent Advances and Limitations. *J. Allergy* **2011**, *2011*, 424203–424213, doi:10.1155/2011/424203.
12. Gerberick, G.F.; House, R.V.; Fletcher, E.R.; Ryan, C.A. Examination of the Local Lymph Node Assay for Use in Contact Sensitization Risk Assessment. *Fundam. Appl. Toxicol.* **1992**, *19*, 438–445, doi:10.1016/0272-0590(92)90183-i.
13. Santín, E.P.; Solana, R.R.; García, M.G.; Del Mar García Suárez, M.; Díaz, G.D.B.; Cabal, M.D.C.; Rojas, J.M.M.; Sánchez, J.I.L. Toxicity Prediction Based on Artificial Intelligence: A Multidisciplinary Overview. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1516, doi:10.1002/wcms.1516.

14. Pistollato, F.; Madia, F.; Corvi, R.; Munn, S.; Grignard, E.; Paini, A.; Worth, A.; Bal-Price, A.; Prieto, P.; Casati, S.; et al. Current EU Regulatory Requirements for the Assessment of Chemicals and Cosmetic Products: Challenges and Opportunities for Introducing New Approach Methodologies. *Arch. Toxicol.* **2021**, *95*, 1867–1897, doi:10.1007/s00204-021-03034-y.
15. Ta, G.H.; Weng, C.F.; Leong, M. In Silico Prediction of Skin Sensitization: Quo Vadis? *Front. Pharmacol.* **2021**, *12*, 1052, doi:10.22541/au.159652623.38717529.
16. Mehling, A.; Eriksson, T.; Eltze, T.; Kolle, S.; Ramirez, T.; Teubner, W.; van Ravenzwaay, B.; Landsiedel, R. Non-Animal Test Methods for Predicting Skin Sensitization Potentials. *Arch. Toxicol.* **2012**, *86*, 1273–1295, doi:10.1007/s00204-012-0867-6.
17. Reisinger, K.; Hoffmann, S.; Alépée, N.; Ashikaga, T.; Barroso, J.; Elcombe, C.; Gellatly, N.; Galbiati, V.; Gibbs, S.; Groux, H.; et al. Systematic Evaluation of Non-Animal Test Methods for Skin Sensitisation Safety Assessment. *Toxicol.* In Vitro **2015**, *29*, 259–270, doi:10.1016/j.tiv.2014.10.018.
18. Ezendam, J.; Braakhuis, H.M.; Vandebriel, R.J. State of the Art in Non-Animal Approaches for Skin Sensitization Testing: From Individual Test Methods towards Testing Strategies. *Arch. Toxicol.* **2016**, *90*, 2861–2883, doi:10.1007/s00204-016-1842-4.
19. Thyssen, J.P.; Giménez-Arnau, E.; Lepoittevin, J.-P.; Menné, T.; Boman, A.; Schnuch, A. The Critical Review of Methodologies and Approaches to Assess the Inherent Skin Sensitization Potential (skin Allergies) of Chemicals. Part I. *Contact Dermat.* **2012**, *66* (Suppl. 1), 11–24., doi:10.1111/j.1600-0536.2011.02004_2.x.
20. Wilm, A.; Kühnl, J.; Kirchmair, J. Computational Approaches for Skin Sensitization Prediction. *Crit. Rev. Toxicol.* **2018**, *48*, 738–760, doi:10.1080/10408444.2018.1528207.
21. ECHA (European Chemicals Agency). The Use of Alternatives to Testing on Animals for the REACH Regulation, Third Report under Article 117(3) of the REACH Regulation. Available online: https://echa.europa.eu/documents/10162/13639/alternatives_test_animals_2017_en.pdf (accessed on 10 July 2019).
22. OECD. Test No. 442C: In Chemico Skin Sensitisation. Available online: http://www.oecd.org/env/test-no-442c-in-chemico-skin-sensitisation-9789264229709-en.htm (accessed on 10 July 2019).
23. OECD. Test No. 442D: In Vitro Skin Sensitisation. Available online: http://www.oecd.org/env/test-no-442d-in-vitro-skin-sensitisation-9789264229822-en.htm (accessed on 17 April 2018).
24. OECD. Test No. 442E: In Vitro Skin Sensitisation. Available online: http://www.oecd.org/env/test-no-442e-in-vitro-skin-sensitisation-9789264264359-en.htm (accessed on 17 April 2018).
25. OECD. OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. Available online: https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (accessed on July 2019).
26. Jowsey, I.R.; Basketter, D.A.; Westmoreland, C.; Kimber, I. A Future Approach to Measuring Relative Skin Sensitising Potency: A Proposal. *J. Appl. Toxicol.* **2006**, *26*, 341–350, doi:10.1002/jat.1146.
27. Hoffmann, S.; Kleinstreuer, N.; Alépée, N.; Allen, D.; Api, A.M.; Ashikaga, T.; Clouet, E.; Cluzel, M.; Desprez, B.; Gellatly, N.; et al. Non-Animal Methods to Predict Skin Sensitization (I): The Cosmetics Europe Database. *Crit. Rev. Toxicol.* **2018**, *48*, 344–358, doi:10.1080/10408444.2018.1429385.
28. Safford, R.J.; Api, A.M.; Roberts, D.W.; Lalko, J.F. Extension of the Dermal Sensitisation Threshold (DST) Approach to Incorporate Chemicals Classified as Reactive. *Regul. Toxicol. Pharmacol.* **2015**, *72*, 694–701, doi:10.1016/j.yrtph.2015.04.020.
29. Wilm, A.; Norinder, U.; Agea, M.I.; de Bruyn Kops, C.; Stork, C.; Kühnl, J.; Kirchmair, J. Skin Doctor CP: Conformal Prediction of the Skin Sensitization Potential of Small Organic Molecules. *Chem. Res. Toxicol.* **2020**, *34*, 330–344, doi:10.1021/acs.chemrestox.0c00253.
30. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer Science & Business Media: New York, NY, USA, 2005; ISBN 9780387001524
31. Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling for Regulatory Purposes. A Transparent and Flexible Alternative to Applicability Domain Determination. *Regul. Toxicol. Pharmacol.* **2015**, *71*, 279–284, doi:10.1016/j.yrtph.2014.12.021.
32. Norinder, U.; Rybacka, A.; Andersson, P.L. Conformal Prediction to Define Applicability Domain—A Case Study on Predicting ER and AR Binding. *SAR QSAR Environ. Res.* **2016**, *27*, 303–316, doi:10.1080/1062936x.2016.1172665.
33. Vovk, V. Conditional Validity of Inductive Conformal Predictors. *Mach. Learn.* **2013**, *92*, 349–376, doi:10.1007/s10994-013-5355-6.
34. Concepts and Applications of Conformal Prediction in Computational Drug Discovery. Available online: https://arxiv.org/pdf/1908.03569.pdf (accessed on 17 March 2020).
35. Carlsson, L.; Eklund, M.; Norinder, U. Aggregated Conformal Prediction. In Proceedings of IFIP International Conference on Artificial Intelligence Applications and Innovations, Rhodes, Greece, 19–21 September *2014*; pp. 231–240.
36. Svensson, F.; Afzal, A.M.; Norinder, U.; Bender, A. Maximizing Gain in High-Throughput Screening Using Conformal Prediction. *J. Cheminform.* **2018**, *10*, 7, doi:10.1186/s13321-018-0260-4.
37. Norinder, U.; Svensson, F. Multitask Modeling with Confidence Using Matrix Factorization and Conformal Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 1598–1604, doi:10.1021/acs.jcim.9b00027.
38. Norinder, U.; Ahlberg, E.; Carlsson, L. Predicting Ames Mutagenicity Using Conformal Prediction in the Ames/QSAR International Challenge Project. *Mutagenesis* **2019**, *34*, 33–40, doi:10.1093/mutage/gey038.
39. Zhang, J.; Norinder, U.; Svensson, F. Deep Learning-Based Conformal Prediction of Toxicity. *J. Chem. Inf. Model.* **2021**, doi:10.1021/acs.jcim.1c00208.

40. Di, P.; Yin, Y.; Jiang, C.; Cai, Y.; Li, W.; Tang, Y.; Liu, G. Prediction of the Skin Sensitising Potential and Potency of Compounds via Mechanism-Based Binary and Ternary Classification Models. *Toxicol. In Vitro* **2019**, *59*, 204–214, doi:10.1016/j.tiv.2019.01.004.
41. Borba, J.V.B.; Braga, R.C.; Alves, V.M.; Muratov, E.N.; Kleinstreuer, N.; Tropsha, A.; Andrade, C.H. Pred-Skin: A Web Portal for Accurate Prediction of Human Skin Sensitizers. *Chem. Res. Toxicol.* **2021**, *34*, 258–267, doi:10.1021/acs.chemrestox.0c00186.
42. Liu, J.; Kern, P.S.; Gerberick, G.F.; Santos-Filho, O.A.; Esposito, E.X.; Hopfinger, A.J.; Tseng, Y.J. Categorical QSAR Models for Skin Sensitization Based on Local Lymph Node Assay Measures and Both Ground and Excited State 4D-Fingerprint Descriptors. *J. Comput. Aided Mol. Des.* **2008**, *22*, 345–366, doi:10.1007/s10822-008-9190-y.
43. Riniker, S.; Landrum, G.A. Similarity Maps—A Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminform.* **2013**, *5*, 43, doi:10.1186/1758-2946-5-43.
44. Wilm, A.; Stork, C.; Bauer, C.; Schepky, A.; Kühnl, J.; Kirchmair, J. Skin Doctor: Machine Learning Models for Skin Sensitization Prediction That Provide Estimates and Indicators of Prediction Reliability. *Int. J. Mol. Sci.* **2019**, *20*, 4833, doi:10.3390/ijms20194833.
45. Kleinstreuer, N.C.; Hoffmann, S.; Alépée, N.; Allen, D.; Ashikaga, T.; Casey, W.; Clouet, E.; Cluzel, M.; Desprez, B.; Gellatly, N.; et al. Non-Animal Methods to Predict Skin Sensitization (II): An Assessment of Defined Approaches. *Crit. Rev. Toxicol.* **2018**, *48*, 359–374, doi:10.1080/10408444.2018.1429386.
46. Zhang, J.; Hsieh, J.-H.; Zhu, H. Profiling Animal Toxicants by Automatically Mining Public Bioassay Data: A Big Data Approach for Computational Toxicology. *PLoS ONE* **2014**, *9*, e99863, doi:10.1371/journal.pone.0099863.
47. Ribay, K.; Kim, M.T.; Wang, W.; Pinolini, D.; Zhu, H. Predictive Modeling of Estrogen Receptor Binding Agents Using Advanced Cheminformatics Tools and Massive Public Data. *Front. Environ. Sci.* **2016**, *4*, 12, doi:10.3389/fenvs.2016.00012.
48. Zhu, H.; Zhang, J.; Kim, M.T.; Boison, A.; Sedykh, A.; Moran, K. Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays to Identify Potential Toxicants. *Chem. Res. Toxicol.* **2014**, *27*, 1643–1651, doi:10.1021/tx500145h.
49. Kim, M.T.; Huang, R.; Sedykh, A.; Wang, W.; Xia, M.; Zhu, H. Mechanism Profiling of Hepatotoxicity Caused by Oxidative Stress Using Antioxidant Response Element Reporter Gene Assay Models and Big Data. *Environ. Health Perspect.* **2016**, *124*, 634–641, doi:10.1289/ehp.1509763.
50. Riniker, S.; Wang, Y.; Jenkins, J.L.; Landrum, G.A. Using Information from Historical High-Throughput Screens to Predict Active Compounds. *J. Chem. Inf. Model.* **2014**, *54*, 1880–1891, doi:10.1021/ci500190p.
51. Guo, Y.; Zhao, L.; Zhang, X.; Zhu, H. Using a Hybrid Read-across Method to Evaluate Chemical Toxicity Based on Chemical Structure and Biological Data. *Ecotoxicol. Environ. Saf.* **2019**, *178*, 178–187, doi:10.1016/j.ecoenv.2019.04.019.
52. Zhu, H.; Bouhifd, M.; Donley, E.; Egnash, L.; Kleinstreuer, N.; Kroese, E.D.; Liu, Z.; Luechtefeld, T.; Palmer, J.; Pamies, D.; et al. Supporting Read-across Using Biological Data. *ALTEX* **2016**, *33*, 167–182, doi:10.14573/altex.1601252.
53. Luechtefeld, T.; Marsh, D.; Rowlands, C.; Hartung, T. Machine Learning of Toxicological Big Data Enables Read-Across Structure Activity Relationships (RASAR) Outperforming Animal Test Reproducibility. *Toxicol. Sci.* **2018**, *165*, 198–212, doi:10.1093/toxsci/kfy152.
54. Ball, T.; Barber, C.G.; Cayley, A.; Chilton, M.L.; Foster, R.; Fowkes, A.; Heghes, C.; Hill, E.; Hill, N.; Kane, S.; et al. Beyond Adverse Outcome Pathways: Making Toxicity Predictions from Event Networks, SAR Models, Data and Knowledge. *Toxicol. Res.* **2021**, *10*, 102–122, doi:10.1093/toxres/tfaa099.
55. Norinder, U.; Spjuth, O.; Svensson, F. Using Predicted Bioactivity Profiles to Improve Predictive Modeling. *J. Chem. Inf. Model.* **2020**, *60*, 2830–2837, doi:10.1021/acs.jcim.0c00250.
56. Garcia de Lomana, M.; Morger, A.; Norinder, U.; Buesen, R.; Landsiedel, R.; Volkamer, A.; Kirchmair, J.; Mathea, M. ChemBioSim: Enhancing Conformal Prediction of In Vivo Toxicity by Use of Predicted Bioactivities. *J. Chem. Inf. Model.* **2021**, *61*, 3255–3272, doi:10.1021/acs.jcim.1c00451.
57. Alves, V.M.; Capuzzi, S.J.; Braga, R.C.; Borba, J.V.B.; Silva, A.C.; Luechtefeld, T.; Hartung, T.; Andrade, C.H.; Muratov, E.N.; Tropsha, A. A Perspective and a New Integrated Computational Strategy for Skin Sensitization Assessment. *ACS Sustain. Chem. Eng.* **2018**, *6*, 2845–2859, doi:10.1021/acssuschemeng.7b04220.
58. CompTox Chemicals Dashboard. Available online: https://comptox.epa.gov/dashboard/ (accessed on 20 February 2021).
59. DrugBank Release Version 5.1.8. Available online: https://go.drugbank.com/releases/latest (accessed on 20 February 2021).
60. Epa, U.S. ToxCast & Tox21 Data Spreadsheet from Invitrodb_v3.3. Available online: https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data (accessed on 7 September 2020).
61. Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. eMolTox: Prediction of Molecular Toxicity with Confidence. *Bioinformatics* **2018**, *34*, 2508–2509, doi:10.1093/bioinformatics/bty135.
62. eChemPortal. Available online: https://www.echemportal.org/echemportal/ (accessed on 6 August 2020).
63. Scikit-Learn. Available online: https://scikit-learn.org/stable/ (accessed on 20 February 2021).
64. Landrum, G. RDKit. Available online: http://www.rdkit.org (accessed on 20 February 2021).
65. Matthews, B.W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451, doi:10.1016/0005-2795(75)90109-9.
66. Otsubo, Y.; Nishijo, T.; Miyazawa, M.; Saito, K.; Mizumachi, H.; Sakaguchi, H. Binary Test Battery with KeratinoSens™ and H-CLAT as Part of a Bottom-up Approach for Skin Sensitization Hazard Prediction. *Regul. Toxicol. Pharmacol.* **2017**, *88*, 118–124, doi:10.1016/j.yrtph.2017.06.002.
67. Asturiol, D.; Casati, S.; Worth, A. Consensus of Classification Trees for Skin Sensitisation Hazard Prediction. *Toxicol.* In Vitro **2016**, *36*, 197–209, doi:10.1016/j.tiv.2016.07.014.

