Imputation of Sensory Properties Using Deep Learning

Samar Mahmoud†, Benedict Irwin†, Dmitriy Chekmarev*, Shyam Vyas*, Jeff Kattas*, Thomas Whitehead§, Tamsin Mansley†, Jack Bikker*, Gareth Conduit§‡, Matthew Segall†

† Optibrium Limited, Cambridge, UK
*** Research and Development, International Flavors & Fragrances, Union Beach, NJ**
§ Intellegens Limited, Cambridge, UK
‡ University of Cambridge, Cambridge, UK

Corresponding Author: Samar Mahmoud, samar@optibrium.com

Abstract

Predicting the sensory properties of compounds is challenging due to the subjective nature of the experimental measurements. This testing relies on a panel of human participants and is therefore also expensive and time-consuming. We describe the application of a state-of-the-art deep learning method, Alchemite™, to the imputation of sparse physicochemical and sensory data and compare the results with conventional quantitative structure-activity relationship methods and a multi-target graph convolutional neural network. The imputation model achieved a substantially higher accuracy of prediction, with improvements in R^2 between 0.26 and 0.45 over the next best method for each sensory property. We also demonstrate that robust uncertainty estimates generated by the imputation model enable the most accurate predictions to be identified and that imputation also more accurately predicts activity cliffs, where small changes in compound structure result in large changes in sensory properties. In combination, these results demonstrate that the use of imputation, based on data from less expensive, early experiments, enables better selection of compounds for more costly studies, saving experimental time and resources.

Keywords

Sensory properties, *in silico* model, deep learning, imputation, quantitative structure-activity relationship

Declarations

Funding

GC would like to acknowledge financial support from the Royal Society.

Conflicts of Interest/Competing Interests

SM, BI, TM and MS are employees of Optibrium Limited. DM, SV, JK and JB are employees of International Flavors & Fragrances, Inc. TW and GC are employees of Intellegens Limited.

Availability of data and material

We are unable to publish the full data set used in this study due to the proprietary nature and high cost of the data. Nevertheless, we believe that the first demonstration of the potential for deep learning imputation to make accurate predictions of such challenging in vivo properties warrants publication. To ensure reproducibility and enable comparisons with other methods, we have previously published benchmarking studies accompanied with public domain data sets including all predicted values [Whitehead *et al.* J. Chem. Inf. Model. (2019) **59**(3) pp. 1197-1204, Irwin *et al.* App. AI Lett. (2021) **2**(3) e31 DOI:10.1002/ail2.31].

Code availability

The code used in this study is proprietary.

Authors' contributions

All authors contributed to the study conception and design. Data curation was performed by Dmitriy Chekmarev; modelling and analysis of the results was performed by Samar Mahmoud; the Alchemite method was developed by Gareth Conduit and Tom Whitehead.. The first draft of the manuscript was written primarily by Samar Mahmoud with contributions from all authors. All authors read and approved the final manuscript.

Introduction

The olfactive system is highly evolved to translate chemical information in food and our surroundings into essential impressions of beauty, events, or even imminent danger [1, 2, 3]. The chemical information in the environment is generally sensed through the nose directly (orthonasally), whereas the impact of food aroma is sensed both orthonasally and retronasally (through the back of the throat) [4]. The basis of smell and aroma perception in humans is an array of sensory neurons that extend into the sensory epithelium and present an array of approximately 400 unique olfactory receptors, with each sensory neuron expressing a single receptor type [5, 6, 7]. The intensity of smell or flavour so sensed correlates with concentration. The pattern of functional activation of olfactory receptors connects to evoked memory to support scent and aroma classification and assessment [8]. Finally, odor adaptation biases the systems to detect novelty allowing us to focus on potentially important changes happening around us [9]. The science and art of perfumery and flavour seeks to accentuate and mimic key sensory experiences in our surroundings. It can evoke a sense of beauty, soothe with a sense of familiarity, reinforce authenticity, or can startle with creative novelty [10].

