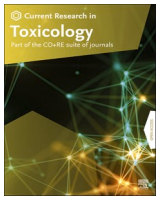
Current Research in Toxicology 7 (2024) 100183

Contents lists available at ScienceDirect

Current Research in Toxicology

journal homepage: www.journals.elsevier.com/current-research-in-toxicology



Research Paper

Low-cost quantum mechanical descriptors for data efficient skin sensitization QSAR models

Davy Guan, Raymond Lui, Slade T. Mattthews [*,1]

Computational Pharmacology & Toxicology Laboratory, Discipline of Pharmacology, Sydney Pharmacy School, Faculty of Medicine and Health, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

Keywords:
Skin sensitization
Quantum mechanical descriptors
QSAR
Machine learning

ABSTRACT

Quantitative Structure Activity Relationship modelling methodologies need to incorporate relevant mechanistic information to have high predictive performance and validity. Electrophilic reactivity is a common mechanistic feature of skin sensitization endpoints which could be concisely characterized with electronic descriptors which is key to enabling the modelling of small datasets in this domain. However, quantum mechanical methodologies have previously featured high computational costs which would exclude the use of large datasets. Consequently, we investigate the use of electronic descriptors calculated using the Hartree Fock with 3 corrections (Hf-3c) method, a low-cost *ab initio* methodology that has higher chemical accuracy than previous semiempirical methodologies for modelling *in vitro* skin sensitization assay outcomes. We also model the Ames assay as a surrogate for determining skin sensitization outcomes. The quantum chemical descriptors calculated using the Hf-3c method with conductor-like polarizable continuum model (CPCM) implicit solvation found improved QSAR model performance for the *in vitro* Ames ($n = 6049$, 0.770 AUC), KeratinoSens ($n = 164$, 0.763 AUC), and Direct Peptide Reactivity Assay ($n = 122$, 0.750 AUC) datasets, with their combination producing high predictive performance for unseen *in vivo* Local Lymph Node Assay ($n = 86$, 0.789 AUC) and Human Repeated Insult Patch Test ($n = 86$, 0.791 AUC) assay toxicant outcomes.

Introduction

Skin sensitization is a toxicological endpoint that is initiated by the covalent chemical interactions between the toxicant and predominantly nucleophilic protein moieties leading to the development of lifelong allergic contact dermatitis following dermal chemical exposure. This phenomena has been characterized as the Molecular Initiating Event (MIE) in the skin sensitization Adverse Outcome Pathway (AOP) which is defined as the rate limiting event that initiates the pathway leading to the toxicity outcome and has been previous studied using *in vitro* assays such as the Direct Peptide Reactivity Assay (DPRA) with the aim of reducing or replacing the use of animal models, most predominately the Local Lymph Node Assay (LLNA). Previous studies have found insufficient predictive performance with the use of only one assay compared to the "two of three" approach which involves the use of two *in vitro* assays modelling the MIE and Key Events further downstream together to enhance *in vivo* LLNA predictive performance (Benigni et al., 2016; Gadarowska et al., 2022; Kleinstreuer et al., 2018; Ta et al., 2021; Tung et al., 2018). The high predictive performance of the "two of three" approach has indicated the MIE and closely associated Key Events to have greater importance in the modelling of skin sensitization outcomes as including the later Key Events does not meaningfully contribute additional performance.

Quantitative Structure Activity Relationship (QSAR) models are the foremost *in silico* methodology used to model toxicological outcomes. They have been widely used to model *in vitro* and *in vivo* toxicological assays for the prediction of potential toxicity in untested compounds. This is achieved by statistically relating properties derived from the chemical structure to a biological outcome. The predictive performance of a QSAR model depends on the quantity of chemical structures, input chemical representation quality, and the selection and tuning of the algorithm used to parameterize the input representation to the toxicological outcome. All three factors are key to generating QSAR models that feature high predictive performance. The applicability domain defines the limits of a QSAR model and can be used to assess whether a model can be applied to a new compound. Establishing the applicability

* Corresponding author.
E-mail address: slade.matthews@sydney.edu.au (S.T. Mattthews).
[1] https://orcid.org/0000-0002-1652-543X.

https://doi.org/10.1016/j.crtox.2024.100183
Received 30 December 2023; Received in revised form 15 June 2024; Accepted 17 June 2024
Available online 18 June 2024

domain is not straightforward and may be established using one of several similarity measures based on molecular descriptors.

The limited data availability for skin sensitization *in vitro* assays has been challenging for the application of *in silico* methodologies to support the aim of reducing or replacing the use of *in vivo* LLNA assays (Hoffmann et al., 2018; Wang et al., 2017). This would restrict the applicability domain of any global QSAR model, which is the defined chemical space for which QSAR model predictions are reliable. Current skin sensitization models exhibit poor predictive performance for untested chemicals of toxicological concern, such as pesticides (Braeuning et al., 2018). The development of a global QSAR that is predictive for *in vivo* outcomes is important following the European ban on the use of animal assays in cosmetics. To that end, this project will implement extensive modifications to the conventional QSAR modelling methodology to address data availability limitations.

Mutagenicity features substantial mechanistic similarity with the MIE in skin sensitization toxicity outcomes by sharing similar electrophilic covalent interactions between the toxicant and biological target (Ashby et al., 1993) such as Michael addition for covalent bonding to epidermal proteins (Enoch et al., 2008) that can also be present in Ames mutagenicity (Townsend and Grayson, 2020). This is supported by the high concordance between the *in vitro* Ames mutagenicity assay and the *in vivo* LLNA skin sensitization assay outcomes (Patlewicz et al., 2010; Wolfreys and Basketter, 2005). Patlewitcz et al., further supports the relevance of reactions comprised within mutagenic effects with the inclusion of the same reaction types as the first step in their workflow summarising testing and assessment strategies for identifying indirectly acting sensitizers (Patlewicz et al., 2016). The use of Ames mutagenicity information to assist in detection of skin sensitisers is further supported by its inclusion in the OECD guidance document on reporting of defined approaches and individual information sources to be used within integrated approaches to testing and assessment (IATA) for skin sensitisation (OECD, 2016). The key difference is the orders of magnitude greater data availability of the Ames mutagenicity assay compared to *in vitro* skin sensitization assays in the AOP recognized by regulatory authorities (Hansen et al., 2009). This presents an opportunity to investigate the viability of using the abundant Ames mutagenicity data as a surrogate for the limited available *in vitro* skin sensitization data in building high performance QSAR models for predicting *in vivo* skin sensitization outcomes. This project hypothesizes a computational QSAR model of the Ames assay can reproduce the *in vitro* concordance rate while addressing the applicability domain limitations with current QSAR models of skin sensitization assays.

While previous work found chemical reactivity to be the key determinant in the solicitation of the immune response (Enoch and Roberts, 2013), the use of *in silico* methodologies to mechanistically study chemical reactivity from first principles have been limited due to the high computational costs, low throughput, and limited chemical applicability. Quantum mechanics is the only theory that could explain the inherent reactivity of a potential toxicant which could be used to derive relevant descriptors. This approach would produce a chemical representation that would feature relatively few dimensions as the chemical reactivity is directly quantified which reduces model complexity to improve generalizability. This project aims to investigate chemical representations derived from quantum mechanical calculations.

Quantum mechanical descriptors readily quantify the reactivity of a molecule which is a determinant of the mechanisms of toxicity for both skin sensitization and genotoxicity (Chipinda et al., 2011). These descriptors are calculated from the electronic properties of a ground state molecular structure using *ab initio* computational chemistry methodologies. Quantum mechanical descriptors derived from the molecular orbitals such as the Lowest Unoccupied Molecular Orbital (LUMO) energy and polarization have seen extensive use in previous toxicological QSAR studies (Can et al., 2013; Kostal and Voutchkova-Kostal, 2016; Pandith et al., 2010). The level of theory, basis set, and input structure used in computational chemistry methodologies determines the precision of the calculated quantum mechanical descriptor.

While quantum mechanical descriptors have shown high predictive power in past studies, their use has been restricted to small datasets owing to their high computational cost. Many previous studies have used rudimentary semiempirical or Hartree Fock methodology implementations with small basis sets to reduce computational cost at the expense of a restricted chemical space and descriptor robustness (Can et al., 2013; Puzyn et al., 2008). The development of a quantum mechanical descriptor set that could reliably be used for a diverse chemical space is desirable for constructing robust QSAR models of toxicological datasets as it directly models the covalent interactions that form the mechanistic basis of skin sensitization. This approach might mitigate the problem of predicting outcomes for molecules outside the applicability domain. Previous implementations of these descriptors utilize the molecular structure optimized in the gas phase without implicit or explicit solvation to avoid the computational cost of incorporating solvation. However, this reduces the biological relevance of the calculated quantum mechanical descriptors which could in turn reduce the resulting model performance and reliability. Previous studies using quantum mechanical descriptors in toxicological QSAR models feature narrow applicability due to the computational cost of molecular characterization using quantum methodologies.

Contemporary computational chemistry has developed sufficiently to address previous limitations of using quantum mechanical descriptors. These developments include the “low-cost” Hartree Fock with three corrections (Hf-3c) method that incorporates dispersion corrections to account for van Der Waals forces, and the use of an optimized basis set that can be used to calculate quantum mechanical descriptors across the periodic table with lower computational cost compared to the conventional Hartree Fock method (Sure and Grimme, 2013). Additionally, solvation could be readily incorporated using implicit solvation models which simulate the bulk polarization effects of solvation while omitting the explicit inclusion of solvent molecules to reduce computational cost. High concordance with experimental results has been achieved with implicit solvation models for simulating condensed phase chemical milieus (Barone and Cossi, 1998; Marenich et al., 2009; Silva et al., 2016). The incorporation of these developments into a quantum mechanical descriptor implementation is hypothesized to enhance model reliability and prediction performance and simplify model mechanistic interpretation in the toxicological domain.