68. Roberts, D.W.; Patlewicz, G. Non-Animal Assessment of Skin Sensitization Hazard: Is an Integrated Testing Strategy Needed, and If so What Should Be Integrated? *J. Appl. Toxicol.* **2018**, *38*, 41–50, doi:10.1002/jat.3479.
69. Wilm, A.; Garcia de Lomana, M.; Stork, C.; Mathai, N.; Hirte, S.; Norinder, U.; Kühnl, J.; Kirchmair, J. Predicting the skin sensitization potential of small molecules with machine learning models trained on biologically meaningful descriptors. *Zenodo* **2021**, doi:10.5281/zenodo.4761226.

7. Conclusion

Skin sensitization is a complex endpoint that may result in the development of ACD. In order to minimize this risk, skin sensitization should be addressed for the development and approval of new chemicals or consumer products. Although animal experiments have been considered the gold standard for the assessment of skin sensitization potential and potency for decades, it is now desired and partialy legally required to assess this endpoint by non-animal alternatives. To that end, several in vitro and in chemico assays address the different key events of the skin sensitization AOP experimentally. In addition, a variety of computational prediction tools are capable to predict the skin sensitization potential (and to a limited degree potency) from existing data. Such non-testing methods allow for a large number of compounds to be evaluated in a short time and with low financial and logistical expenses. At the beginning of this thesis, we extensively reviewed the existing data basis and the available computational tools to predict skin sensitization potential and potency [P1]. Even though a variety of computational tools for skin sensitization prediction are available, none of them are capable of serving as a stand-alone method for risk assessment. Transparency, defined reliability, and interpretability were identified as three important pillars for a useful and applicable model for the prediction of skin sensitization potential or potency.

Within this thesis, we aimed to advance computational methods for the prediction of skin sensitization potential by the developing and evaluating predictive and transparent ML models, accompanied by solid measures of reliability.

In our first study, we compiled and standardized the currently largest LLNA data set comprising 1416 compounds labeled with binary skin sensitization potential. A comparison of the new LLNA data set with three reference data sets comprising cosmetics, approved drugs, and pesticides, respectively, revealed a high overlap, especially between the LLNA data set and the cosmetics space (30%, 10% and 13% overlap between the LLNA data set and cosmetics, approved drugs, and pesticides, respectively). These findings indicate that a model built on the LLNA data will be of high relevance for most molecules commonly used in cosmetic products. Nevertheless, not all areas in chemical space covered by the reference data sets are well represented by LLNA. Therefore, a solid definition of the AD is strongly needed for a model based on LLNA data. We also evaluated the capability of different ML algorithms (SVM and RF) and different

sets of descriptors (eight sets of binary and continuous descriptors and different combinations of two of them) to correctly predict skin sensitization potential measured in the LLNA. In total, 58 different ML models were optimized regarding their hyperparameters and evaluated in 10-fold CV. By a clear margin, the lowest performance was found for the SVM and RF models based on the five bit long fingerprint encoding the absence or presence of five OASIS alerts for skin sensitization (ACC of 0.67 and 0.66, respectively and MCC of 0.29 and 0.27, respectively), whereas all other models showed comparable performance measures with ACCs between 0.71 and 0.76 and MCCs between 0.39 and 0.55. The combination of two sets of descriptors was only beneficial compared to the single descriptor sets if two complementary types of descriptors (i.e. physicochemical descriptors and structural fingerprint) were combined. In PCA a cluster of halogenated aliphatic chains with differing class labels could be identified. Since these molecules cannot be differentiated by Morgan2 fingerprints, physicochemical descriptors will be needed to predict activity of these molecules. Finally, five models were selected for further evaluation and complemented with a solid definition of the AD (based on the mean Tanimoto similarity to the five nearest neighbors in Morgan2 space) as well as two reliability measures (comprising the distance to the models decision threshold and the number of nearest neighbors with concurrent class label). Our findings could be confirmed on unseen test data. Two models accompanied by their AD and reliability measures have been implemented on a publicly available web server for easy and fast usability. Compared to existing models, our models have the advantage of high coverage and relevance due to the largest publicly available data basis and the advantage of a solid definition of the AD and two transparent and well proven measures of reliability that can select reliable predictions based on differently strict cutoffs.

In our second study, we refined our LLNA data set by an additional manual data curation step. Consequently, we reduced the size of our binary data set to 1285 compounds, while remarkably increasing the quality of the data. A subset of the data was labeled with ternary class information and comprised 634 non-sensitizers, 345 weak to moderate sensitizers and 84 strong to extreme sensitizers.
One of the best performing combinations of ML algorithms, descriptors, and hyperparameters found in the previous study (a RF model based on MACCS key fingerprint), was applied to the new data set and enveloped into a CP framework. The model was valid on all error significances ε investigated (ε = 0.05, 0.10, 0.20 and 0.30). At an error significance ε of 0.05 high performance (ACC and MCC of 0.89 and 0.78, respectively) could be achieved at the cost of reduced efficiency of 0.32. At an error significance ε of 0.30, a high efficiency of 0.92 could be achieved at still reasonable performance of ACC and MCC of 0.75 and 0.51, respectively. Investigation of the ternary class labels (if available), revealed that, by far, most of the compounds falsely predicted as non-sensitizers are experimentally assigned to the weak to moderate sensitizers (5% and 9%

of predictions at the significance of $\varepsilon = 0.10$ and 0.30, respectively), whereas the more dangerous wrong prediction of strong to extreme sensitizers as non-sensitizers occurs only rarely (0% and 2% of predictions at the significance of $\varepsilon = 0.10$ and 0.30, respectively). To enable ternary classification, the binary classifier was combined with a second binary CP classifier differentiating weak to moderate from strong to extreme sensitizers. The combined model was overall valid at the error significance of $\varepsilon = 0.20$ and 0.30, but suffered from poor PPV for strong and extreme sensitizers at the significance levels of 0.20 and 0.30. Nevertheless, comparison to another ternary classifier for skin sensitization prediction based on a subset of our data revealed that shortcomings in class-wise performance might be a general problem in multi-class prediction of LLNA outcomes. The binary CP model differentiating sensitizers from non-sensitizers was published on our web service under the name Skin Doctor CP.

In our third study, we evaluated the capacity of a newly developed set of 750 calculated bioactivity descriptors to describe skin sensitization potential of small molecules measured in the LLNA. A strict feature selection process, allowed us to select only ten of these descriptors for further modeling and analysis. Three of the selected descriptors directly encoded for the predicted activity of a molecule in three non-animal testing approaches related to skin sensitization (DPRA, KeratinoSens assay, and h-CLAT). Also for most of the other selected descriptors, a biological relation to the skin sensitization AOP could be observed. Reconstruction of our CP framework with a RF model based on the ten bioactivity descriptors only, achieved a performance comparable our former model based on the less interpretable MACCS key fingerprint: ACC, MCC and efficiency of the new model are 0.86, 0.72 and 0.39 at the error significance of 0.05 and 0.74, 0.49 and 0.95 at the error significance of 0.30, respectively. These results underscore the relevance of the descriptors selected for the skin sensitization outcome and pave the way for more interpretable models. Two different approaches combining the results from the old model based on MACCS key fingerprints and the new model based on bioactivity descriptors did not achieve a clear improvement of the final performance and are thus not recommended for further usage.
In addition, high relevance of the LLNA data set for compounds labeled as approved drugs, cosmetics, or pesticides could be demonstrated by a PCA in the chemical space spanned by the ten bioactivity descriptors. To some extent, a separation between sensitizers and non-sensitizers could be observed in the PCA plot of the LLNA data set within this descriptor space, also demonstrating the value of the selected descriptors. A further comparison of the MACCS key model and the bioactivity model revealed that the performance of the former is more dependent on the number of training samples available than the latter. This makes the bioactivity descriptors also promising candidates for modeling tasks with limited number of training instances.

As for many other endpoints [43], predictivity of the ML models for skin sensitization potential is strongly limited by the available experimental data. With the current data basis we could neither develop nor identify a trustworthy model with global ACCs higher than 0.8. In several examples, we demonstrated that higher performance could be achieved by further limiting the applicability of the models by a classic definition of the AD, an alternative reliability measure, or the application of a desired level of error significance within a CP framework. All these improvements come along with decreased coverage of the corresponding models. A more substantial improvement in a model's performance could only be achieved with additional experimental data. Since further human or animal data for cosmetic ingredients can only be expected in exceptional cases, the inclusion of experimental non-animal data into the modeling process (as for example in DAs or IATAs) could be a promising route to increase model's predictivity or coverage without a simultaneous decrease in the other corresponding values. The simultaneous learning of several endpoints within one ML model (called multi-task modeling) might also address the challenge resulting from the limited data basis for skin sensitization prediction (as well as for other toxicological endpoints) and increase model's performance [111].