Perfumers and flavorists draw from a large palette of ingredients to affect these experiences. These molecules can be drawn from nature if they are already components of natural oils and spices [11], or they can arise from the long history of aromatic molecule discovery, especially to support perfumery [12, 13, 14]. Scent creations include fine perfumes and consumer applications such as shampoos, fabric conditioners [15], detergents, candles, cleaning solutions and lotions. Molecules in flavours appear in foods that range from baked goods, beverages, savory snacks, dairy products, and sweets. The intended use can constrain the ingredients used and dictates the nature and evolution of the sensory experience. By understanding chemical and sensorial properties, ingredient and mixture performance can be predicted, and an optimal selection of ingredients can be chosen. In this way, predictions can allow better selection of existing materials and design of new materials to identify and fill gaps that exist historically or that emerge as ingredients are withdrawn from use for e.g. regulatory reasons.

Olfactive properties are the ultimate benchmark of a chemical's value in fragrance and flavour. The odor intensity/concentration relationship enables perfumers and flavorists to appropriately dose chemicals in complementary ways. When designing or selecting new molecules for use, being able to predict the intensity of a candidate's odour or flavour at low concentrations can aid in prioritizing the high-value targets. Unfortunately, this is remarkably challenging.

Part of the difficulty in developing accurate predictive models lies in the inherent challenges in collecting reproducible psychophysical data. The methods for determining a quantification of a chemical's intensity are primarily subjective and, because they rely on a panel of human participants, they are expensive and time-consuming. Subjects are asked to rate an odor on a particular scale that contains landmarks or qualifiers. A fair amount of training is required to ensure any individual subject can reproducibly use the same rating for identical stimuli across tests. Even while minimizing the error here, physiological differences across subjects introduce a high amount of variability into any one measurement. Sensitivities to an odorant will vary greatly within a population [16]. Ergo, one needs to obtain a fairly large sample size to achieve reproducibility. Even then, what ends up being an averaged value per molecule is really a representation of a normal distribution of responses. Training a model on these types of data is quite challenging, and the inherent variability must be kept in mind when judging success. In addition to dose/response intensity data, the other important variable is the detection threshold. Typically, this is set as the lowest concentration that is detectable 50% of the time. It is often used to quantify the relative strength of ingredients, although it may not reliably indicate the maximum intensity of the ingredient. Together, the odor threshold, the dose/response behaviour, the character of the ingredient, and key physical properties represent the attributes that characterize an ingredient’s behaviour in use [17].

Volatile aromatic molecules are necessarily small molecules; molar mass ranges up to about 300, with rare exceptions. These typically have only one hydrogen bond donor and only up to three hydrogen bond acceptors [18, 19]. Their complexity arises from subtle changes in the hydrophobic portion of the molecule, with the relocation of even a single methyl group affecting intensity or even character [20]. Structure-odor activity is notoriously sensitive to these changes, and analogue series can show significant activity cliffs. Structure-property data is generally more well-behaved, with important attributes that include vapor pressure, water solubility and LogP. Unfortunately, there are no good public comprehensive repositories of key attributes, and corporate data sets can often be sparse due to the cost and care needed to obtain reliable data, especially of low volatile materials.   There is, therefore, significant interest in generating functional predictive models to support fragrance and flavour creation, to reduce experimental effort, and to guide the design or selection of new molecules.

The development of computational models that link the structure of compounds to their activities or properties is commonplace, and these models are described as quantitative structure-activity relationship (QSAR) or quantitative structure-property relationship (QSPR) models.