This project implements quantum mechanical descriptors using the Hf-3c method with and without aqueous CPCM implicit solvation for modelling skin sensitization and Ames mutagenicity datasets. These predictive models are evaluated on *in vivo* LLNA outcomes compared to conventional descriptors, existing skin sensitization QSAR models, and the *in vitro* Ames assay. The applicability of using the Ames assay as a surrogate skin sensitization QSAR model data source will also be investigated and compared with *in vitro* Ames assay results. It is necessary to remove manual model optimization to avoid investigator bias since this project assesses and compares quantum mechanical and conventional descriptors through QSAR model performance. To that end, an automated machine learning methodology using the TPOT library was adopted to conduct all data transformation and scaling, machine learning algorithm selection, and hyperparameter tuning to produce a pipeline for each dataset and descriptor variant. Lastly, the highest performing *in silico* model produced in this project will be compared to the *in vivo* Local Lymph Node Assay in their prediction of *in vivo* Human Repeat Insult Patch Test (HRIPT) outcomes to characterize *in silico* skin sensitization QSAR models in the wider context and to guide future development efforts.

Materials and methods

Dataset preparation

The following datasets were obtained for each toxicity outcome:

1. **Ames Mutagenicity.** The Hansen Ames Mutagenicity Benchmark dataset (Hansen et al., 2009) was selected with all 6512 chemicals and binary Ames assay outcomes. This dataset comprised results labeled as four sources, CCRIS (2542), VITIC (1197), EPA (2747), and 26 from GENETOX. In the Hansen paper they assert that Ames data are generally affected by an error rate of around 15 % in terms of interlaboratory reproducibility but that the CCIRS data is more likely to be only around 11 % in accordance with a study by Kazius et al. (2005). A recent study by Li et al., examined the Hansen dataset comparing it with the training data for the second Ames/QSAR international challenge from the Division of Genetics and Mutagenesis, National Institute of Health Sciences, Japan (Li et al., 2023). They found 175 overlapping molecules in the Hansen dataset and of these 25 had discordant results. This roughly corresponds to the 85 % reproducibility rate observed in laboratory Ames data (Kamber et al., 2009).
2. **KeratinoSens (KRS).** A 164 chemical dataset with binary KeratinoSens assay results was extracted from the Skin Sensitization Database (SkinSensDB) published on 19 October 2017 (Wang et al., 2017).
3. **Direct Peptide Reactivity Assay (DPRA).** 122 chemicals from the Cosmetics Europe database were selected with categorical DPRA outcomes and assigned mechanistic domains (Hoffmann et al., 2018). *In vivo* Human Repeat Insult Patch Test (HRIPT) and corresponding LLNA outcomes were also extracted from this database. The six HRIPT outcome categories were quantized to binary qualitative (positive, negative) and ternary ordinal (none, weak, strong) skin sensitization categories.
4. **Skin Sensitization (SKS).** A 92 chemical dataset was composed with structures, mutagenicity, and LLNA skin sensitization toxicity outcomes from Patlewicz et al. (2010) and held out as an external validation dataset.

Dataset curation

All SMILES structures of each dataset were curated with the removal of salts and solvents, neutralization of charges and the addition of explicit hydrogens using Standardizer 18.22.0, 2019, ChemAxon (https://www.chemaxon.com/). A KNIME (Berthold et al., 2008) workflow was used to convert all two-dimensional structures to 3D with initial 3D structures generated by OpenBabel (O'Boyle et al., 2011) and optimized with the Universal Force Field for 50,000 steps with the RDKit nodes (Landrum). OpenBabel was then used to convert the SDF file to MOL2 format files for quantum chemical descriptor calculation.

Conventional descriptor calculation

The Mordred 1.1.2 Python library (Moriwaki et al., 2018) was used to calculate 1,825 molecular descriptors from the 3D structures for each dataset. Feature selection was conducted using the CfsSubsetEval filter implemented in the Weka 3.9.3 machine learning program (Witten et al., 2016).

Quantum mechanical descriptor calculation

A performance optimized fork of the MaPhi descriptor package was implemented with Cython (Behnel et al., 2011) and parallelization (Moritz et al., 2018). This modified MaPhi package was configured to further optimize the geometry of each structure at the PM7 level of theory using the MOPAC2016 semiempirical chemical program (Stewart, 2016) before further calculation. 21 quantum chemical descriptors were then calculated for each dataset at the Hf-3c level of theory (Sure and Grimme, 2013) either in vacuum or with aqueous implicit solvation using the Conductor-like Polarizable Continuum Model (Barone and Cossi, 1998) within the Orca 4.2.0 computational chemistry program (Neese, 2012, 2018). This MaPhi fork has been made available at https://github.com/IamDavyG/FasterMaPhi.

Applicability domain analysis

The applicability domain of the Ames, KeratinoSens, and DPRA modelling datasets were assessed against the SKS dataset that was to be used for external validation. This project selected the distance from centroid (dist. centroid), leverage, fixed and variable kNN measures implemented in the Applicability Domain Toolbox (Sahigara et al., 2014; Sahigara et al., 2012) in MATLAB 2019b (The MathWorks, 2019) to quantify the applicability domain between each corresponding dataset variant. An additional unified AD measure consisting of only counting chemicals that were within all four measures was implemented to provide an aggregated and conservative AD estimate (Eq. (1)).

$$\textit{For each dataset descriptor combination}: \textit{Unified AD} = AD_{\textit{Dist. Centroid}} \bigcap AD_{\textit{Leverage}} \bigcap AD_{\textit{Fixed kNN}} \bigcap AD_{\textit{Variable kNN}} \quad (1)$$

Machine learning methodology

The TPOT Python library (Le et al., 2019; Olson et al., 2016a; Olson et al., 2016b) automated the optimization of toxicological models using Darwinian evolutionary theory for each dataset. This was achieved by the generation of an initial population of 100 model configurations that differed in dataset preprocessing and machine learning algorithm. All model configurations for each dataset were trained and scored using ROC AUC from 10-fold cross validation with the worse performing half of the population removed. Additional models are generated with crossover, where models combine configurations, mutation, where model configuration parameters randomly change, and reinitialized from scratch, up to the original 100 model configuration population. This concludes and increments a single generation, with further model training repeating this process. The optimization process is carried out for 10 generations.

QSAR model combination

Three QSAR model combinations were composed by averaging the prediction probabilities output from each individual QSAR generated using Hf-3c CPCM descriptors. These three variants consisted of averaging: all three QSAR models, only the DPRA and KeratinoSens QSARs to replicate the literature, and only the Ames mutagenicity and KeratinoSens QSAR model outputs.

Additional performance enhancement

The unified AD measure (Equation (1) was used to compose SKS external validation dataset subsets that only include predictions for molecules considered to be within the applicability domain of each combined QSAR model described above. Threshold values for assigning the binary positive/negative class on the model output probability values were also optimized and compared to the naïve 0.5 threshold. These values were manually optimized for balanced accuracy by selecting threshold values between 0.35 and 0.65 in 0.01 increments.

CAESAR positive control

The CAESAR Skin Sensitization QSAR model version 2.1.6 (Chaudhry et al., 2010) in the VEGA *in silico* platform version 1.1.5 evaluated the SKS external validation dataset to provide a baseline for model performance comparison. The integrated Applicability Domain Index was used with a 0.8 threshold value to select a subset of SKS external validation dataset predictions that were considered within the applicability domain of this model.

Table 1
Summary of the datasets and the total number of molecules before and after descriptor calculation.

		Chemical Representations (% Total)		
Dataset	n	Hf-3c vacuum	Hf-3c CPCM	Mordred PM7
Hansen Ames	6512	6361 (97.7 %)	6049 (92.9 %)	6450 (99.0 %)
SSDB KeratinoSens	164	161 (98.2 %)	160 (97.6 %)	161 (98.2 %)
CosEU DPRA	122	110 (90.2 %)	109 (89.3 %)	111 (91.0 %)
ExtVal SKS	92	87 (94.6 %)	86 (93.5 %)	88 (95.7 %)

Permutation feature importance analysis for model interpretation

Permutation feature importance analysis assessed the contribution of each Hf-3c descriptor, both vacuum and solvated variants for each *in vitro* dataset, in predicting the external validation dataset. The eli5 0.10.1 Python library was used to iteratively shuffle each descriptor column of the external validation molecules which were predicted again; the change in external validation ROC AUC signified the relative importance weight of that shuffled descriptor. This process was repeated 20 times for each *in vitro* dataset and quantum mechanical descriptor pair, with the mean weight of each descriptor used to rank their relative contributions.

Binary to ternary QSAR model output conversion

The probability output of the Hansen Ames, KeratinoSens, and DPRA Hf-3c CPCM QSAR models was arithmetically averaged to generate combined model predictions for predicting HRIPT outcomes before two thresholds were chosen to enable the current models to produce a ternary output. They consisted of a non-sensitizing classification if the averaged probability was below 0.33, weak sensitizers between 0.33 and 0.66, and strongly sensitizing if 0.66 probability was exceeded.

Results

Outline

This study assessed the feasibility of quantum mechanical descriptors calculated using the Hf-3c level of theory compared to conventional molecular modelling-derived descriptors as chemical representations in QSAR models for two small skin sensitization assay datasets and one large Ames mutagenicity assay dataset. This consisted of comparing the proportion of each dataset for which quantum mechanical descriptors could be calculated to conventional descriptors and visualizing the resulting chemical space. The predictive concordance between the LLNA outcomes in the SKS external validation dataset and the Ames mutagenicity, KRS, and DPRA skin sensitization QSAR model predictions were characterized and compared to CAESAR, an established QSAR model known to be predictive for skin sensitization outcomes. A comparison between the current *in silico* models and the *in vivo* LLNA assay for predicting *in vivo* HRIPT outcomes is also presented.