Apart from the quantity of the data available for modeling, the quality of these data pose an upper limit for the predictivity of ML models. In this context, computational methods can assist the evaluation and reliability assessment of existing experimental data. For example, Roberts [112] reinterpreted existing LLNA data also by the aid of chemistry-based read-across, which could be assisted by computational tools. Similarly, our PCA of the LLNA data set in the chemical space spanned by the ten bioactivity descriptors gives some insight into regions with possible outliers. If applied carefully, computational methods could therefore be applied to increase the quality of the existing data sets on which future models could be built.

The bioactivity descriptors developed within this thesis demonstrated a promising alternative to classical descriptor sets for the development of a reliable and interpretable model, especially when applied to a sparse data set. Transfer of the 750 bioactivity descriptors combined with the feature selection process might result in similarly small, interpretable, and biologically meaningful sets of descriptors for other toxicological endpoints, which could support further predictive modeling as well as the investigation of the chemical space and linkage between modeling and the underlying biological processes. In the case of small data sets, better performance and/or larger coverage might be achieved by this process. The success achieved when applying the bioactivity descriptors also demonstrates the possibilities that come along with small purposive sets of descriptors in general. The development of new sets of descriptors which are well suited for toxicology prediction might therefore also come into focus for

further research.

This thesis has contributed to several valuable objectives on the way to an animal-free assessment of skin sensitization potential and will be helpful in paving the way for computational tools to become an accepted pillar in non-animal risk assessment.

Bibliography

[1] **Wilm, A.**, Kühnl, J., Kirchmair, J., Computational approaches for skin sensitization prediction, *Critical Reviews in Toxicology*, **48**(9) (2018) 738–760.

[2] **Wilm, A.**, Stork, C., Bauer, C., Schepky, A., Kühnl, J., Kirchmair, J., Skin doctor: Machine learning models for skin sensitization prediction that provide estimates and indicators of prediction reliability, *International Journal of Molecular Sciences*, **20**(19) (2019) 4833.

[3] **Wilm, A.**, Norinder, U., Agea, M. I., de Bruyn Kops, C., Stork, C., Kühnl, J., Kirchmair, J., Skin doctor CP: Conformal prediction of the skin sensitization potential of small organic molecules, *Chemical Research in Toxicology*, **34**(2) (2020) 330–344.

[4] **Wilm, A.**, Garcia de Lomana, M., Stork, C., Mathai, N., Hirte, S., Norinder, U., Kühnl, J., Kirchmair, J., Predicting the skin sensitization potential of small molecules with machine learning models trained on biologically meaningful descriptors, *Pharmaceuticals*, **14**(8) (2021) 790.

[5] Thyssen, J. P., Linneberg, A., Menné, T., Johansen, J. D., The epidemiology of contact allergy in the general population–prevalence and main findings, *Contact dermatitis*, **57**(5) (2007) 287–299.

[6] Sætterstrøm, B., Olsen, J., Johansen, J. D., Cost-of-illness of patients with contact dermatitis in denmark, *Contact Dermatitis*, **71**(3) (2014) 154–161.

[7] Diepgen, T. L., Scheidt, R., Weisshaar, E., John, S. M., Hieke, K., Cost of illness from occupational hand eczema in germany, *Contact Dermatitis*, **69**(2) (2013) 99–106.

[8] Thyssen, J. P., Giménez-Arnau, E., Lepoittevin, J.-P., Menné, T., Boman, A., Schnuch, A., The critical review of methodologies and approaches to assess the inherent skin sensitization potential (skin allergies) of chemicals part I, *Contact Dermatitis*, **66**(s1) (2012) 11–24.

[9] EU – European Union, Directive 2010/63/EU of the european parliament and of the council of 22 september 2010 on the protection of animals used for scientific purposes, *Official Journal of the European Union*, **L 276/34** (2010) 20.10.2010.

[10] Animal testing under reach. https://echa.europa.eu/animal-testing-under-reach. accessed on 22 February 2022.

[11] EU – European Union, Regulation (ec) no 1223/2009 of the european parliament and of the council of 30 november 2009 on cosmetic products, *Official Journal of the European Union*, **L 342/59** (2009) 22.12.2009.

[12] Tox21 program. https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html. accessed on 22 February 2022.

[13] Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., Kavlock, R. J., The ToxCast program for prioritizing toxicity testing of environmental chemicals, *Toxicological Sciences*, **95**(1) (2006) 5–12.

[14] Pistollato, F., Madia, F., Corvi, R., Munn, S., Grignard, E., Paini, A., Worth, A., Bal-Price, A., Prieto, P., Casati, S., Berggren, E., Bopp, S. K., Zuang, V., Current EU regulatory requirements for the assessment of chemicals and cosmetic products: challenges and opportunities for introducing new approach methodologies, *Archives of Toxicology*, **95**(6) (2021) 1867–1897.

[15] Pérez Santín, E., Rodríguez Solana, R., González García, M., García Suárez, M. D. M., Blanco Díaz, G. D., Cima Cabal, M. D., Moreno Rojas, J. M., López Sánchez, J. I., Toxicity prediction based on artificial intelligence: A multidisciplinary overview, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **11**(5) (2021) e1516.

[16] Di, P., Yin, Y., Jiang, C., Cai, Y., Li, W., Tang, Y., Liu, G., Prediction of the skin sensitising potential and potency of compounds via mechanism-based binary and ternary classification models, *Toxicology in Vitro*, **59** (2019) 204–214.

[17] Salah, S., Taieb, C., Demessant, A., Haftek, M. et al., Prevalence of skin reactions and self-reported allergies in 5 countries with their social impact measured through quality of life impairment, *International Journal of Environmental Research and Public Health*, **18**(9) (2021) 4501.

[18] Winkler, G. C., Perino, C., Araya, S. H., Bechter, R., Kuster, M., Barle, E. L., Classification of dermal sensitizers in pharmaceutical manufacturing, *Regulatory Toxicology and Pharmacology*, **72**(3) (2015) 501–505.

[19] Kimber, I., Gerberick, G. F. Skin sensitization: What is it? why is it important? what are the challenges? http://alttox.org/skin-sensitization-what-is-it-why-is-it-important-what-are-the-challenges/, 2007. accessed on 9 February 2022.

[20] OECD. The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins. https://www.oecd.org/env/the-adverse-outcome-pathway-for-skin-sensitisation-initiated-by-covalent-binding-to-proteins-9789264221444-en.htm, 2012. accessed on 17 April 2018.

[21] Karlberg, A.-T., Bergström, M. A., Börje, A., Luthman, K., Nilsson, J. L. G., Allergic contact dermatitis–formation, structural requirements, and reactivity of skin sensitizers, *Chemical Research in Toxicology*, **21**(1) (2008) 53–69.

[22] de Ávila, R. I., Lindstedt, M., Valadares, M. C., The 21st century movement within the area of skin sensitization assessment: From the animal context towards current human-relevant in vitro solutions, *Regulatory Toxicology and Pharmacology*, **108** (2019) 104445.

[23] Wittwehr, C., Aladjov, H., Ankley, G., Byrne, H. J., de Knecht, J., Heinzle, E., Klambauer, G., Landesmann, B., Luijten, M., MacKay, C., Maxwell, G., Meek, M. E. B., Paini, A., Perkins, E., Sobanski, T., Villeneuve Dan Waters,, K. M., Whelan, M., How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology, *Toxicological Sciences*, **155**(2) (2017) 326–336.

[24] Kimber, I., Hilton, J., Weisenberger, C., The murine local lymph node assay for identification of contact allergens: a preliminary evaluation of in situ measurement of lymphocyte proliferation, *Contact Dermatitis*, **21**(4) (1989) 215–220.

[25] Anderson, S. E., Siegel, P. D., Meade, B., The LLNA: a brief review of recent advances and limitations, *Journal of Allergy*, **2011** (2011) 1–10.

[26] Felter, S., Kern, P., Ryan, C., Allergic contact dermatitis: Adequacy of the default 10x assessment factor for human variability to protect infants and children, *Regulatory Toxicology and Pharmacology*, **99** (2018) 116–121.

[27] Basketter, D. A., Alépée, N., Ashikaga, T., Barroso, J., Gilmour, N., Goebel, C., Hibatallah, J., Hoffmann, S., Kern, P., Martinozzi-Teissier, S., Maxwell, G., Reisinger, K., Sakaguchi, H., Schepky, A., Tailhardat, M., Templier, M., Categorization of chemicals according to their relative human skin sensitizing potency, *Dermatitis*, **25**(1) (2014) 11–21.

[28] Api, A. M., Parakhia, R. A., O'Brien, D., Basketter, D. A., Fragrances categorized according to relative human skin sensitization potency, *Dermatitis*, **28** (2017) 299–307.

[29] Natsch, A., Landsiedel, R., Kolle, S. N., A triangular approach for the validation of new approach methods for skin sensitization, *ALTEX - Alternatives to Animal Experimentation*, **38**(4) (2021) 669–677.

[30] Hoffmann, S., Kleinstreuer, N., Alépée, N., Allen, D., Api, A. M., Ashikaga, T., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Goebel, C., Kern, P. S., Klaric, M., Kühnl, J., Lalko, J. F., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Parakhia, R., van Vliet, E., Zang, Q., Petersohn, D., Non-animal methods to predict skin sensitization (I): the cosmetics europe database, *Critical Reviews in Toxicology*, **48**(5) (2018) 344–358, PMID: 29474128.

[31] Mehling, A., Eriksson, T., Eltze, T., Kolle, S., Ramirez, T., Teubner, W., van Ravenzwaay, B., Landsiedel, R., Non-animal test methods for predicting skin sensitization potentials, *Archives of Toxicology*, **86**(8) (2012) 1273–1295.

[32] Reisinger, K., Hoffmann, S., Alépée, N., Ashikaga, T., Barroso, J., Elcombe, C., Gellatly, N., Galbiati, V., Gibbs, S., Groux, H., Hibatallah, J., Keller, D., Kern, P. S., Klaric, M., Kolle, S., Kuehnl, J., Lambrechts, N., Lindstedt, M., Millet, M., Martinozzi-Teissier, S., Natsch, A., Petersohn, D., Pike, I., Sakaguchi, H., Schepky, A. G., Tailhardat, M., Templier, M., Van Vliet, E., Maxwell, G., Systematic evaluation of non-animal test methods for skin sensitisation safety assessment, *Toxicology in Vitro*, **29**(1) (2015) 259–270.

[33] OECD, Test no. 442C: In chemico skin sensitisation: Assays addressing the adverse outcome pathway key event on covalent binding to proteins, *OECD Guidelines for the Testing of Chemicals*, **4** (2021) 40.

[34] OECD, Test no. 442D: In vitro skin sensitisation: ARE-Nrf2 Luciferase test method, *OECD Guidelines for the Testing of Chemicals*, **4** (2018) 51.

[35] OECD, Test no. 442E: In vitro skin sensitisation: In vitro skin sensitisation assays addressing the key event on activation of dendritic cells on the adverse outcome pathway for skin sensitisation, *OECD Guidelines for the Testing of Chemicals*, **4** (2018) 65.

[36] Richter, A., Schmucker, S. S., Esser, P. R., Traska, V., Weber, V., Dietz, L., Thierse, H.-J., Pennino, D., Cavani, A., Martin, S. F., Human T cell priming assay (hTCPA) for the identification of contact allergens based on naive T cells and DC–IFN-γ and TNF-α readout, *Toxicology in vitro*, **27**(3) (2013) 1180–1185.

[37] Kleinstreuer, N. C., Hoffmann, S., Alépée, N., Allen, D., Ashikaga, T., Casey, W., Clouet, E., Cluzel, M., Desprez, B., Gellatly, N., Göbel, C., Kern, P. S., Klaric, M., Kühnl, J., Martinozzi-Teissier, S., Mewes, K., Miyazawa, M., Strickland, J., van Vliet, E., Zang, Q., Petersohn, D., Non-animal methods to predict skin sensitization (II): an assessment of defined approaches, *Critical Reviews in Toxicology*, **48**(5) (2018) 359–374.

[38] Ezendam, J., Braakhuis, H. M., Vandebriel, R. J., State of the art in non-animal approaches for skin sensitization testing: from individual test methods towards testing strategies, *Archives of Toxicology*, **90**(12) (2016) 2861–2883.

[39] Gilmour, N., Kern, P. S., Alépée, N., Boislève, F., Bury, D., Clouet, E., Hirota, M., Hoffmann, S., Kühnl, J., Lalko, J. F., Miyazawa, M., Nishida, H., Osmani, A., Petersohn, D., Vliet, E. V., Klaric, M., Development of a next generation risk assessment framework for the evaluation of skin sensitisation of cosmetic ingredients, *Regulatory Toxicology and Pharmacology*, **116** (2020) 104721.

[40] Alves, V. M., Capuzzi, S. J., Braga, R. C., Borba, J. V., Silva, A. C., Luechtefeld, T., Hartung, T., Andrade, C. H., Muratov, E. N., Tropsha, A., A perspective and a new integrated computational strategy for skin sensitization assessment, *ACS Sustainable Chemistry & Engineering*, **6**(3) (2018) 2845–2859.

[41] Leontaridou, M., Gabbert, S., Van Ierland, E. C., Worth, A. P., Landsiedel, R., Evaluation of non-animal methods for assessing skin sensitisation hazard: A bayesian value-of-information analysis, *Alternatives to Laboratory Animals*, **44**(3) (2016) 255–269.

[42] Alves, V. M., Auerbach, S. S., Kleinstreuer, N., Rooney, J. P., Muratov, E. N., Rusyn, I., Tropsha, A., Schmitt, C., Curated data in—trustworthy in silico models out: The impact of data quality on the reliability of artificial intelligence models as alternatives to animal testing, *Alternatives to Laboratory Animals*, **49**(3) (2021) 73–82.

[43] Kramer, C., Kalliokoski, T., Gedeck, P., Vulpetti, A., The experimental uncertainty of heterogeneous public KI data, *Journal of Medicinal Chemistry*, **55**(11) (2012) 5165–5173.

[44] Hansch, C., Maloney, P. P., Fujita, T., Muir, R. M., Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients, *Nature*, **194**(4824) (1962) 178–180.

[45] Roberts, D., Williams, D., The derivation of quantitative correlations between skin sensitisation and physio-chemical parameters for alkylating agents, and their application to experimental data for sultones, *Journal of Theoretical Biology*, **99**(4) (1982) 807–825.

[46] Hartigan, J. A. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[47] Johnson, S. C., Hierarchical clustering schemes, *Psychometrika*, **32**(3) (1967) 241–254.

[48] McInnes, L., Healy, J., Melville, J., UMAP: Uniform manifold approximation and projection for dimension reduction, *arXiv preprint arXiv:1802.03426*, 2018.

[49] Jolliffe, I. T. *Principal Component Analysis*. Springer, 2002.

[50] Setosa. https://setosa.io/ev/principal-component-analysis/. accessed on 10 December 2022.

[51] Ho, T. K. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Volume 1, p. 278–282. IEEE, 1995.

[52] Fix, E., Hodges, J. L., Discriminatory analysis. nonparametric discrimination: Consistency properties, *International Statistical Review*, **57**(3) (1989) 238–247.

[53] Cortes, C., Vapnik, V., Support-vector networks, *Machine learning*, **20**(3) (1995) 273–297.

[54] Extensive guide to support vector machines. https://www.inovex.de/de/blog/support-vector-machines-guide/. accessed on 22 February 2022.

[55] Tropsha, A., Golbraikh, A., Predictive qsar modeling workflow, model applicability domains, and virtual screening, *Current Pharmaceutical Design*, **13**(34) (2007) 3494–3504.

[56] Klingspohn, W., Mathea, M., Ter Laak, A., Heinrich, N., Baumann, K., Efficiency of different measures for defining the applicability domain of classification models, *Journal of Cheminformatics*, **9**(1) (2017) 1–17.

[57] Vovk, V., Gammerman, A., Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[58] Norinder, U., Carlsson, L., Boyer, S., Eklund, M., Introducing conformal prediction in predictive modeling for regulatory purposes. a transparent and flexible alternative to applicability domain determination, *Regulatory Toxicology and Pharmacology*, **71**(2) (2015) 279–284.

[59] Norinder, U., Boyer, S., Conformal prediction classification of a large data set of environmental chemicals from ToxCast and Tox21 estrogen receptor assays, *Chemical Research in Toxicology*, **29**(6) (2016) 1003–1010.

[60] Bosc, N., Atkinson, F., Felix, E., Gaulton, A., Hersey, A., Leach, A. R., Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery, *Journal of Cheminformatics*, **11**(1) (2019) 1–16.

[61] Linusson, H., Norinder, U., Boström, H., Johansson, U., Löfström, T., On the Calibration of Aggregated Conformal Predictors, *Proceedings of the Sixth Workshop on Conformal and Probabilistic Prediction and Applications*, **60** (2017) 154–173.

[62] Jmol wiki – file formats. https://wiki.jmol.org/index.php/File_formats/Coordinates#XYZ. accessed on 21 March 2022.

[63] Heller, S. R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D., InChI, the IUPAC international chemical identifier, *Journal of Cheminformatics*, **7**(1) (2015) 1–34.

[64] Weininger, D., SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules, *Journal of Chemical Information and Computer Sciences*, **28**(1) (1988) 31–36.

[65] Todeschini, R., Consonni, V. *Handbook of molecular descriptors*, Volume 11. John Wiley & Sons, 2008.

[66] Durant, J. L., Leland, B. A., Henry, D. R., Nourse, J. G., Reoptimization of mdl keys for use in drug discovery, *Journal of chemical information and computer sciences*, **42**(6) (2002) 1273–1280.

[67] Pubchem substructure fingerprint. https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt. accessed on 12 December 2021.

[68] Daylight theory: fingerprints. https://www.daylight.com/dayhtml/doc/theory/theory.finger.html. accessed on 17 January 2022.

[69] Rogers, D., Hahn, M., Extended-connectivity fingerprints, *Journal of Chemical Information and Modeling*, **50**(5) (2010) 742–754.

[70] Morgan, H. L., The generation of a unique machine description for chemical structures – a technique developed at chemical abstracts service, *Journal of Chemical Documentation*, **5**(2) (1965) 107–113.

[71] Glen, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S., Smith, J., Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to adme, *IDrugs*, **9**(3) (2006) 199–204.

[72] Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C. T., Guha, R., Steinbeck, C., The chemistry development kit (CDK) v2. 0: Atom typing, depiction, molecular formulas, and substructure searching, *Journal of Cheminformatics*, **9**(1) (2017) 1–19.

[73] PaDEL-Descriptor. http://www.yapcwsoft.com/dd/padeldescriptor/. accessed on 10 Mai 2019.

[74] Landrum, G., RDKit. http://www.rdkit.org. accessed on 26 April 2019.

[75] Chemical Computing Group, Molecular Operating Environment (MOE). https://www.chemcomp.com/Products.htm. accessed on 12 June 2019.

[76] Yap, C. W., Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints, *Journal of Computational Chemistry*, **32**(7) (2011) 1466–1474.

[77] Mauri, A., Consonni, V., Pavan, M., Todeschini, R., Dragon software: An easy approach to molecular descriptor calculations, *Match*, **56**(2) (2006) 237–248.

[78] Kumar, V., Minz, S., Feature selection: a literature review, *SmartCR*, **4**(3) (2014) 211–229.

[79] Tibshirani, R., Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1) (1996) 267–288.

[80] Borba, J. V., Alves, V. M., Braga, R. C., Korn, D. R., Overdahl, K., Silva, A. C., Hall, S. U., Overdahl, E., Kleinstreuer, N., Strickland, J., Allen, D., Andrade, C. H., Muratov, E. N., Tropsha, A., STopTox: An in silico alternative to animal testing for acute systemic and topical toxicity, *Environmental Health Perspectives*, **130**(2) (2022) 027012.

[81] Borba, J. V., Braga, R. C., Alves, V. M., Muratov, E. N., Kleinstreuer, N., Tropsha, A., Andrade, C. H., Pred-skin: a web portal for accurate prediction of human skin sensitizers, *Chemical Research in Toxicology*, **34**(2) (2020) 258–267.

[82] Gleeson, D., Gleeson, M. P., Theoretical studies to estimate the skin sensitization potential of chemicals of the schiff base domain, *International Journal of Quantum Chemistry*, **120**(12) (2020) e26218.

[83] Kim, J. Y., Kim, M. K., Kim, K.-B., Kim, H. S., Lee, B.-M., Quantitative structure–activity and quantitative structure–property relationship approaches as alternative skin sensitization risk assessment methods, *Journal of Toxicology and Environmental Health, Part A*, **82**(7) (2019) 447–472.

[84] Kim, J. Y., Kim, K.-B., Lee, B.-M., Validation of quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) approaches as alternatives to skin sensitization risk assessment, *Journal of Toxicology and Environmental Health, Part A*, **84**(23) (2021) 945–959.

[85] Silva, F. A., Brites, G., Ferreira, I., Silva, A., Miguel Neves, B., Costa Pereira, J. L., Cruz, M. T., Evaluating skin sensitization via soft and hard multivariate modeling, *International Journal of Toxicology*, **39**(6) (2020) 547–559.

[86] Li, H., Bai, J., Zhong, G., Lin, H., He, C., Dai, R., Du, H., Huang, L., Improved defined approaches for predicting skin sensitization hazard and potency in humans, *ALTEX-Alternatives to Animal Experimentation*, **36**(3) (2019) 363–372.

[87] Ambe, K., Suzuki, M., Ashikaga, T., Tohkin, M., Development of quantitative model of a local lymph node assay for evaluating skin sensitization potency applying machine learning catboost, *Regulatory Toxicology and Pharmacology*, **125** (2021) 105019.

[88] Wang, C.-C., Lin, Y.-C., Wang, S.-S., Shih, C., Lin, Y.-H., Tung, C.-W., SkinSensDB: a curated database for skin sensitization assays, *Journal of Cheminformatics*, **9**(1) (2017) 1–6.

[89] Tung, C.-W., Wang, C.-C., Wang, S.-S., Mechanism-informed read-across assessment of skin sensitizers based on SkinSensDB, *Regulatory Toxicology and Pharmacology*, **94** (2018) 276–282.

[90] Tung, C.-W., Lin, Y.-H., Wang, S.-S., Transfer learning for predicting human skin sensitizers, *Archives of Toxicology*, **39** (2019) 931–940.

[91] Jaworska, J. S., Natsch, A., Ryan, C., Strickland, J., Ashikaga, T., Miyazawa, M., Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy, *Archives of Toxicology*, **89**(12) (2015) 2355–2383.

[92] Otsubo, Y., Nishijo, T., Mizumachi, H., Saito, K., Miyazawa, M., Sakaguchi, H., Adjustment of a no expected sensitization induction level derived from bayesian network integrated testing strategy for skin sensitization risk assessment, *The Journal of Toxicological Sciences*, **45**(1) (2020) 57–67.

[93] Golden, E., Macmillan, D. S., Dameron, G., Kern, P., Hartung, T., Maertens, A., Evaluation of the global performance of eight in silico skin sensitization models using human data, *ALTEX-Alternatives to animal experimentation*, **38**(1) (2021) 33–48.

[94] Ta, G. H., Weng, C.-F., Leong, M. K., In silico prediction of skin sensitization: Quo vadis?, *Frontiers in Pharmacology*, **12** (2021) 1052.

[95] Ball, T., Barber, C. G., Cayley, A., Chilton, M. L., Foster, R., Fowkes, A., Heghes, C., Hill, E., Hill, N., Kane, S. et al., Beyond adverse outcome pathways: making toxicity predictions from event networks, SAR models, data and knowledge, *Toxicology Research*, **10**(1) (2021) 102–122.

[96] Johnson, C., Ahlberg, E., Anger, L. T., Beilke, L., Benigni, R., Bercu, J., Bobst, S., Bower, D., Brigo, A., Campbell, S. et al., Skin sensitization in silico protocol, *Regulatory Toxicology and Pharmacology*, **116** (2020) 104688.

[97] Kuseva, C., Schultz, T. W., Yordanova, D., Tankova, K., Kutsarova, S., Pavlov, T., Chapkanov, A., Georgiev, M., Gissi, A., Sobanski, T. et al., The implementation of RAAF in the OECD QSAR toolbox, *Regulatory Toxicology and Pharmacology*, **105** (2019) 51–61.