Exploratory data analysis

Both Hf-3c quantum mechanical representations in vacuum and CPCM solvation feature a lower yield of molecules for which descriptors could be calculated compared to conventional Mordred descriptors (Table 1). This difference is most substantial in the Hansen Ames dataset with the greatest difference of 451 between the Hf-3c CPCM and

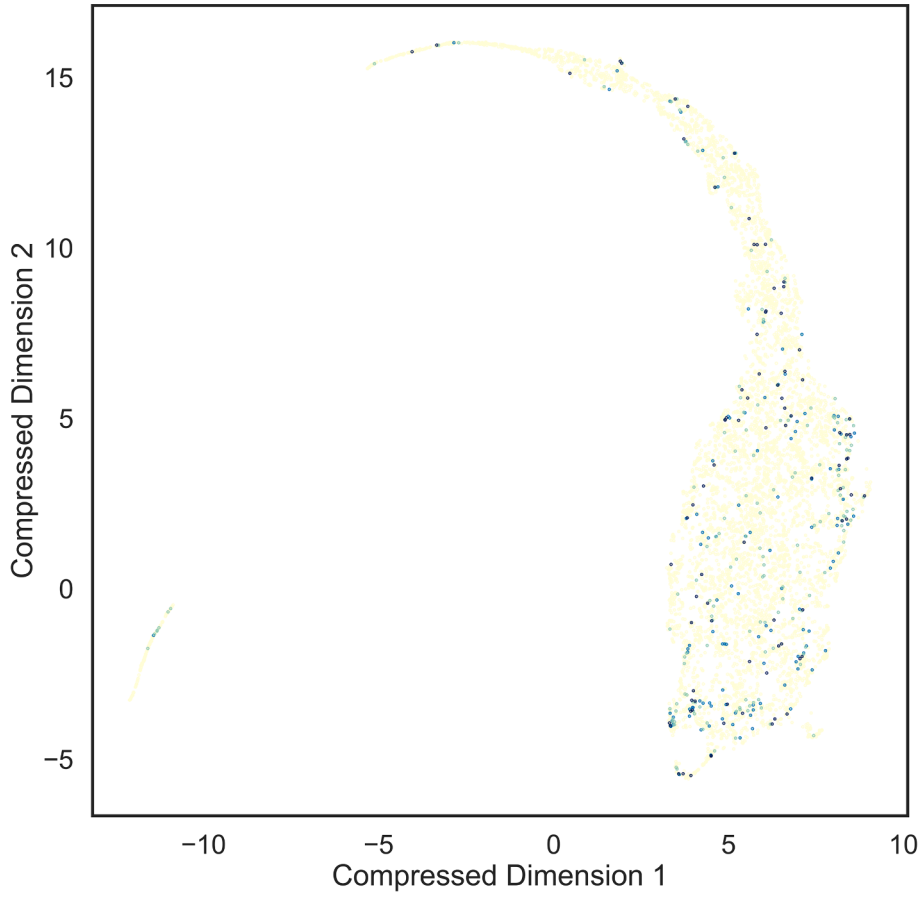


Fig. 1. Chemical space of the Ames (yellow), DPRA (teal), KeratinoSens (green), and SKS (navy) datasets with Hf-3c descriptors calculated in vacuum visualized using UMAP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

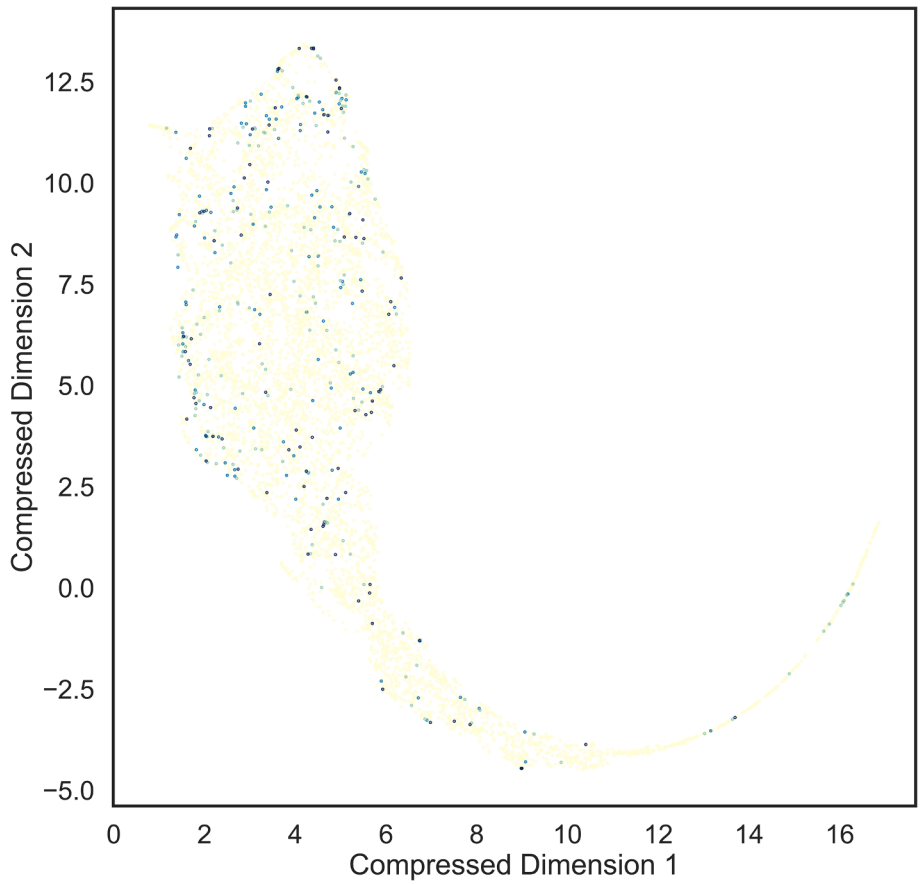


Fig. 2. Chemical space of the Ames (yellow), DPRA (teal), KeratinoSens (green), and SKS (navy) datasets with Hf-3c descriptors with CPCM implicit solvation visualized using UMAP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Mordred datasets. The other datasets show a smaller difference with Hf-3c vacuum descriptors yielding one molecule less than Mordred descriptors in the DPRA and SKS datasets while having the same yield in the KeratinoSens dataset. The Hf-3c CPCM descriptors feature one less molecule compared to the Hf-3c vacuum descriptors on the KeratinoSens, DPRA, and SKS datasets.

Chemical space visualization

The chemical space of each descriptor type was visualized by transforming the descriptors from each dataset to two dimensions using UMAP analysis. The chemical space of the Hansen Ames dataset is substantially larger than those of all the other datasets in this project for all descriptor types (Figs. 1, 2, and 3). The chemical space of this dataset also covers all the datasets indicating all the other datasets are qualitatively within the applicability domain of models constructed from this dataset. The KeratinoSens, DPRA, and SKS datasets are dispersed throughout the chemical space with many singletons that do not feature chemicals from other datasets nearby for all descriptor types.

Automated model training results

Most models achieved similar internal validation performance across descriptor types for each dataset (Table 2). The Hf-3c CPCM descriptor type had lower model performance for the Ames and DPRA datasets and the best model performance in the KeratinoSens dataset compared to Hf-3c vacuum descriptors. Models based on Mordred descriptors achieved the best training performance in the remaining Ames and DPRA datasets.

External validation performance

The performance of both the Hf-3c quantum chemical descriptor variants and the DPRA Mordred descriptor models in classifying the LLNA outcomes in the SKS external validation dataset was better than random (ROC AUC 95 % CI > 0.5 (Table 3)) and visualized in Fig. 6. Hf-3c descriptors calculated with CPCM implicit solvation achieved the best SKS external validation classification performance in the models using the Hansen Ames (0.744 ROC AUC) and DPRA (0.705 ROC AUC) datasets (Table 3). A statistically significant difference in model performance between the Hf-3c CPCM and Mordred descriptors was found in the Ames dataset (bold values in Table 3) and visualized in Fig. 4. Models using Hf-3c descriptors calculated in vacuum showed inconsistent performance with the best external validation performance using the KeratinoSens dataset (0.701 ROC AUC), comparable to the Hf-3c CPCM descriptors in Hansen Ames (0.730 ROC AUC) and the DPRA datasets (0.688 ROC AUC). Mordred descriptors ranked last in external validation performance in the Ames (0.509 ROC AUC) and KeratinoSens (0.530 ROC AUC) datasets with lower performance than the Vega CAESAR model (0.633 ROC AUC) and shown in Figs. 4 and 5.

Both Hf-3c descriptor variants produced models that outperformed models using Mordred descriptors trained on the Hansen Ames and KeratinoSens datasets in terms of balanced accuracy at a naïve 0.5 probability threshold (Table 4). The model using Mordred descriptors and the DPRA dataset found the higher performance at a naïve 0.5 probability threshold (Table 4). Selecting molecule predictions that were considered within the applicability domain of each model generally increased performance by between 1 and 2 % balanced accuracy. There were two exceptions to this trend where only including molecules inside the applicability domain reduced performance with the Hf-3c

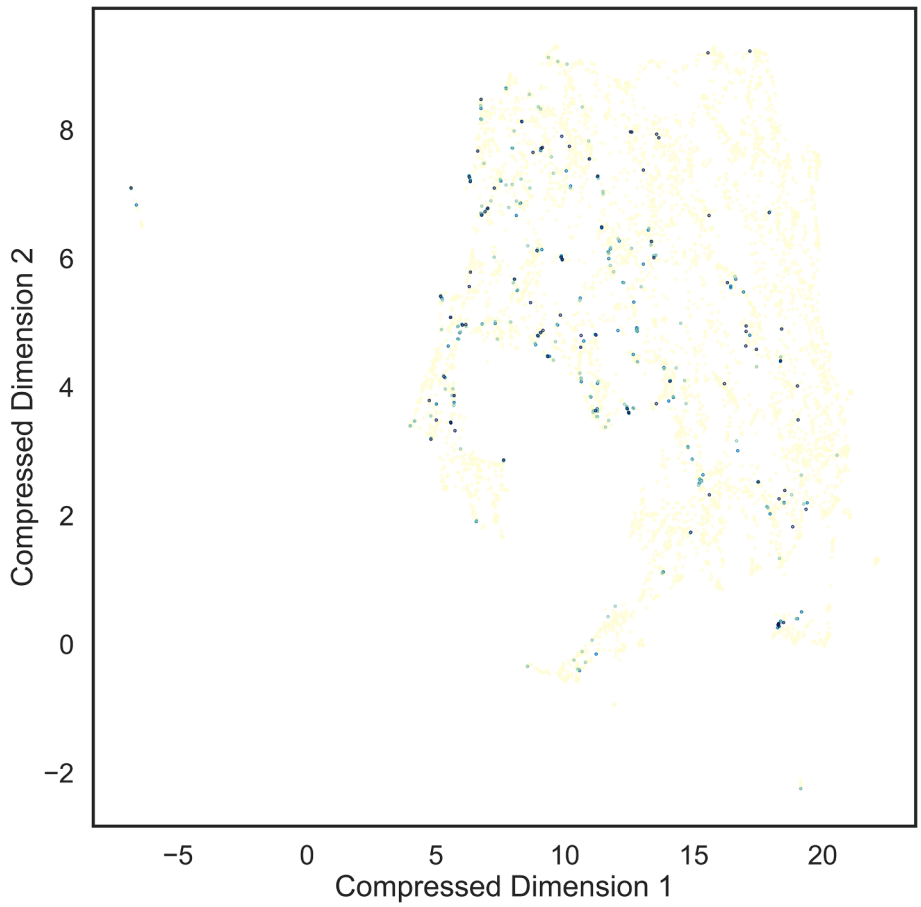


Fig. 3. Chemical space of the Ames (yellow), DPRA (teal), KeratinoSens (green), and SKS (navy) datasets with Mordred descriptors visualized using UMAP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
10-fold cross validation model training performance (ROC AUC) for TPOT pipelines of each dataset and descriptor type.