[98] Selvestrel, G., Robino, F., Baderna, D., Manganelli, S., Asturiol, D., Manganaro, A., Russo, M. Z., Lavado, G., Toma, C., Roncaglioni, A. et al., SpheraCosmolife: a new tool for the risk assessment of cosmetic products, *ALTEX-Alternatives to Animal Experimentation*, **38**(4) (2021) 565–579.

[99] Di, P., Wu, Z., Yang, H., Li, W., Tang, Y., Liu, G., Prediction of the allergic mechanism of haptens via a reaction-substructure-compound-target-pathway network system, *Toxicology Letters*, **317** (2019) 68–81.

[100] Cosing – cosmetics – growth – european commission. http://ec.europa.eu/growth/tools-databases/cosing/. accessed on 26 April 2019.

[101] Comptox chemicals dashboard. https://comptox.epa.gov/dashboard/. (accessed on 20 February 2021).

[102] Drugbank release version 5.1.2. https://www.drugbank.ca. (accessed on 7 May 2019).

[103] Drugbank release version 5.1.8. https://www.drugbank.ca. (accessed on 20 February 2021).

[104] Eu pesticides database — european commission. http://ec.europa.eu/food/plant/pesticides/eu-pesticides-database/public/?event=activesubstance.selection&language=EN. (accessed on 25 February 2019.

[105] Garcia de Lomana, M., Morger, A., Norinder, U., Buesen, R., Landsiedel, R., Volkamer, A., Kirchmair, J., Mathea, M., ChemBioSim: Enhancing conformal prediction of in vivo toxicity by use of predicted bioactivities, *Journal of Chemical Information and Modeling*, **61**(7) (2021) 3255–3272.

[106] Stork, C., Wagner, J., Friedrich, N.-O., de Bruyn Kops, C., Šícho, M., Kirchmair, J., Hit dexter: A machine-learning model for the prediction of frequent hitters, *ChemMedChem*, **13**(6) (2018) 564–571.

[107] Molvs. molvs version 0.1.1. https://github.com/mcs07/MolVS. accessed on 26 April 2019.

[108] Apt Systemst Ltd. Aptsys.net OASIS. QSAR Toolbox 4.3. http://oasis-lmc.org/products/software/toolbox.aspx. accessed on 10 July 2019.

[109] Scikit-learn: Machine learning in python—scikit-learn 0.21.0 documentation. https: //scikit-learn.org/stable/. accessed on 10 Mai 2019.

[110] OECD. Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf, 2004. accessed on 10 January 2022.

[111] Sosnin, S., Vashurina, M., Withnall, M., Karpov, P., Fedorov, M., Tetko, I. V., A survey of multi-task learning methods in chemoinformatics, *Molecular Informatics*, **38**(4) (2019) 1800108.

[112] Roberts, D. W., Interpretation of murine local lymph node assay (LLNA) data for skin sensitization: Overload effects, danger signals and chemistry-based read-across, *Current Research in Toxicology*, **2** (2021) 53–63.

9. Appendix

A Gefahrstoffe nach GHS

In this work no hazardous compounds according to the GHS (Globally Harmonized System Of Classification and Labeling of Chemicals) were used.

B Supporting information for publications originating from this work

B.1 Supporting information for publication [P2]

This appendix contains the supporting information for the publication:

Wilm, A., Stork, C., Bauer, C., Schepky, A., Kühnl, J., Kirchmair, J., Skin doctor: Machine learning models for skin sensitization prediction that provide estimates and indicators of prediction reliability, *International Journal of Molecular Sciences*, **20**(19) (2019) 4833

International Journal of
Molecular Sciences

MDPI

Supporting Information for

Skin Doctor: Machine Learning Models for Skin Sensitization Prediction that Provide Estimates and Indicators of Prediction Reliability

Anke Wilm [1,2], Conrad Stork [1], Christoph Bauer [3,4], Andreas Schepky [5], Jochen Kühnl [5] and Johannes Kirchmair [1,3,4*]

[1] Center for Bioinformatics, Universität Hamburg, Hamburg, Germany
[2] HITeC e.V, Hamburg, Germany
[3] Department of Chemistry, University of Bergen, Bergen, Norway
[4] Computational Biology Unit (CBU), University of Bergen, Bergen, Norway
[5] Front End Innovation, Beiersdorf AG, Hamburg, Germany

* Correspondence: kirchmair@zbh.uni-hamburg.de; Tel.: +49-40-42838-7303.

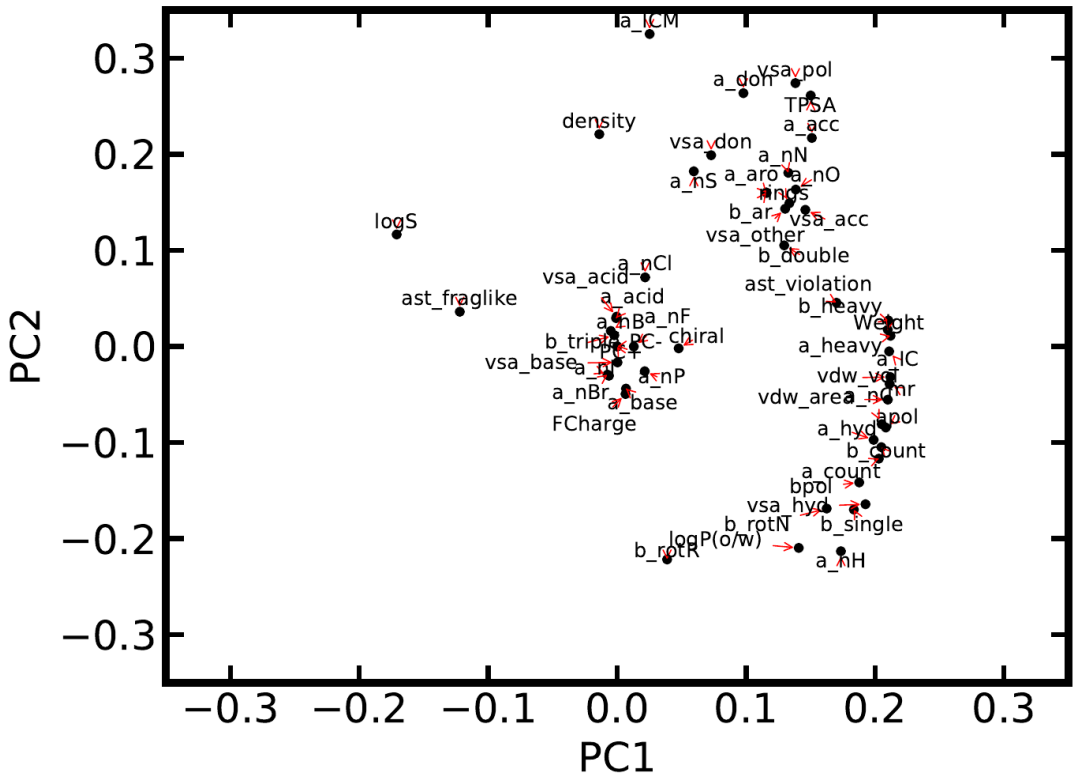


Figure S1. Enlarged version of the loadings plot from Figure 5B. For an explanation of the abbreviations see Table S1.

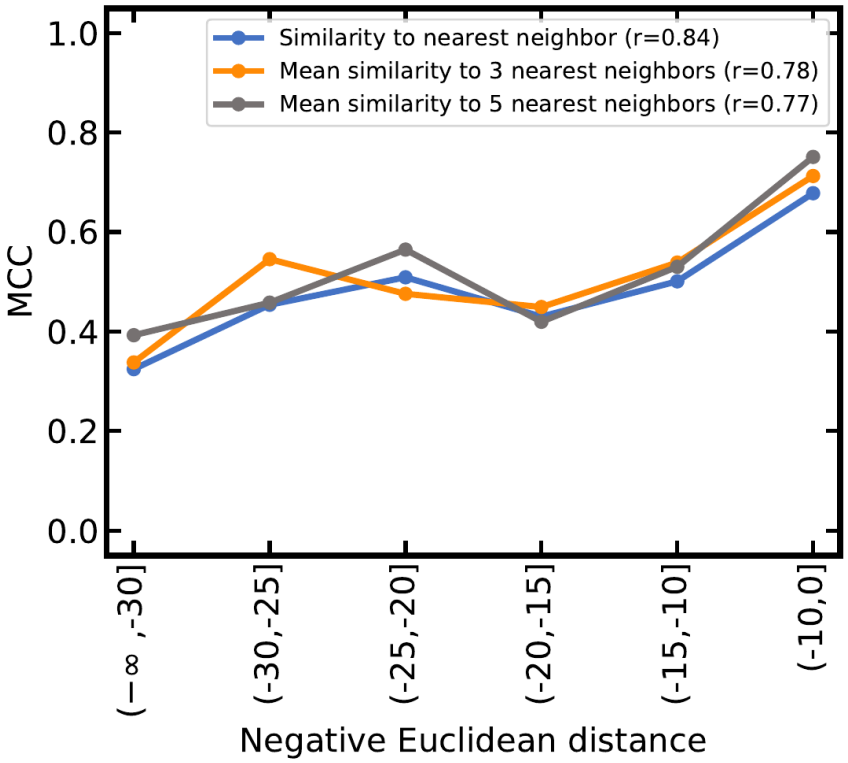


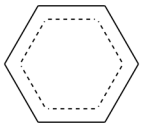
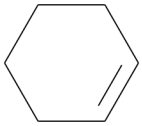
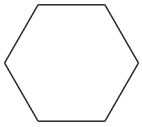
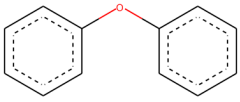
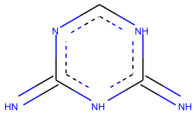
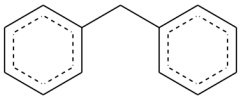
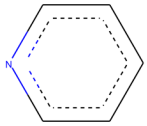
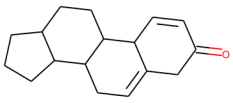
Figure S2. Correlation between molecular similarity measured as negative Euclidean distance in PaDEL space for the SVM_PaDEL model. Number of compounds in each bin are reported in Table S6.

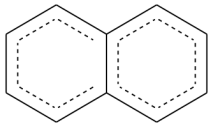
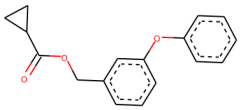
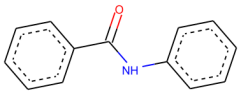
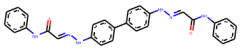
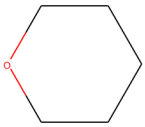
Table S1. Descriptors Used for the PCA and Explanation of the Abbreviations.

Descriptor	Explanation
apol	Polarizabilities of all atoms in molecule (as sum)
ast_fraglike	Binary Astex fragment-likeness
ast_violation	Number of Astex fragment-likeness violations
a_acc	H-bond acceptor atom count
a_acid	Acidic atom count
a_aro	Aromatic atom count
a_base	Basic atom count
a_count	Atom count
a_don	H-bond donor count
a_heavy	Heavy atom count
a_hyd	Hydrophobic atom count
a_IC	Total atom information content
a_ICM	Mean atom information content
a_nB	Boron atom count
a_nBr	Bromine atom count
a_nC	Carbon atom count
a_nCl	Chlorine atom count
a_nF	Fluorine atom count
a_nH	Hydrogen atom count
a_nI	Iodine atom count
a_nN	Nitrogen atom count
a_nO	Oxygen atom count
a_nP	Phosphorus atom count
a_nS	Sulfur atom count
bpol	Bonded atom polarizability difference
b_ar	Number of aromatic bonds
b_count	Number of bonds
b_double	Number of double bonds

b_heavy	Number of bonds between heavy atoms
b_rotN	Number of rotatable bonds
b_rotR	Fraction of rotatable bonds
b_single	Number of single bonds
b_triple	Number of triple bonds
chiral	Number of chiral centers
density	Molecular mass density
FCharge	Total charge of the molecule
logP(o/w)	Log of the octanol/water partition coefficient
logS	Log of the aqueous solubility (mol/L)
mr	Molecular refractivity
PC+	Total positive partial charge
PC-	Total negative partial charge
rings	Number of rings
TPSA	Polar surface area (Å^2)
vdw_area	Area of van der Waals surface (Å^2)
vdw_vol	Van der Waals volume (Å^3)
vsa_acc	Approximation to the sum of VDW surface areas (Å^2) of pure hydrogen bond acceptors
vsa_acid	Approximation to the sum of VDW surface areas of acidic atoms (Å^2)
vsa_base	Approximation to the sum of VDW surface areas of basic atoms (Å^2)
vsa_don	Approximation to the sum of VDW surface areas of pure hydrogen bond donors
vsa_hyd	Approximation to the sum of VDW surface areas of hydrophobic atoms (Å^2)
vsa_other	Approximation to the sum of VDW surface areas (Å^2) of atoms typed as "other"
vsa_pol	Approximation to the sum of VDW surface areas (Å^2) of polar atoms
Weight	Molecular weight

Table S2. Ten Most Prevalent Murcko Scaffolds in the LLNA Data Set.[1]

	LLNA of Alves et al.	LLNA of Di et al.	Merged LLNA	Cosmetics	Drugs	Pesticides
	30.12%	23.44%	27.04%	27.50%	10.72%	23.46%
	1.32%	1.85%	1.73%	3.59%	0.21%	
	2.05%	1.56%	2.04%	3.55%	0.54%	0.49%
	0.29%	0.28%	0.20%	0.30%	0.32%	2.75%
	0.15%	0.14%	0.10%	0.07%		1.94%
	1.75%	1.14%	1.53%	1.04%	0.80%	1.62%
	1.90%	1.56%	1.53%	0.85%	0.75%	1.13%
	0.15%	0.14%	0.10%	0.04%	1.39%	

	0.58%	0.28%	0.51%	0.89%	0.38%	1.29%
		0.14%	0.10%	0.07%	0.05%	1.29%
	0.29%	1.28%	0.92%	0.15%	0.05%	1.13%
	1.46%	1.56%	1.12%	0.11%		
	0.29%	0.28%	0.41%	1.04%	0.32%	

[1] Reported are the percentages of compounds based on the indicated Murcko scaffolds among all compounds having a Murcko scaffold.

Table S3. Hyperparameters Selected During Grid Search.[1]

	RF		SVM	
Name	**n_estimators**	**max_features**	**C**	**gamma**
MOE2D	250	0.4	1000	0.0001
MOE2D53	250	0.4	1000	0.001
Padel	250	0.8	1	0.001
MACCS	1000	sqrt	1	0.1
Morgan2	100	0.2	100	0.1
OASIS	10	sqrt	1	0.1
Padel-Est	1000	0.4	10	0.1
Padel-Ext	100	0.4	1	0.01
MOE2D+Padel	500	None	1	0.001
MOE2D+MACCS	500	0.2	10	0.01
MOE2D+Morgan2	500	0.4	10	0.001
MOE2D+OASIS	100	None	100	0.001
MOE2D+Padel-Est	1000	0.4	10	0.01
MOE2D+Padel-Ext	1000	sqrt	10	0.001
Padel+MACCS	500	0.4	1	0.001
Padel+Morgan2	1000	0.2	100	0.001
Padel+OASIS	500	0.6	1	0.001
Padel+Padel-Est	1000	sqrt	1	0.001
Padel+Padel-Ext	50	0.8	1	0.001
MACCS+Morgan2	50	0.8	10	0.01
MACCS+OASIS	50	None	1	0.1
MACCS+Padel-Est	250	sqrt	1	0.1
MACCS+Padel-Ext	50	0.2	1	0.01
Morgan2+OASIS	100	sqrt	100	0.1
Morgan2+Padel-Est	250	sqrt	10	0.01
Morgan2+Padel-Ext	1000	sqrt	1	0.01
OASIS+Padel-Est	50	0.4	10	0.1
OASIS+Padel-Ext	1000	0.6	1	0.01
Padel-Est+Padel-Ext	250	0.6	1	0.01

[1] Definitions of the individual descriptor sets are provided in Table 2.

Table S4. Matthews Correlation Coefficients for the RF Models.[1]

	MOE2D	PaDEL	Morgan2	PaDEL-Ext	PaDEL-Est	MACCS	OASIS
MOE2D	0.44	0.48	0.46	0.45	0.45	0.44	0.45
PaDEL		0.48	0.49	0.49	0.47	0.49	0.49
Morgan2			0.46	0.44	0.48	0.44	0.44
PaDEL-Ext				0.42	0.43	0.43	0.43
PaDEL-Est					0.43	0.46	0.48
MACCS						0.47	0.47
OASIS							0.27

[1] The diagonal reports MCC values for models based on a single set of descriptors.

Table S5. Matthews Correlation Coefficients for the SVM Models.[1]

	MOE2D	PaDEL	Morgan2	PaDEL-Ext	PaDEL-Est	MACCS	OASIS
MOE2D	0.48	0.5	0.5	0.5	0.5	0.5	0.55
PaDEL		0.5	0.51	0.51	0.5	0.51	0.5
Morgan2			0.39	0.48	0.43	0.46	0.43
PaDEL-Ext				0.47	0.47	0.46	0.47
PaDEL-Est					0.44	0.49	0.47
MACCS						0.47	0.48
OASIS							0.29

[1]The diagonal reports MCC values for models based on a single set of descriptors.

Table S6. Number of Compounds with Specified negative Euclidean distance to 1, 3 and 5 Nearest Neighbors of SVM_PaDEL model in PaDEL space.

	(-∞ ,-30]	(-30,-25]	(-25,-20]	(-20,-15]	(-15,-10]	(-10,0]
Similarity to nearest neighbor	138	112	193	267	259	140
Mean similarity to 3 nearest neighbors	174	148	237	295	207	48
mean similarity to 5 nearest neighbors	200	174	259	288	171	17

Table S7. Number of Compounds with Specified Mean Tanimoto Similarity to 1, 3 and 5 Nearest Neighbors.

		Mean Tanimoto similarity					
Model	**Number of neighbors considered**	**[0 ,0.5]**	**(0.5,0.6]**	**(0.6,0.7]**	**(0.7,0.8]**	**(0.8,0.9]**	**(0.9,1]**
SVM_MOE2D+OASIS	1	24	89	218	327	244	226
	3	38	140	339	334	154	123
	5	53	207	374	289	140	65
SVM_PaDEL+OASIS	1	19	92	198	320	252	228
	3	34	134	317	343	164	117
	5	44	204	362	297	132	70
SVM_PaDEL	1	19	92	198	320	252	228
	3	34	134	317	343	164	117
	5	44	204	362	297	132	70
RF_MACCS	1	25	89	207	335	242	234
	3	38	138	329	344	168	115
	5	56	204	374	296	135	67
SVM_PaDEL+MACCS	1	19	92	198	320	252	228
	3	34	134	317	343	164	117
	5	44	204	362	297	132	70

Table S8. Number of Compounds with Specified Distances Between the Prediction Probability and the Decision Threshold.

	Distance							
Model	**[0 ,0.25]**	**(0.25 - 0.5]**	**(0.5 - 0.75]**	**(0.75 - 1]**	**(1 - 1.25]**	**(1.25 - 1.5]**	**(1.5 - 1.75]**	**(1.75,∞)**
SVM_MOE2D+OASIS	124	159	133	125	121	100	88	278
SVM_PaDEL+OASIS	183	233	198	174	173	91	48	9
SVM_PaDEL	180	238	193	177	172	92	47	10
SVM_PaDEL+MACCS	182	237	198	172	174	90	48	8
	[0 ,0.1]	**(0.1,0.15]**	**(0.15,0.2]**	**(0.2,0.25]**	**(0.25,0.3]**	**(0.3,0.35]**	**(0.35,0.4]**	**(0.4,0.5)**
RF_MACCS	237	134	126	128	126	113	73	195

Table S9. Number of Compounds with Specified Numbers of Consecutive Nearest Neighbors with Same Activity as Predicted.

Model	**0**	**1**	**2**	**3**	**4**	**5 or more**
SVM_MOE2D+OASIS	308	201	142	102	71	304
SVM_PaDEL+OASIS	295	213	124	94	83	300
SVM_PaDELl	295	213	124	94	83	300
RF_MACCS	329	187	135	104	78	299
SVM_PaDEL+MACCS	294	213	124	94	83	301

B.2 Supporting information for publication [P3]

This appendix contains information on the supporting information for the publication:

Wilm, A., Garcia de Lomana, M., Stork, C., Mathai, N., Hirte, S., Norinder, U., Kühnl, J., Kirchmair, J., Predicting the skin sensitization potential of small molecules with machine learning models trained on biologically meaningful descriptors, *Pharmaceuticals*, **14**(8) (2021) 790

The supporting information of this publication [P3] can be downloaded free of charge from the following link:
https://pubs.acs.org/doi/10.1021/acs.chemrestox.0c00253.

B.3 Supporting information for publication [P4]

This appendix contains the supporting information for the publication:

Wilm, A., Garcia de Lomana, M., Stork, C., Mathai, N., Hirte, S., Norinder, U., Kühnl, J., Kirchmair, J., Predicting the skin sensitization potential of small molecules with machine learning models trained on biologically meaningful descriptors, *Pharmaceuticals*, **14**(8) (2021) 790

Supporting Information

Predicting the Skin Sensitization Potential of Small Molecules with Machine Learning Models Trained on Biologically Meaningful Descriptors

Anke Wilm [1,2], Marina Garcia de Lomana [3], Conrad Stork [1], Neann Mathai [4], Steffen Hirte [3], Ulf Norinder [5,6,7], Jochen Kühnl [8] and Johannes Kirchmair [1,3*]

[1] Center for Bioinformatics (ZBH), Department of Informatics, Universität Hamburg, 20146 Hamburg, Germany; wilm@zbh.uni-hamburg.de (A.W.); stork@zbh.uni-hamburg.de (C.S.)
[2] HITeC e.V., 22527 Hamburg, Germany
[3] Department of Pharmaceutical Sciences, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria; a11853333@unet.univie.ac.at (M.G.d.L.); steffen.hirte@univie.ac.at (S.H.)
[4] Computational Biology Unit (CBU), Department of Chemistry, University of Bergen, N-5020 Bergen, Norway; neann.mathai@uib.no
[5] MTM Research Centre, School of Science and Technology, Örebro University, SE-70182 Örebro, Sweden; ulf.norinder@farmbio.uu.se
[6] Department of Computer and Systems Sciences, Stockholm University, SE-16407 Kista, Sweden
[7] Department of Pharmaceutical Biosciences, Uppsala University, SE-75124 Uppsala, Sweden
[8] Front End Innovation, Beiersdorf AG, 22529 Hamburg, Germany; Jochen.Kuehnl@Beiersdorf.com
* Correspondence: johannes.kirchmair@univie.ac.at; Tel.: +43-1-4277-55104

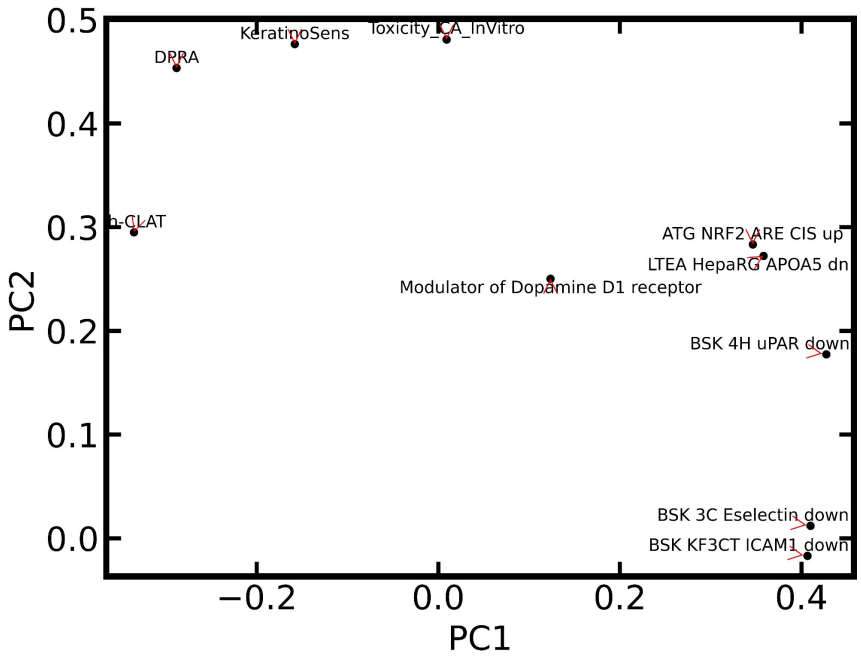


Figure S1. Loadings plot for the PCA on the LLNA and the three reference data sets, based on the ten selected bioactivity descriptors.