	Chemical Representations		
Dataset	Hf-3c vacuum	Hf-3c CPCM	Mordred
Hansen Ames	0.858	0.770	0.893
SSDB KeratinoSens	0.760	0.763	0.716
CosEU DPRA	0.762	0.750	0.793

Table 3
External validation classification performance (ROC AUC [95 % CI], **bold = non-overlapping CI) of each model for predicting the SKS dataset.**

	Chemical Representations			CAESAR
Dataset	Hf-3c vacuum	Hf-3c CPCM	Mordred	
Ames	0.730 [0.626, 0.834]	**0.744 [0.642, 0.846]**	**0.509 [0.389, 0.629]**	–
KRS	0.701 [0.592, 0.810]	0.683 [0.572, 0.794]	0.530 [0.409, 0.651]	–
DPRA	0.688 [0.578, 0.798]	0.705 [0.597, 0.813]	0.664 [0.550, 0.778]	–
CAESAR	–	–	–	0.633 [0.504, 0.762]
Mean AUC	0.706	0.711	0.568	–

vacuum descriptor set on KeratinoSens dataset reducing balanced accuracy by 0.5 % and the Mordred DPRA model that lost 6.2 % balanced accuracy (Table 4).

Applicability domain analysis

All descriptor types found high applicability domain scores for all three modelling datasets compared to the SKS external validation dataset (Tables 5, 6, and 7). There was a trend with reduced scores for smaller datasets (Table 6 and 7) compared to the Hansen Ames dataset (Table 5). The KeratinoSens dataset represented with Hf-3c vacuum descriptor set is an exception to this trend with a higher Unified AD Measure score (Table 6) than the corresponding Hansen Ames variant (Table 5) owing to the identification of the same molecules deemed to be out of the applicability domain across different measures.

Combined model performance

The combination of multiple models using the average of their output probabilities improved external validation performance over the performance of the individual model constituents in ROC AUC point estimates (Table 8, Fig. 7). Consideration of the applicability domain further improved combined model performance for DPRA and KeratinoSens and the “all” model combinations and the Ames and KeratinoSens models (Table 8, Fig. 8). The Vega CAESAR model found a statistically significant difference compared to the *in vitro* Ames assay for predicting LLNA outcomes for the entire SKS dataset. CAESAR model performance improved when the applicability domain was taken into consideration. However, this performance was still lower than the point

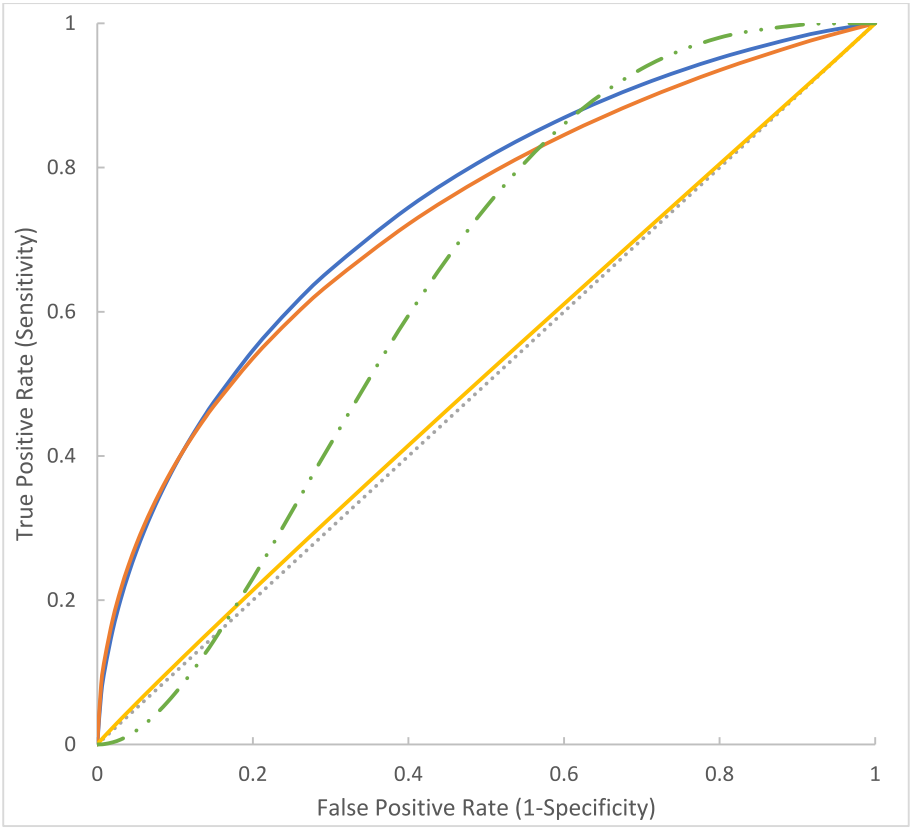


Fig. 4. ROC curves for models based on the Hansen Ames dataset using Hf-3c vacuum (orange), Hf-3c CPCM (blue), or Mordred descriptors (yellow) and the Vega CAESAR model (green dotted) predicting the SKS external validation dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

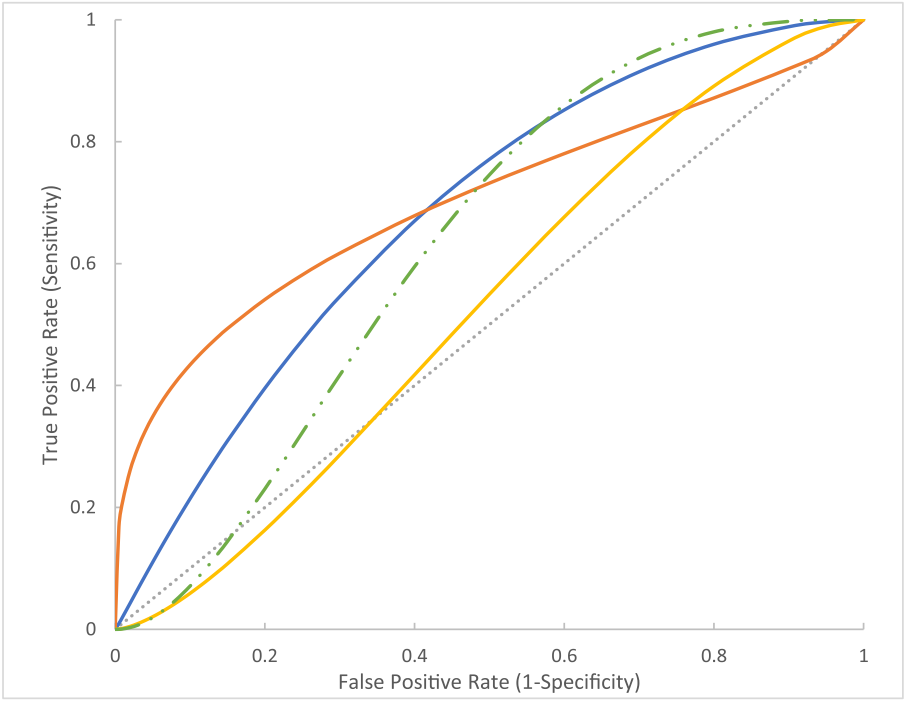


Fig. 5. ROC curves for models based on the KeratinoSens dataset using Hf-3c vacuum (orange), Hf-3c CPCM (blue), or Mordred descriptors (yellow) and the Vega CAESAR model (green dotted) predicting the SKS external validation dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

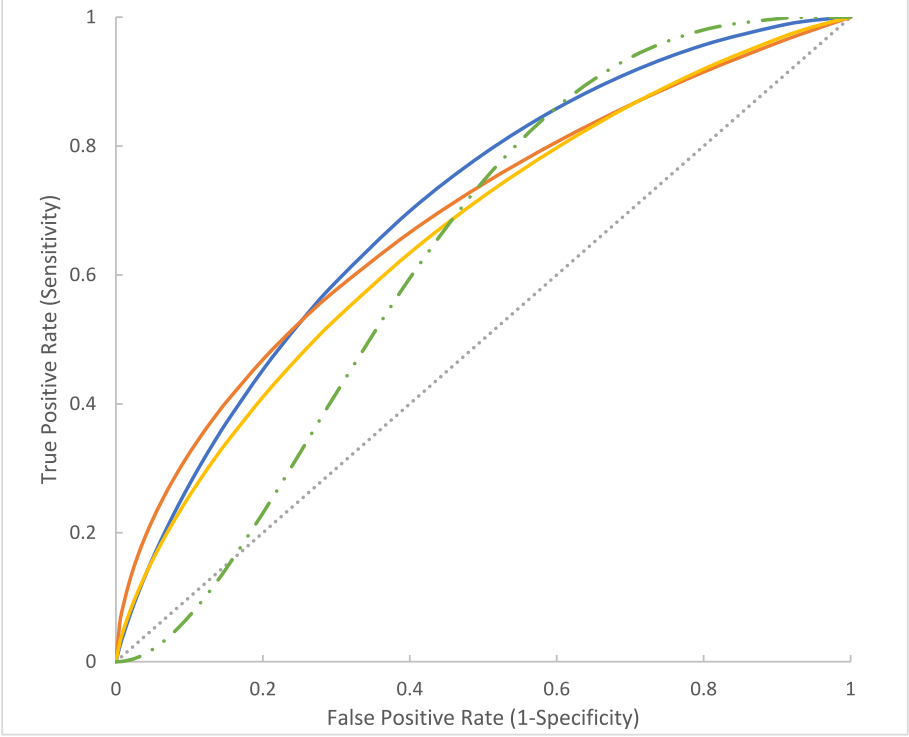


Fig. 6. ROC curves for models based on the DPRA dataset using Hf-3c vacuum (orange), Hf-3c CPCM (blue), or Mordred descriptors (yellow) and the Vega CAESAR model (green dotted) predicting the SKS external validation dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Performance statistics (%) for predicting LLNA outcomes in the SKS external validation dataset for each model of the Hansen, KeratinoSens, and DPRA dataset represented using conventional Mordred, or Hf-3c quantum mechanical descriptors in vacuum or with aqueous CPCM solvation.

Dataset	Representation	Variant	Sensitivity	Specificity	Balanced Accuracy
Ames	Hf-3c CPCM	Global	60.0	82.9	71.5
		Inside AD	61.5	82.9	72.2
	Hf-3c Vacuum	Global	48.9	85.4	67.1
		Inside AD	51.2	86.5	68.8
	Mordred	Global	62.2	29.3	45.7
		Inside AD	62.2	30.0	46.1
KRS	Hf-3c CPCM	Global	93.3	24.4	58.9
		Inside AD	94.6	24.4	59.5
	Hf-3c Vacuum	Global	84.4	24.4	54.4
		Inside AD	84.1	23.7	53.9
	Mordred	Global	64.4	39.0	51.7
		Inside AD	66.7	40.6	53.6
DPRA	Hf-3c CPCM	Global	71.1	43.9	57.5
		Inside AD	72.2	45.0	58.6
	Hf-3c Vacuum	Global	71.1	41.5	56.3
		Inside AD	70.7	45.5	58.1
	Mordred	Global	80.0	51.2	65.6
		Inside AD	79.5	39.3	59.4

Table 5
Quantitative applicability domain comparison for the Hansen Ames dataset represented using Hf-3c vacuum, Hf-3c CPCM, and Mordred descriptors compared to the corresponding SKS external validation dataset variant.

AD Measure	Hf-3c vacuum (%)	Hf-3c CPCM (%)	Mordred (%)
Dist. Centroid	94.2	94.2	100
Leverage	98.8	98.8	98.8
Fixed kNN	98.8	98.8	100
Variable kNN	95.3	96.5	100
Unified AD Measure	89.5 (n = 77)	91.9 (n = 80)	98.8 (n = 85)

Table 6
Quantitative applicability domain comparison for the KeratinoSens dataset represented using Hf-3c vacuum, Hf-3c CPCM, and Mordred descriptors compared to the corresponding SKS external validation dataset variant.

AD Measure	Hf-3c vacuum (%)	Hf-3c CPCM (%)	Mordred (%)
Dist. Centroid	95.3	95.3	98.8
Leverage	96.5	95.3	86.0
Fixed kNN	97.7	96.5	97.7
Variable kNN	97.7	96.5	95.3
Unified AD Measure	94.2 (n = 81)	90.7 (n = 78)	84.9 (n = 73)

Table 7
Quantitative applicability domain comparison for the DPRA dataset represented using Hf-3c vacuum, Hf-3c CPCM, and Mordred descriptors compared to the corresponding SKS external validation dataset variant.

AD Measure	Hf-3c vacuum (%)	Hf-3c CPCM (%)	Mordred (%)
Dist. Centroid	95.3	96.5	96.5
Leverage	93.0	93.0	83.7
Fixed kNN	96.5	97.7	94.2
Variable kNN	89.5	94.2	84.9
Unified AD Measure	83.7 (n = 72)	88.4 (n = 76)	76.7 (n = 66)

Table 8
External validation performance (ROC AUC [95 % CI], **bold = non-overlapping CI**) for Hf-3c CPCM models combined by averaging the prediction probabilities for all datasets, the DPRA and KeratinoSens, or Ames and KeratinoSens datasets, and the Vega CAESAR model for the entire SKS dataset (Global) or only within the applicability domain (Inside AD) alongside *in vitro* Ames assay performance.

Model Combination	Global	Inside AD	In vitro Ames
All (Hf-3c CPCM)	0.789 [0.685, 0.893]	0.806 [0.712, 0.900]	–
DPRA + KRS (Hf-3c CPCM)	0.760 [0.660, 0.860]	0.764 [0.657, 0.871]	–
Ames + KRS (Hf-3c CPCM)	0.770 [0.672, 0.868]	0.777 [0.674, 0.880]	–
Vega CAESAR	**0.633 [0.504, 0.762]**	0.669 [0.481, 0.857]	–
In vitro Ames	–	–	**0.828**

estimate of any combined model variants (Table 8, Figs. 7 and 8). The *in vitro* Ames assay was the most predictive for LLNA outcomes of the SKS external validation dataset with a ROC AUC point estimate of 0.828 (Table 8, Figs. 7 and 8).

The external validation balanced accuracy results of all models did not follow the rank order demonstrated in the ROC AUC results (Table 8) when predicting on all SKS molecules (Table 9). Most combined models and the Vega CAESAR model found performance improvements after molecules were selected within the applicability domain; however, the effect was greater in Vega CAESAR with an 8.9 % increase in balanced accuracy compared to up to 3.3 % for the combined models (Table 9). This resulted in the Vega CAESAR model achieving 71.9 % balanced accuracy at the expense of only classifying 47 molecules compared to at least 74 molecules being within the applicability domain of the combined models (Table 9). The single exception was the Ames and KeratinoSens combined model which found a performance decrease of −1.0 % balanced accuracy after selecting molecules inside the applicability domain (Table 9). Threshold tuning improved performance for all combined models with a larger effect than considering the applicability domain only, with up to 11.7 % balanced accuracy improvement observed for the DPRA and KeratinoSens combined model (Table 9). The *in vitro* Ames assay featured the highest balanced accuracy at 82.8 % followed with the applicability domain-controlled combination of all Hf-3c models with threshold tuning at 74.6 % balanced accuracy.

LLNA prediction permutation feature importance analysis

The three highest ranked Hf-3c CPCM descriptors (hardness, LUMO, and HOMO) are conserved between the DPRA and KeratinoSens datasets (Table 10). Some of these descriptors also ranked in the top three when calculated at the Hf-3c vacuum level of theory in the Ames and KeratinoSens datasets sharing hardness. The Hf-3c vacuum variant of the LUMO molecular orbital descriptor was present in the DPRA dataset with the KeratinoSens dataset featuring HOMO in the top three ranking descriptors. The Ames dataset for both descriptor variants share HOMO-1 molecular orbital and are largely distinct from the DPRA and KeratinoSens datasets.

Human repeated Insult Patch Test (HRIPT) prediction performance

The *in silico* Ames and combined Hf-3c models found a statistically significant difference in ROC AUC performance when predicting HRIPT

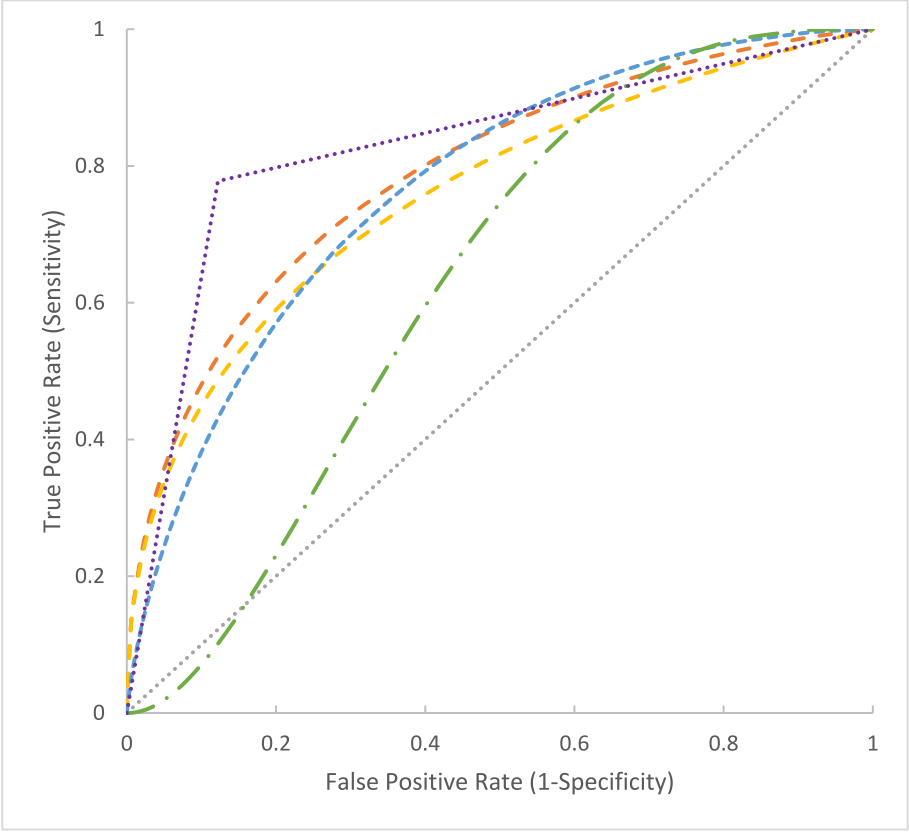


Fig. 7. ROC curves depicting external validation performance of the combined models using Hf-3c descriptors averaging model prediction probabilities for all models (orange striped), the combined DPRA and KeratinoSens models (yellow striped), and the combined Ames and KeratinoSens models (blue striped). The Vega CAESAR model (green dotted) and *in vitro* Ames assay (purple striped) for classifying the entire SKS dataset are also shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