Table S1. Mean absolute Lasso coefficients and standard deviation σ retrieved from the 10-fold cross-validation.

Assay name	Mean Lasso coefficient	σ(Lasso coefficient)	Correlation to positive assay outcome
p0 BSK KF3CT ICAM1 down	0.074	0.0088	positive
p1 BSK 4H uPAR down	0.051	0.0454	negative
p0 CA	0.049	0.0096	positive
p1 DPRA	0.047	0.0125	positive
p1 Modulator of Dopamine D1 receptor	0.045	0.0064	positive
p1-h-CLAT	0.043	0.0134	positive
p1 BSK 3C Eselectin down	0.043	0.0210	positive
p1 LTEA HepaRG APOA5 dn	0.040	0.0123	negative
p1-KeratinoSens	0.039	0.0036	positive
p0 ATG NRF2 ARE CIS up	0.036	0.0142	positive
p0 Modulator of Muscarinic acetylcholine receptor M1	0.036	0.0145	positive
p0 Inhibitors and Substrates of Cytochrome P450 2C9	0.032	0.0064	positive

p1 OT ER ERaERb 1440	0.026	0.0129	positive
p1 AMES	0.026	0.0098	positive
p1 LTEA HepaRG FABP1 dn	0.025	0.0144	negative
p1 BSK hDFCGF IP10 down	0.025	0.0200	positive
p1 Activators of the human pregnane X receptor (PXR) signaling pathway	0.025	0.0164	negative
p0 TOX21 RAR LUC Agonist	0.022	0.0095	negative
p1 BSK LPS TNFa down	0.022	0.0212	negative
p1 TOX21 MMP ratio up	0.022	0.0125	negative
p1 TOX21 ERa BLA Agonist ratio	0.021	0.0154	negative
p1 UPITT HCI U2OS AR TIF2 Nucleoli Antagonist	0.021	0.0145	positive
p0 Modulator of Muscarinic acetylcholine receptor M4	0.020	0.0074	positive
p0 OT AR ARSRC1 0480	0.020	0.0204	positive
p1 Modulator of Melatonin receptor 1B	0.019	0.0069	negative
p1 LTEA HepaRG ABCB1 up	0.019	0.0100	negative
p0 Induce genoin human embryonic kidney cells	0.019	0.0092	negative
p1 TOX21 HDAC Inhibition	0.018	0.0156	positive
p0 Modulator of Monoamine oxidase A	0.018	0.0044	positive
p0 TOX21 TR LUC GH3 Antagonist	0.017	0.0234	positive
p0 Mutagenicity	0.016	0.0134	negative
p0 LTEA HepaRG CYP2E1 dn	0.015	0.0167	positive
p1 ATG RORE CIS up	0.015	0.0107	negative
p1 ATG DR4 LXR CIS dn	0.014	0.0092	positive
p1 Modulator of Androgen Receptor	0.013	0.0070	negative
p1 Differential cyto(isogenic chicken DT40 Rev3 mutant cell line)	0.013	0.0104	positive
p1 Block Bile Salt Export Pump	0.013	0.0103	negative
p0 Modulator of Adenosine A1 receptor	0.013	0.0066	negative
p0 Agonist of the AP-1 signaling pathway	0.013	0.0166	positive
p1 LTEA HepaRG CYP1A1 up	0.012	0.0093	positive
p1 Inhibitors and Substrates of Cytochrome P450 2D6	0.012	0.0111	positive
p1 TOX21 FXR BLA antagonist ratio	0.011	0.0182	positive
p0 UPITT HCI U2OS AR TIF2 Nucleoli Agonist	0.011	0.0142	negative

p0 LTEA HepaRG CYP1A2 up	0.011	0.0087	positive
p0 BSK 3C Eselectin down	0.011	0.0141	positive
p0 Modulator of Platelet activating factor receptor	0.011	0.0072	negative
p0 NHEERL ZF 144hpf TERATOSCORE up	0.011	0.0077	positive
p1 Agonist of the RXR signaling pathway	0.010	0.0074	negative
p1 TOX21 AP1 BLA Agonist ratio	0.010	0.0138	negative
p0 TOX21 PR BLA Antagonist ratio	0.010	0.0125	negative
p1 Caco2	0.009	0.0113	positive
p1 BSK hDFCGF MCSF down	0.009	0.0069	positive
p1 Differential cytoagainst isogenic chicken DT40 cell lines with known DNA damage response pathways Rad54Ku70 mutant cell line	0.008	0.0086	positive
p1 TOX21 AhR LUC Agonist	0.008	0.0107	negative
p0 NCCT HEK293T CellTiterGLO	0.008	0.0096	positive
p0 Antagonist of the retinoic acid receptor (RAR) signaling pathway	0.008	0.0108	negative
p1 TOX21 ERa LUC VM7 Agonist	0.008	0.0036	negative
p1 ATG RXRb TRANS up	0.007	0.0065	positive
p1 TOX21 MMP ratio down	0.007	0.0127	positive
p1 Modulator of Calcitonin gene-related peptide type 1 receptor	0.007	0.0061	positive
p0 Modulator of Glutamate NMDA receptor	0.007	0.0067	negative
p0 Modulator of Neurokinin 2 receptor	0.007	0.0066	negative
p1 BSK hDFCGF TIMP1 down	0.007	0.0139	positive
p0 Modulator of Adenosine A3 receptor	0.007	0.0101	negative
p1 ATG NRF2 ARE CIS up	0.006	0.0079	positive
p1 Modulator of Dopamine transporter	0.006	0.0063	positive
p1 ATG Ets CIS dn	0.006	0.0084	negative
p0 Cytoin HepG2 cells 40 hour	0.006	0.0106	negative
p1 ATG PBREM CIS up	0.006	0.0101	negative
p0 Inhibit CYP1A2 Activity	0.006	0.0119	positive
p1 LTEA HepaRG ALPP dn	0.006	0.0169	negative
p1 CA	0.006	0.0104	positive

p0 Modulator of Neuronal acetylcholine receptor alpha4beta2	0.006	0.0072	positive
p0 Block Bile Salt Export Pump	0.005	0.0108	negative
p1 TOX21 RXR BLA Agonist ratio	0.005	0.0060	negative
p1 BSK BE3C IL1a down	0.005	0.0141	negative
p0 Modulator of Melatonin receptor 1B	0.005	0.0040	negative
p1 ATG HIF1a CIS up	0.005	0.0053	negative
p0 Modulator of Receptor protein-tyrosine kinase erbB-2	0.005	0.0084	positive
p0 OT ER ERaERb 1440	0.005	0.0117	positive
p0 Modulator of Cholecystokinin A receptor	0.005	0.0051	negative
p1 Disruptors of the mitochondrial membrane potential	0.005	0.0069	positive
p0 Modulator of Sodium channel protein type IX alpha subunit	0.004	0.0046	negative
p1 UPITT HCI U2OS AR TIF2 Nucleoli Agonist	0.004	0.0065	negative
p0 BSK CASM3C MCP1 down	0.004	0.0074	positive
p0 Modulator of GABA-A receptor alpha-1beta-3gamma-2	0.003	0.0059	negative
p0 LTEA HepaRG CYP1A1 up	0.003	0.0075	positive
p0 Modulator of Neuronal acetylcholine receptor protein alpha-7 subunit	0.003	0.0060	negative
p0 Cytoin HepG2 cells 32 hour	0.003	0.0070	negative
p1 Modulator of Sodium channel protein type IX alpha subunit	0.003	0.0033	negative
p1 ATG C EBP CIS up	0.003	0.0055	negative
p1 Modulator of Acetylcholinesterase	0.003	0.0034	positive
p1 BSK hDFCGF Proliferation down	0.003	0.0044	positive
p1 OT FXR FXRSRC1 1440	0.003	0.0085	negative
p0 Modulator of Serotonin 7 (5-HT7) receptor	0.003	0.0050	positive
p1 Modulator of GABA-A receptor alpha-2beta-3gamma-2	0.003	0.0038	negative
p1 Antagonist of the estrogen receptor alpha (ER-alpha) signaling pathway	0.003	0.0052	negative
p1 ATG E Box CIS dn	0.003	0.0080	positive
p1 Modulator of Serotonin 2b (5-HT2b) receptor	0.003	0.0046	negative

p1 ATG ERa TRANS up	0.003	0.0039	positive
p1 TOX21 TSHR Agonist ratio	0.002	0.0061	positive
p1 Modulator of Serotonin 7 (5-HT7) receptor	0.002	0.0026	negative
p0 Modulator of Dopamine transporter	0.002	0.0044	positive
p1 BSK SAg CD69 down	0.002	0.0068	positive
p1 ATG BRE CIS up	0.002	0.0040	negative
p1 ACEA ER 80hr	0.002	0.0052	negative
p1 Modulator of Adenosine A1 receptor	0.002	0.0032	negative
p1 APR HepG2 CellLoss 72h dn	0.002	0.0059	negative
p0 Activators of the human pregnane X receptor (PXR) signaling pathway	0.002	0.0043	negative
p0 Modulator of Norepinephrine transporter	0.002	0.0030	positive
p0 Modulator of Vascular endothelial growth factor receptor 2	0.002	0.0054	positive
p0 BSK CASM3C MCSF down	0.002	0.0029	positive
p1 Modulator of Alpha-1a adrenergic receptor	0.002	0.0035	positive
p1 BSK hDFCGF CollagenIII down	0.002	0.0034	positive
p0 Modulator of Serotonin 2b (5-HT2b) receptor	0.002	0.0030	negative
p0 Modulators of myocardial damage	0.002	0.0026	positive
p0 Modulator of HERG	0.002	0.0048	negative
p1 BSK CASM3C MCSF down	0.002	0.0048	positive
p1 ATG PXR TRANS up	0.002	0.0048	positive
p1 Modulator of Alpha-2a adrenergic receptor	0.002	0.0024	positive
p0 Modulator of Serotonin 1b (5-HT1b) receptor	0.002	0.0037	negative
p0 Modulator of Peroxisome proliferator-activated receptor gamma	0.001	0.0041	negative
p1 Modulator of P2X purinoceptor 7	0.001	0.0019	negative
p0 Modulator of Cannabinoid CB2 receptor	0.001	0.0043	positive
p0 Modulator of P2X purinoceptor 3	0.001	0.0042	positive
p1 Activator the aryl hydrocarbon receptor (AhR) signaling pathway	0.001	0.0028	negative
p1 Modulator of Serotonin 1b (5-HT1b) receptor	0.001	0.0027	negative
p1 ATG PPARg TRANS up	0.001	0.0028	positive
p0 Modulator of Delta opioid receptor	0.001	0.0032	positive

p1 ATG ISRE CIS dn	0.001	0.0025	negative
p1 Modulator of Histamine H1 receptor	0.001	0.0024	positive
p1 Modulator of Platelet-derived growth factor receptor beta	0.001	0.0026	positive
p1 ACEA AR antagonist 80hr	0.001	0.0035	negative
p1 DIO1	0.001	0.0032	positive
p0 Differential cytoagainst isogenic chicken DT40 cell lines with known DNA damage response pathways Rad54Ku70 mutant cell line	0.001	0.0033	positive
p0 Modulator of Calcitonin gene-related peptide type 1 receptor	0.001	0.0032	negative
p1 TOX21 ERR Agonist	0.001	0.0032	positive
p1 TOX21 DT40	0.001	0.0032	positive
p1 Modulator of Neuronal acetylcholine receptor alpha4beta2	0.001	0.0014	negative
p0 Caco2	0.001	0.0032	positive
p1 TOX21 AR LUC MDAKB2 Agonist	0.001	0.0032	negative
p1 Inhibitors of Hepatocyte nuclear factor 4 (HNF4) dimerization	0.001	0.0031	positive
p0 Modulator of Neurokinin 1 receptor	0.001	0.0029	negative
p1 Modulator of Adenosine A2a receptor	0.001	0.0026	negative
p1 Antagonist of the farnesoid-X-receptor (FXR) signaling pathway	0.001	0.0021	negative
p1 Modulator of Dopamine D2 receptor	0.001	0.0020	positive
p0 AMES	0.001	0.0014	positive
p0 LTEA HepaRG UGT1A1 up	0.001	0.0018	positive
p1 Modulator of GABA-A receptor alpha-1beta-3gamma-2	0.001	0.0011	negative
p0 TOX21 PGC ERR Agonist	0.001	0.0016	negative
p1 TOX21 CAR Agonist	0.001	0.0016	negative
p1 TOX21 DT40 657	0.001	0.0012	positive
p0 Modulator of Angiotensin-converting enzyme	0.001	0.0016	positive
p1 Antagonist of the vitamin D receptor (VDR) signaling pathway	0.001	0.0015	positive
p1 Modulator of Serotonin 4 (5-HT4) receptor	0.001	0.0011	negative

p0 ATG DR4 LXR CIS dn	0.000	0.0015	positive
p0 TOX21 TSHR Agonist ratio	0.000	0.0014	positive
p0 TOX21 MMP ratio up	0.000	0.0014	negative
p1 Modulator of GABA-A receptor alpha-5beta-3gamma-2	0.000	0.0014	negative
p1 ATG TA CIS up	0.000	0.0012	negative
p1 Modulator of Alpha-1b adrenergic receptor	0.000	0.0012	positive
p1 Agonist of H2AX	0.000	0.0012	positive
p1 Modulator of Urotensin II receptor	0.000	0.0012	negative
p1 Modulator of Adenosine A3 receptor	0.000	0.0012	negative
p0 MammMutagenicity	0.000	0.0011	positive
p0 Modulator of Serotonin 4 (5-HT4) receptor	0.000	0.0011	positive
p0 LTEA HepaRG CYP7A1 dn	0.000	0.0010	positive
p0 TOX21 HSE BLA agonist ratio	0.000	0.0009	negative
p0 BSK CASM3C VCAM1 down	0.000	0.0009	positive
p0 Bioavailability	0.000	0.0009	negative
p1 Modulator of Serotonin transporter	0.000	0.0008	positive
p1 Induce genoin human embryonic kidney cells	0.000	0.0008	negative
p0 Modulator of Alpha-1a adrenergic receptor	0.000	0.0006	negative
p1 Antagonist of the androgen receptor (AR) signaling pathway dup	0.000	0.0006	negative
p0 BSK hDFCGF IP10 down	0.000	0.0006	positive
p1 Modulator of Angiotensin-converting enzyme	0.000	0.0006	positive
p0 Modulator of Sigma opioid receptor	0.000	0.0006	positive
p1 BSK 4H MCP1 down	0.000	0.0005	positive
p0 Modulator of Vascular endothelial growth factor receptor 3	0.000	0.0004	negative
p0 BSK KF3CT TGFb1 down	0.000	0.0004	positive
p1 ATG NF kB CIS dn	0.000	0.0003	positive
p0 Modulator of Serotonin 3a (5-HT3a) receptor	0.000	0.0003	negative
p1 ATG RARa TRANS dn	0.000	0.0003	positive
p1 TOX21 p53 BLA p2 ratio	0.000	0.0002	positive
p1 Modulator of Cannabinoid CB2 receptor	0.000	0.0002	positive
p1 Cytoin HEK293 cells 32 hour	0.000	0.0002	positive

p1 Modulator of Serotonin 1a (5-HT1a) receptor	0.000	0.0001	negative
p1 Modulator of Sigma opioid receptor	0.000	0.0001	positive
p0 Modulator of P2X purinoceptor 7	0.000	0.0001	negative
p0 Modulator of TNF-alpha	0.000	0.0001	negative
p1 Antagonist of the estrogen receptor alpha (ER-alpha) signaling pathway dup	0.000	0.0001	negative
p0 ATG ISRE CIS dn	0.000	0.0000	negative
p1 Inhibitors and Substrates of Cytochrome P450 3A4	0.000	0.0000	negative

Table S2. Full name of the assays with high correlation to the ten selected bioactivity descriptors.

Descriptor Name	Assay title
AMES	Ames test for mammalian environmental mutagenicity
Caco2	Caco-2 permeability assay to investigate intestinal permeability
Inhibit CYP1A2 Activity	Inhibitors of CYP1A2 activity assay
Inhibit CYP2C19 Activity	Inhibitors of CYP2C19 activity assay
Inhibitors of Hepatocyte nuclear factor 4 (HNF4) dimerization	Inhibitors of Hepatocyte nuclear factor 4 (HNF4) dimerization assay
Modulator of Alpha-2a adrenergic receptor	Modulator of alpha-2a adrenergic receptor assay
Modulator of Alpha-2b adrenergic receptor	Modulator of alpha-2b adrenergic receptor assay
Modulator of Bradykinin B2 receptor	Modulator of bradykinin B2 receptor assay
Modulator of Monoamine oxidase A	Modulator of monoamine oxidase A assay
Modulator of Muscarinic acetylcholine receptor M4	Modulator of muscarinic acetylcholine receptor M4 assay
Modulator of P2X purinoceptor 3	Modulator of P2X purinoceptor 3 assay
Modulator of Peroxisome proliferator-activated receptor gamma	Modulator of peroxisome proliferator-activated receptor gamma assay
Modulator of Serotonin 1a (5-HT1a) receptor	Modulator of serotonin 1a (5-HT1a) receptor assay
Modulator of Serotonin 2a (5-HT2a) receptor	Modulator of serotonin 2a (5-HT2a) receptor assay
Modulators of myocardial damage	Modulators of myocardial damage assay
MammMutagenicity	Mammalian cell gene mutation assay
PGPinhibition	P-glycoprotein (Pgp) inhibition assay
ATG AP 1 CIS up	Attagene human HepG2 FBJ murine osteosarcoma viral oncogene homolog | jun proto-oncogene assay
ATG MRE CIS up	Attagene human HepG2 metal-regulatory transcription factor 1 assay
ATG PPARg TRANS up	Attagene TRANS-FACTORIAL HepG2 Human Peroxisome Proliferator-activated Receptor Gamma (PPARg) Activation Assay
ATG PXR TRANS up	Attagene human HepG2 nuclear receptor subfamily 1, group I,

	member 2 assay
ATG TA CIS up	Attagene human HepG2 unspecified assay
ATG VDRE CIS up	Attagene human HepG2 vitamin D (1,25-dihydroxyvitamin D3) receptor assay
BSK 3C MCP1 down	Bioseek human umbilical vein endothelium chemokine (C-C motif) ligand 2 assay
BSK 3C uPAR down	Bioseek human umbilical vein endothelium plasminogen activator, urokinase receptor assay
BSK 3C VCAM1 down	Bioseek human umbilical vein endothelium vascular cell adhesion molecule 1 assay
BSK 4H Pselectin down	Bioseek human umbilical vein endothelium selectin P (granule membrane protein 140kDa, antigen CD62) assay
BSK 4H SRB down	Bioseek human umbilical vein endothelium selectin P (granule membrane protein 140kDa, antigen CD62) assay
BSK 4H VCAM1 down	Bioseek human umbilical vein endothelium vascular cell adhesion molecule 1 assay
BSK hDFCGF TIMP1 down	Bioseek human foreskin fibroblast TIMP metallopeptidase inhibitor 1 assay
BSK KF3CT MCP1 down	Bioseek human keratinocytes and foreskin fibroblasts chemokine (C-C motif) ligand 2 assay
BSK KF3CT SRB down	Bioseek human keratinocytes and foreskin fibroblasts unspecified assay
BSK KF3CT TGFb1 down	Bioseek human keratinocytes and foreskin fibroblasts transforming growth factor, beta 1 assay
BSK KF3CT uPA down	Bioseek human keratinocytes and foreskin fibroblasts plasminogen activator, urokinase assay
BSK LPS SRB down	Bioseek human umbilical vein endothelium and peripheral blood mononuclear cells unspecified assay
BSK SAg MCP1 down	Bioseek human umbilical vein endothelium and peripheral blood mononuclear cells chemokine (C-C motif) ligand 2 assay
LTEA HepaRG CYP4A11 dn	LifeTech/Expression Analysis human HepaRG cytochrome P450, family 4, subfamily A, polypeptide 11 assay
LTEA HepaRG CYP4A22 dn	LifeTech/Expression Analysis human HepaRG cytochrome P450, family 4, subfamily A, polypeptide 22 assay
LTEA HepaRG DDIT3 up	LifeTech/Expression Analysis human HepaRG DNA-damage-inducible transcript 3 assay
LTEA HepaRG FMO3 dn	LifeTech/Expression Analysis human HepaRG flavin

	containing monooxygenase 3 assay
LTEA HepaRG GSTA2 dn	LifeTech/Expression Analysis human HepaRG glutathione S-transferase alpha 2 assay
LTEA HepaRG HMGCS2 dn	LifeTech/Expression Analysis human HepaRG 3-hydroxy-3-methylglutaryl-CoA synthase 2 (mitochondrial) assay

Table S3: Comparison of the Skin Doctor CP and Skin Doctor CP:Bio approaches.

	Skin Doctor CP	**Skin Doctor CP:Bio**
type of descriptors	MACCS Keys	Bioactivity descriptors
number of descriptors	166	10
n estimators	1000	500
max features	"sqrt"	"auto"
random state	43	43
number of compounds in the test set	257	257
number of compounds in the training set	1028	1021

Table S4: Results of Skin Doctor CP on the test set.

Significance level ε	Validity	Efficiency	ACC	MCC	CCR	SE	SP	NPV	PPV
0.05	0.96	0.32	0.89	0.78	0.89	0.91	0.88	0.94	0.83
0.10	0.91	0.49	0.83	0.66	0.84	0.90	0.78	0.92	0.72
0.20	0.82	0.79	0.77	0.55	0.78	0.84	0.72	0.88	0.65
0.30	0.69	0.92	0.75	0.51	0.76	0.81	0.70	0.84	0.65

C Scientific contribution

Several scientific publications and oral presentations have been originating from this thesis.

C.1 Publications

P1 **Wilm, A.**, Kühnl, J., Kirchmair, J., Computational approaches for skin sensitization prediction, *Critical Reviews in Toxicology*, **48**(9) (2018) 738–760.

P2 **Wilm, A.**, Stork, C., Bauer, C., Schepky, A., Kühnl, J., Kirchmair, J., Skin doctor: Machine learning models for skin sensitization prediction that provide estimates and indicators of prediction reliability, *International Journal of Molecular Sciences*, **20**(19) (2019) 4833.

P3 **Wilm, A.**, Norinder, U., Agea, M. I., de Bruyn Kops, C., Stork, C., Kühnl, J., Kirchmair, J., Skin doctor CP: Conformal prediction of the skin sensitization potential of small organic molecules, *Chemical Research in Toxicology*, **34**(2) (2020) 330–344.

P4 **Wilm, A.**, Garcia de Lomana, M., Stork, C., Mathai, N., Hirte, S., Norinder, U., Kühnl, J., Kirchmair, J., Predicting the skin sensitization potential of small molecules with machine learning models trained on biologically meaningful descriptors, *Pharmaceuticals*, **14**(8) (2021) 790.

C.2 Oral Presentations

O1 **Wilm, A.**, In silico prediction of skin sensitization potential and potency, Science and Technology Forum, Beiersdorf Hamburg, Oct. 2018.

O2 **Wilm, A.**, Skin Doctor: Machine Learning Models for Skin Sensitization Prediction, Science and Technology Forum, Beiersdorf Hamburg, online, Oct. 2020.

O3 **Wilm, A.**, de Lomana, M. G., Stork, C., Mathai, N., Hirte, S., Norinder, U., Kühnl, J., Kirchmair, J., Biologically meaningful descriptors for the prediction of skin sensitization potential of small molecules, OpenTox conference, online, Sep. 2021.

O4 **Wilm, A.**, de Lomana, M. G., Stork, C., Mathai, N., Hirte, S., Norinder, U., Kühnl, J., Kirchmair, J., Computational approaches for skin sensitization prediction, ÖGMBT annual meeting, online, Sep. 2021.

10. Eidesstattliche Versicherung

Declaration on oath
I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Eidesstattliche Versicherung
Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Anke Wilm
Hamburg, April 03, 2022