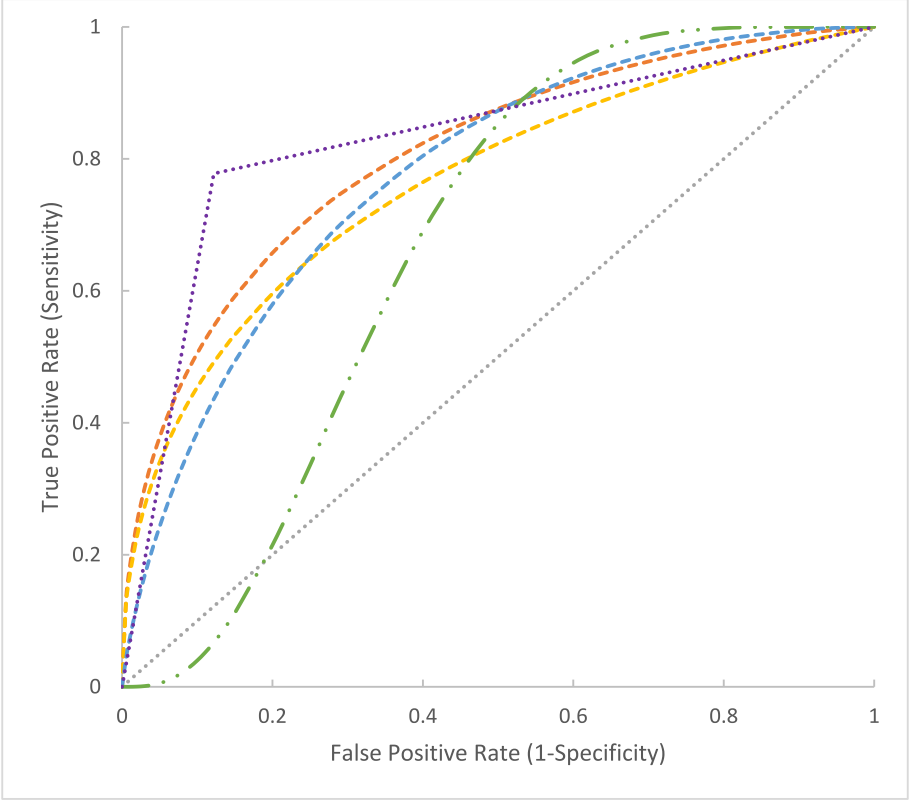


Fig. 8. ROC curves depicting applicability domain-controlled external validation performance of the combined models using Hf-3c CPCM descriptors averaging model prediction probabilities for all models (orange striped), the DPRA and KeratinoSens models (yellow striped), or the Ames and KeratinoSens models (blue striped), followed by the Vega CAESAR model (green dotted) and *in vitro* Ames assay (purple striped) for classifying a subset of the SKS dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 9
External validation performance statistics (%) for the combined Hf-3c CPCM models averaging the prediction probabilities of all models, DPRA and KeratinoSens models, and Ames and KeratinoSens models for predicting all 86 SKS molecules (Global), a subset within the applicability domain (Inside AD), and with threshold probability tuning compared to the Vega CAESAR model and *in vitro* Ames assay.

Model	Variant	Sensitivity	Specificity	Balanced Accuracy
All (Hf-3c CPCM)	Global	84.4	51.2	67.8
	Inside AD (n = 74)	85.3	52.5	68.9
	T = 0.59 Inside AD	61.8	87.5	74.6
DPRA + KRS (Hf-3c CPCM)	Global	91.1	24.4	57.8
	Inside AD (n = 76)	91.7	25.0	58.3
	T = 0.61 Inside AD	75.0	65.0	70.0
Ames + KRS (Hf-3c CPCM)	Global	73.3	65.9	69.6
	Inside AD (n = 76)	71.4	65.9	68.6
	T = 0.54 Inside AD	60.0	78.0	69.0
Vega CAESAR	Global	95.9	30.0	63.0
	Inside AD (n = 47)	96.2	47.6	71.9
In vitro Ames	–	77.8	87.8	82.8

compared to *in vivo* LLNA outcomes (Table 11), with the Combined model finding the best binary skin sensitization prediction performance (Table 11, Fig. 9). The *in vivo* LLNA is more predictive for ternary outcomes compared to the combined model (Table 12).

Discussion

This study examined the use of quantum mechanical descriptors to improve skin sensitization and mutagenicity QSAR model performance for the prediction of *in vivo* LLNA outcomes for a set of unseen chemicals.

QSAR models based on the Hansen Ames mutagenicity benchmark dataset demonstrated high predictive performance for predicting *in vivo* skin sensitization outcomes when used in conjunction with Hf-3c quantum mechanical descriptors calculated in vacuum or with CPCM solvation. Neither Hf-3c descriptor variant could produce QSAR models with statistically significant predictive performance (Table 3) compared to the actual *in vitro* Ames mutagenicity assay results (Table 8). However, the Hf-3c descriptor variant with aqueous solvation showed a statistically significant difference compared to conventional descriptors (Table 3). Although beyond the scope of the present study, we also found 0.843 ROC AUC for the Hf-3c with CPCM solvation Ames QSAR model when predicting the Ames mutagenicity outcomes included within the SKS external validation dataset. This is close to the 84 % interlaboratory reproducibility of the *in vitro* Ames mutagenicity assay itself (Benigni and Bossa, 2011; Piegorsch and Zeiger, 1991).

The pure *in silico* "2 of 3" implementation in this project, combining the DPRA and KeratinoSens QSAR models, found similar performance to the "2 of 3" defined approach implemented using *in vitro* assay results from the literature (Kleinstreuer et al., 2018). Together, these three findings show QSAR models can become computational facsimiles of *in vitro* assays when given abundant data, appropriate features that

Table 10
Permutation feature importance rankings for both Hf-3c descriptor variants calculated in vacuum or with CPCM solvation for predicting the external validation dataset, higher rank (darker greeen and lower value i.e. from 1st rank down) indicates greater relative contribution to LLNA prediction. Quantum mechanical descriptor families include reactivity (*REACT*), molecular orbital (*MO*), polarization (*POLAR*), and total energy (*ENERGY*).

		Hf-3c CPCM Rank			*Hf-3c Vacuum Rank*		
Family	QM Descriptor	Ames	KRS	DPRA	Ames	KRS	DPRA
REACT	**Electron_accepting_power [Eh]**	**1**	7	11	15	21	20
MO	**LUMO+1 [Eh]**	**2**	8	13	16	6	15
MO	**HOMO-1 [Eh]**	**3**	5	21	**1**	10	17
REACT	**Hardness [Eh]**	4	**1**	**2**	**3**	**1**	11
POLAR	**quadrupole XX [Debye]**	5	11	18	14	11	5
MO	**LUMO [Eh]**	6	**3**	**1**	7	4	**1**
MO	**LUMO+2 [Eh]**	7	10	6	8	8	18
POLAR	**IsotropicQuad**	8	13	16	12	19	**2**
POLAR	**quadrupole YY [Debye]**	9	20	19	9	5	6
POLAR	**quadrupole ZZ [Debye]**	10	12	17	5	**2**	4
POLAR	**quadrupole XZ [Debye]**	11	21	12	10	18	9
REACT	**Electronegativity [Eh]**	12	14	4	4	12	7
REACT	**Net_electrophilicity [Eh]**	13	16	8	20	14	14
REACT	**Electrophilicity [Eh]**	14	6	20	17	20	10
POLAR	**quadrupole YZ [Debye]**	15	17	10	13	15	8
REACT	**Electron_donating_power [Eh]**	16	15	15	11	17	12
MO	**HOMO [Eh]**	17	**2**	**3**	6	**3**	13
MO	**HOMO-2 [Eh]**	18	4	5	**2**	7	21
POLAR	**quadrupole XY [Debye]**	19	19	9	19	9	19
POLAR	**Dipole [Debye]**	20	18	7	18	16	16
ENERGY	**Total Energy [Eh]**	21	9	14	21	13	**3**

Table 11
HRIPT binary outcome prediction performance (ROC AUC [95 % CI], * = significant difference vs LLNA) for the LLNA, individual Hf-3c CPCM models, and the combined model averaging the prediction probabilities from each individual model on the molecules for the entire CosEU dataset without filtering for the applicability domain.

Model	Type	ROC AUC [95 % CI]
LLNA	*In vivo*	0.674
Ames (Hf-3c CPCM)*	*In silico*	0.765 [0.718, 0.813]
KRS (Hf-3c CPCM)	*In silico*	0.641 [0.583, 0.699]
DPRA (HF-3c CPCM)	*In silico*	0.721 [0.672, 0.770]
All (Hf-3c CPCM)*	*In silico*	0.791 [0.749, 0.834]

describe mechanistic details within the solvated biological milieu, more rigorous automated model optimization methodologies, and consideration of the applicability domain.

While models based on quantum mechanical descriptors for either DPRA or KeratinoSens datasets did not show statistically significant performance improvements compared to conventional Mordred descriptors, their combination with the Ames QSAR model resulted in the highest LLNA predictive performance of any computational model in this project at 0.806 ROC AUC. This is comparable to existing read across models for the LLNA which feature 0.805 and 0.837 ROC AUC (Tung et al., 2018). Furthermore, the substitution of the DPRA QSAR model with the Ames QSAR model in the computational implementation of the "2 of 3" approach only produced marginally improved performance relative to the substantially increased number of chemicals. This indicates the DPRA dataset with Hf-3c CPCM features may include mechanistic information that is less readily accessible to the Ames mutagenicity assay. These findings show the small skin sensitization assay datasets that are currently available can still be used constructively to improve skin sensitization QSAR models in combination with the Ames mutagenicity QSAR model.

Most individual and combined QSAR models developed in this project display slightly increased performance when the applicability domain was taken into consideration compared to the Vega CAESAR model. CAESAR was selected as the best performing QSAR model that currently exists in the public domain (Braeuning et al., 2018), however, the exclusion of molecules that are outside the applicability domain was essential for CAESAR to perform comparably to the *in vitro* Ames assay without statistically significant difference (Table 8). This indicates the

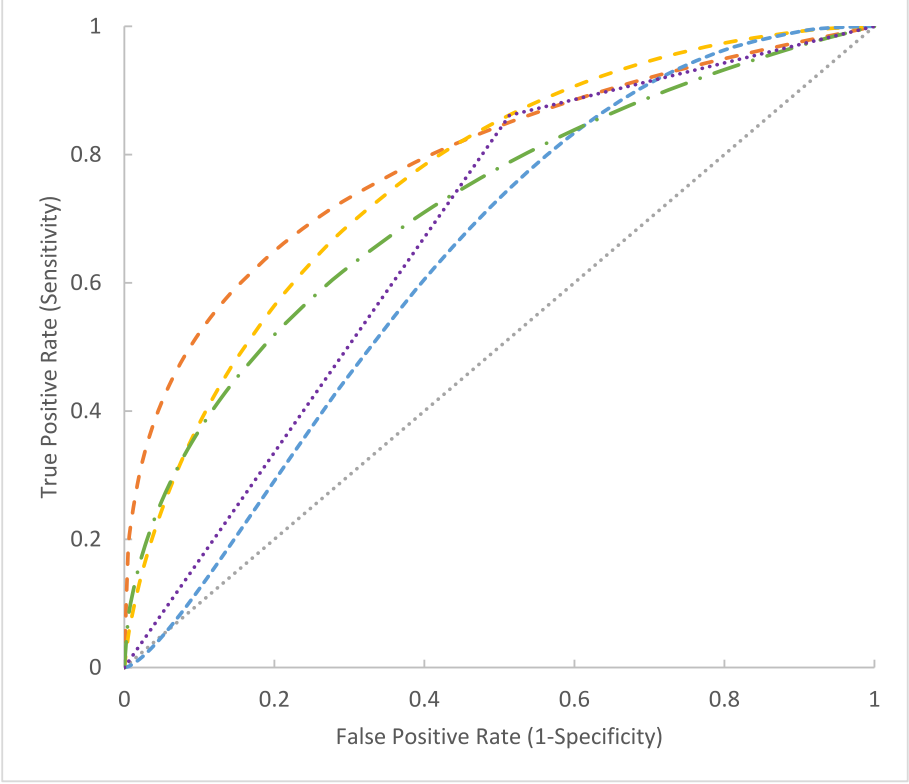


Fig. 9. ROC curves depicting binary HRIPT prediction performance of the individual Ames (yellow striped), KeratinoSens (blue striped), DPRA (green striped), and combined (orange striped) Hf-3c CPCM descriptor models for the entire CosEU dataset compared to the *in vivo* LLNA (purple dotted). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 12
HRIPT ternary outcome prediction accuracy for the LLNA and the combined Hf-3c CPCM model on the entire CosEU dataset.

Model	Accuracy
LLNA	59.6 %
All (Hf-3c CPCM)	53.2 %

QSAR models developed in this project are relatively less sensitive to applicability domain differences between the training and external validation datasets. The small performance difference from omitting the applicability domain to predict on all chemicals is desirable for screening compounds for sensitizing activity where other *in silico* methods are not applicable. The current applicability domain results indicate the combined model (86 % balanced accuracy) is comparable to reported commercial TOPKAT (86 %) and CASE ULTRA (76 %) *in silico* tools (Kostal and Voutchkova-Kostal, 2016). However, this could also be a result of low sensitivity in the unified applicability domain quantification methodology in this project compared to the "Applicability Domain Index" in CAESAR used to determine whether to include or exclude unseen molecules. A more conservative applicability domain quantification methodology could be implemented in future studies to further increase model performance at the expense of predicting a smaller fraction of the external validation dataset.

The Hf-3c quantum mechanical descriptor family was implemented in this project to construct QSAR models with high predictive performance while including a plausible mechanistic interpretation in accordance with Principle 5 of the OECD QSAR Guidelines (OECD, 2014). To that end, we found the high external performance from the Ames QSAR model may be due to improved performance against molecular steric factors that are not directly modelled using either quantum mechanical descriptor set as shown by the inclusion of the HOMO-1 and LUMO + 1 descriptors (Table 10). These molecular orbitals are used when the molecule is unable to interact with the HOMO or LUMO orbitals due to steric hindrance (Enoch, 2010). In comparison, the HOMO and LUMO descriptors are ranked in the top three of the solvated Hf-3c DPRA and KeratinoSens datasets which may be due to their low molecular diversity compared to the Ames dataset as shown in Fig. 1.

The inclusion of solvation using the implicit CPCM solvation model may be another factor that has resulted in the conservation of the HOMO and LUMO descriptors as the top ranking features for both DPRA and KeratinoSens datasets, in contrast to the presence of the IsotropicQuad and quadrupole_ZZ polarization descriptors in the Hf-3c vacuum variant of these datasets. Implicit solvation models represent solvated molecules in an electrostatic cavity that interacts with a dielectric medium parameterized from physicochemical descriptors which serves to generalize bulk solvent interactions to influence the geometry and polarization of a ligand (Barone and Cossi, 1998). As a result, the effects of polarization are reduced in QSAR models using Hf-3c descriptors with solvation in favor of descriptors that more directly related to the mechanism of toxicity. This presented a small external validation performance improvement (Table 3) and is consistent with previous findings in regarding their importance for predicting toxicological outcomes in the related genotoxicity domain (Karelson et al., 2000).

The omission of biotransformation is a general limitation of QSAR models that would more substantially affect quantum mechanical descriptors as they only represent the reactivity, polarization, and molecular orbitals of the parent structure and excludes consideration of metabolites. This limits QSAR model performance in chemical domains where biotransformation is essential for toxicity, such as aromatic amines with azo bonds that currently have poor concordance between *in vitro* Ames assay and their QSAR model results. Current QSAR models implicitly model biotransformation by using endpoints that include metabolism and large conventional descriptor sets with mixed results

Table 13
The OECD QSAR Guidelines and their methodological implementation in the current project.

OECD QSAR Guideline	Methodology
A defined endpoint	Binary LLNA outcomes
An unambiguous algorithm	A pipeline of data scaling and transformation algorithms, machine learning algorithms, and their hyperparameters for each dataset used to construct the model
A defined domain of applicability	The Applicability Domain Toolbox is used to assess the domain of applicability by comparing the unseen chemicals against those in the training dataset
Appropriate measure of goodness of fit, robustness and predictivity	Both internal and external validation is ranked with ROC AUC, an appropriate measure that is resistant to class imbalance
A mechanistic interpretation, if possible	The skin sensitization MIE is defined with non-specific covalent interactions. Quantum mechanics is the only theory that directly describes covalent interactions. The use of QM descriptors to quantify electronic properties related to reactivity readily enables mechanistic interpretation. See Permutation Feature Importance (Table 10).

(Hansen et al., 2009). However, the broad availability of metabolic simulators in the public domain to generate potential metabolites could enable quantum mechanical descriptors to sort and select metabolites based on their reactivity (Dimitrov et al., 2016; Djoumbou-Feunang et al., 2019). This could aid in the construction of biologically mechanistic hypotheses for assessing chemotypes for skin sensitization and mutagenicity. Unlike conventional descriptors, the *ab initio* nature of the Hf-3c method enables descriptor calculation for charged molecules and radicals that constitute toxicological metabolites. The lack of parameterization for ionic molecules in implicit solvation models has potentially been addressed with the recent CMIRS solvation model (Silva et al., 2016). In summary, biotransformation as a QSAR performance limitation could be addressed by explicitly modelling metabolites using quantum chemical descriptors coupled with recent implicit solvation models to enable mechanistic biological pathway construction.

The binary *in vivo* HRIPT results show the highest performing *in silico* model, which is the combined model averaging all individual model probability outputs, can exceed the prediction performance of the *in vivo* LLNA in the Cosmetics Europe database. This is comparable with various other *in silico* methodologies that have also been capable of exceeding LLNA predictive performance for HRIPT outcomes (Kleinstreuer et al., 2018). However, the current QSAR models in this project do not require the use of *in vitro* assays to produce an *in silico* prediction compared to previous Defined Approaches, enabling these QSAR models to be used prospectively in regulatory screening efforts at low fiscal cost. To that end, the QSAR models in this project fulfil the OECD QSAR Validation Guidelines as shown in Table 13.

Ternary HRIPT outcomes are not currently well predicted by either *in vivo* or *in silico* methodologies and requires further investigation in the future. The performance difference between the combined model could be due to the naïve nature of the threshold setting or the LLNA having greater dynamic range from being a model with five categorical outcomes. This limitation necessitates future study as the current REACH regulatory regime is also an ordinal ternary categorical scale for human skin sensitization outcomes of none, 1B (weak), and 1A (strong).

Conclusion

Quantum mechanical descriptors calculated at the Hf-3c level of theory with aqueous CPCM implicit solvation are a novel chemical representation that can generate QSAR models highly predictive of endpoints that incorporate covalent bonding mechanisms such as skin sensitization and mutagenicity. The use of these descriptors with the Ames assay dataset produced QSAR models with equivalent external validation performance to the *in vitro* Ames mutagenicity assay for predicting skin sensitization. This also shows the Ames assay dataset could be used as a surrogate dataset in the skin sensitization domain.

CRediT authorship contribution statement

Davy Guan: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Project administration, Visualization. **Raymond Lui:** Visualization. **Slade T. Mattthews:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code used in this paper is available via Github (link in manuscript). The data was gleaned from the literature as described.

References

Ashby, J., Hilton, J., Dearman, R.J., Callander, R.D., Kimber, I., 1993. Mechanistic relationship among mutagenicity, skin sensitization, and skin carcinogenicity. Environ. Health Perspect. 101 (1), 62–67. https://doi.org/10.1289/ehp.9310162.

Barone, V., Cossi, M., 1998. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. Chem. A Eur. J. 102 (11), 1995–2001. https://doi.org/10.1021/jp9716997.

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D.S., Smith, K., 2011. Cython: The best of both worlds. Comput. Sci. Eng. 13 (2), 31–39. https://doi.org/10.1109/MCSE.2010.118.

Benigni, R., Bossa, C., 2011. Alternative strategies for carcinogenicity assessment: an efficient and simplified approach based on in vitro mutagenicity and cell transformation assays. Mutagenesis 26 (3), 455–460. https://doi.org/10.1093/mutage/ger004.

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. Pharm. 78 (Supplement C), 45–52. https://doi.org/10.1016/j.yrtph.2016.04.003.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B. 2008. KNIME: The Konstanz Information Miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker, *Data Analysis, Machine Learning and Applications* Berlin, Heidelberg.

Braeuning, C., Braeuning, A., Mielke, H., Holzwarth, A., Peiser, M., 2018. Evaluation and improvement of QSAR predictions of skin sensitization for pesticides. SAR QSAR Environ. Res. 29 (10), 823–846. https://doi.org/10.1080/1062936X.2018.1518261.

Can, A., Yildiz, I., Guvendik, G., 2013. The determination of toxicities of sulphonylurea and phenylurea herbicides with quantitative structure–toxicity relationship (QSTR) studies. Environ. Toxicol. Pharmacol. 35 (3), 369–379. https://doi.org/10.1016/j.etap.2013.02.001.

Chaudhry, Q., Piclin, N., Cotterill, J., Pintore, M., Price, N.R., Chrétien, J.R., Roncaglioni, A., 2010. Global QSAR models of skin sensitisers for regulatory purposes. Chem. Cent. J. 4 (Suppl 1), S5–S. https://doi.org/10.1186/1752-153X-4-S1-S5.

Chipinda, I., Hettick, J.M., Siegel, P.D., 2011. Haptenation: chemical reactivity and protein binding. J. Allergy.

Dimitrov, S.D., Diderich, R., Sobanski, T., Pavlov, T.S., Chankov, G.V., Chapkanov, A.S., Karakolev, Y.H., Temelkov, S.G., Vasilev, R.A., Gerova, K.D., Kuseva, C.D., Todorova, N.D., Mehmed, A.M., Rasenberg, M., Mekenyan, O.G., 2016. QSAR Toolbox - workflow and major functionalities. SAR QSAR Environ. Res. 27 (3), 203–219. https://doi.org/10.1080/1062936x.2015.1136680.

Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., & Wishart, D. S. J. J. o. C. (2019). BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification [journal article]. *11*(1), 2. https://doi.org/10.1186/s13321-018-0324-5.

Enoch, S.J., 2010. Chapter 7 The Use of Frontier Molecular Orbital Calculations in Predictive Reactive Toxicology. In: In Silico Toxicology: Principles and Applications. The Royal Society of Chemistry, pp. 193–209. https://doi.org/10.1039/9781849732093-00193.

Enoch, S.J., Cronin, M.T.D., Schultz, T.W., Madden, J.C., 2008. Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via Michael addition. Chem. Res. Toxicol. 21 (2), 513–520. https://doi.org/10.1021/tx700322g.

Enoch, S.J., Roberts, D.W., 2013. Predicting skin sensitization potency for Michael acceptors in the LLNA using quantum mechanics calculations. Chem Res Toxicol 26 (5), 767–774. https://doi.org/10.1021/tx4000655.

Gadarowska, D., Kalka, J., Daniel-Wojcik, A., Mrzyk, I., 2022. Alternative methods for skin-sensitization assessment. Toxics 10 (12). https://doi.org/10.3390/toxics10120740.

Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T., Heinrich, N., Muller, K.R., 2009. Benchmark data set for in silico prediction of Ames mutagenicity. J. Chem. Inf. Model. 49 (9), 2077–2081. https://doi.org/10.1021/ci900161g.

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., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (I): the Cosmetics Europe database. Crit. Rev. Toxicol. 48 (5), 344–358. https://doi.org/10.1080/10408444.2018.1429385.

Kamber, M., Fluckiger-Isler, S., Engelhardt, G., Jaeckh, R., Zeiger, E., 2009. Comparison of the Ames II and traditional Ames test responses with respect to mutagenicity, strain specificities, need for metabolism and correlation with rodent carcinogenicity. Mutagenesis 24 (4), 359–366. https://doi.org/10.1093/mutage/gep017.

Karelson, M., Sild, S., Maran, U., 2000. Non-Linear QSAR Treatment of Genotoxicity. Mol. Simul. 24 (4–6), 229–242. https://doi.org/10.1080/08927020008022373.

Kazius, J., McGuire, R., Bursi, R., 2005. Derivation and validation of toxicophores for mutagenicity prediction. J. Med. Chem. 48 (1), 312–320. https://doi.org/10.1021/jm040835a.

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., Petersohn, D., 2018. Non-animal methods to predict skin sensitization (II): an assessment of defined approaches. Crit. Rev. Toxicol. 48 (5), 359–374. https://doi.org/10.1080/10408444.2018.1429386.

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 (1), 58–64. https://doi.org/10.1021/acs.chemrestox.5b00392.

Landrum, G. *RDKit: Open-source cheminformatics*. In http://www.rdkit.org.

Le, T.T., Fu, W., Moore, J.H., 2019. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics 36 (1), 250–256. https://doi.org/10.1093/bioinformatics/btz470.

Li, T., Liu, Z., Thakkar, S., Roberts, R., Tong, W., 2023. DeepAmes: A deep learning-powered Ames test predictive model with potential for regulatory application. Regul. Toxicol. Pharmacol. 144, 105486 https://doi.org/10.1016/j.yrtph.2023.105486.

Marenich, A.V., Cramer, C.J., Truhlar, D.G., 2009. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J. Phys. Chem. B 113 (18), 6378–6396. https://doi.org/10.1021/jp810292n.

Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M., Stoica, I. 2018. Ray: A Distributed Framework for Emerging AI Applications. *arXiv.org*.

Moriwaki, H., Tian, Y.-S., Kawashita, N., Takagi, T., 2018. Mordred: a molecular descriptor calculator. J. Cheminf. 10 (1), 4. https://doi.org/10.1186/s13321-018-0258-y.

Neese, F., 2012. The ORCA program system. Wiley Interdiscip. Rev.: Comput. Mol. Sci. 2 (1), 73–78. https://doi.org/10.1002/wcms.81.

Neese, F., 2018. Software update: the ORCA program system, version 4.0. Wiley Interdiscip. Rev.: Comput. Mol. Sci. 8 (1), n/a-n/a. https://doi.org/10.1002/wcms.1327.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: An open chemical toolbox. J. Cheminf. 3 (1), 33. https://doi.org/10.1186/1758-2946-3-33.

OECD. (2014). *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. https://doi.org/doi:https://doi.org/10.1787/9789264085442-en.

OECD. (2016). *Series on Testing and Assessment No. 256: Guidance Document on the Reporting of Defined Approaches and Individual Information Sources to Be Used within IATA for Skin Sensitisation.* . Retrieved from http://www.oecd.org/chemicalsafety/testing/series-testing-assessment-publicationsnumber.htm.

Olson, R. S., Bartley, N., Urbanowicz, R. J., Moore, J. H. 2016. *Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science* Proceedings of the Genetic and Evolutionary Computation Conference 2016, Denver, Colorado, USA. https://doi.org/10.1145/2908812.2908918.

Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., Kidd, L. C., Moore, J. H. 2016. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. Applications of Evolutionary Computation, Cham.

Pandith, A.H., Giri, S., Chattaraj, P.K., 2010. A comparative study of two quantum chemical descriptors in predicting toxicity of aliphatic compounds towards Tetrahymena pyriformis. Org. Chem. Int. 1–17 https://doi.org/10.1155/2010/545087.

Patlewicz, G., Mekenyan, O., Dimitrova, G., Kuseva, C., Todorov, M., Kotov, S., Stoeva, S., Donner, E.M., 2010. Can mutagenicity information be useful in an Integrated Testing Strategy (ITS) for skin sensitization? SAR QSAR Environ. Res. 21 (7–8), 619–656. https://doi.org/10.1080/1062936X.2010.528447.

Patlewicz, G., Casati, S., Basketter, D.A., Asturiol, D., Roberts, D.W., Lepoittevin, J.P., Worth, A.P., Aschberger, K., 2016. Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? Regul. Toxicol. Pharmacol. 82, 147–155. https://doi.org/10.1016/j.yrtph.2016.08.007.

Piegorsch, W., Zeiger, E. 1991. Statistical Methods in Toxicology. In *Statistical Methods in Toxicology. Vol. 43* (pp. 35). Springer Heidlberg.

Puzyn, T., Suzuki, N., Haranczyk, M., Rak, J., 2008. Calculation of quantum-mechanical descriptors for QSPR at the DFT Level: Is it necessary? J. Chem. Inf. Model. 48 (6), 1174–1180. https://doi.org/10.1021/ci800021p.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R., 2012. Comparison of different approaches to define the applicability domain of QSAR models. Molecules 17 (5), 4791–4810. https://doi.org/10.3390/molecules17054791.

Sahigara, F., Ballabio, D., Todeschini, R., Consonni, V., 2014. Assessing the validity of QSARs for ready biodegradability of chemicals: an applicability domain perspective. Curr. Comput. Aided Drug Des. 10 (2), 137–147. https://doi.org/10.2174/1573409910666140410110241.

Silva, N.M., Deglmann, P., Pliego, J.R., 2016. CMIRS solvation model for methanol: parametrization, testing, and comparison with SMD, SM8, and COSMO-RS. J. Phys. Chem. B 120 (49), 12660–12668. https://doi.org/10.1021/acs.jpcb.6b10249.

Stewart, J. J. P. 2016. *MOPAC2016*. In Stewart Computational Chemistry. HTTP://OpenMOPAC.net.

Sure, R., Grimme, S., 2013. Corrected small basis set Hartree-Fock method for large systems. J. Comput. Chem. 34 (19), 1672–1685. https://doi.org/10.1002/jcc.23317.

Ta, G.H., Weng, C.F., Leong, M.K., 2021. In silico prediction of skin sensitization: Quo vadis? Front. Pharmacol. 12, 655771 https://doi.org/10.3389/fphar.2021.655771.

The MathWorks, I. 2019. *MATLAB and Statistics Toolbox Release 2019b*. In The MathWorks, Inc.

Townsend, P.A., Grayson, M.N., 2020. Reactivity prediction in aza-Michael additions without transition state calculations: the Ames test for mutagenicity. Chem. Commun. (Camb.) 56 (88), 13661–13664. https://doi.org/10.1039/d0cc05681b.

Tung, C.-W., Wang, C.-C., Wang, S.-S., 2018. Mechanism-informed read-across assessment of skin sensitizers based on SkinSensDB. Regul. Toxicol. Pharm. 94, 276–282. https://doi.org/10.1016/j.yrtph.2018.02.014.

Wang, C.C., Lin, Y.C., Wang, S.S., Shih, C., Lin, Y.H., Tung, C.W., 2017. SkinSensDB: a curated database for skin sensitization assays. J. Cheminform. 9, 5. https://doi.org/10.1186/s13321-017-0194-2.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. 2016. *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc.

Wolfreys, A.M., Basketter, D.A., 2005. Mutagens and sensitizers—An unequal relationship? J. Toxicol.: Cutaneous Ocul. Toxicol. 23 (3), 197–205. https://doi.org/10.1081/CUS-200025577.