QSAR models are mathematical functions that relate simple structural characteristics of a compound (e.g. the dipole moment or molecular weight), known as descriptors, to an experimentally measured endpoint. Early QSAR models relied on linear relationships fitted with methods such as multiple linear regression or partial least squares (PLS) [21]. However, it is now routine to use sophisticated machine learning (ML) methods, such as random forests (RF) [22], support vector machines (SVM) [23], radial basis functions (RBF) [24] or Gaussian processes (GP) [25]. Sophisticated deep learning methods have also been applied to QSAR modelling, including multi-target methods that build a single model that can predict multiple endpoints simultaneously [26]. Multi-target methods offer an advantage where structure-activity relationships are shared between multiple endpoints because this information can improve the accuracy of the resulting predictions. The concept of a QSAR model, relating compound descriptors to one or more experimental endpoints, is illustrated in Fig. 1(a). Recently, QSAR models have also been built using graph convolutional neural networks (GCNN) that learn directly from the two-dimensional graph structure of compounds, without the need to generate descriptors as an intermediate step. GCNNs eliminate the assumption implicit in the selection of descriptors (i.e. that the input descriptors capture the structural information that most strongly correlates with the experimental endpoints) and offer an advantage for modelling some properties [27].

There are several examples of the application of ML methods to generate QSAR models of sensory properties. For example, neural networks were used to predict qualitative sensory properties using mass spectra (MS) data [28] and near-infrared spectra (NIRS) [29] as compound descriptors. While data from MS and NIRS can be useful to capture the structural characteristics of compounds, generating these measurements is time-consuming and impossible for compounds that have not yet been synthesized (virtual compounds). Hence, attempts were made to implement *in silico* chemical descriptors in training machine learning models to predict sensory properties such as taste [30] and odor. For example, odor descriptors were predicted using GCNN for over 5000 compounds achieving 82% accuracy and outperforming random forest models [31]. The challenge in predicting sensory properties can be attributed to the variability in the sensory property data and the subjectivity of the panellist, which cannot be captured solely by the chemical structure.

While the predictive modelling process is commonly understood, a less familiar term 'imputation' describes the process of filling in the gaps in a data set of multiple properties, where values have not yet been measured, using the limited (sparse) data that are already present, as illustrated in Fig. 1(b) [32]. Imputation models learn from the relationships between experimental endpoints and are particularly relevant in chemistry optimization settings, such as drug discovery and flavor & fragrances, where sparse data sets are standard. Data sets of multiple experimental measurements tend to be sparse. Few compounds are measured in all experimental endpoints of interest, and it is rare to perform an experiment on all potential compounds of interest.

In this study, we use a novel deep learning method for imputation, Alchemite, that learns from both structure-activity relationships and correlations between experimental endpoints to impute missing experimental values, as illustrated in Fig. 1(c). In this paper, we also demonstrate that Alchemite can be used to make predictions based only on chemical descriptors in cases where experimental data for compounds are absent, such as for virtual compounds. Alchemite can still exploit correlations between experimental endpoints it learned in training and use these to enhance predictions. In previous publications, we have demonstrated that this approach gains more value from limited and noisy experimental data than QSAR models to improve the accuracy of the resulting predictions [33] [34] [35]. This included benchmarks against a wide variety of machine learning methods using published data sets to enable direct comparisons to be made.

Ideally, any predictive method should provide an estimate of the uncertainty in each prediction, which can be used to determine if the prediction can be used with sufficient confidence to make decisions regarding the selection of a compound for synthesis or testing [36]. Some QSAR methods, such as RF or GP, generate uncertainty estimates for each prediction, and a variety of approaches have been explored to generate uncertainty estimates for modern deep learning methods, but the results are not generally reliable [37]. Conversely, Alchemite generates a probability distribution for each prediction, from which uncertainties can be estimated that correlate strongly with the accuracies of prediction [33] [34].

We present here a novel application of the Alchemite method for deep learning imputation to the prediction of sensory properties. We describe the physicochemical and sensory property data in the data set used and compare the Alchemite models for both imputation and virtual prediction with a variety of widely-applied QSAR methods and a multi-task deep convolutional network (ChemProp). Despite the sparsity and subjectivity of the sensory data used, we demonstrate that Alchemite imputation predicts the values of these properties with significantly higher accuracy than these other methods. Furthermore, we show that the uncertainty estimates generated by Alchemite enable identification of the most accurately predicted values.

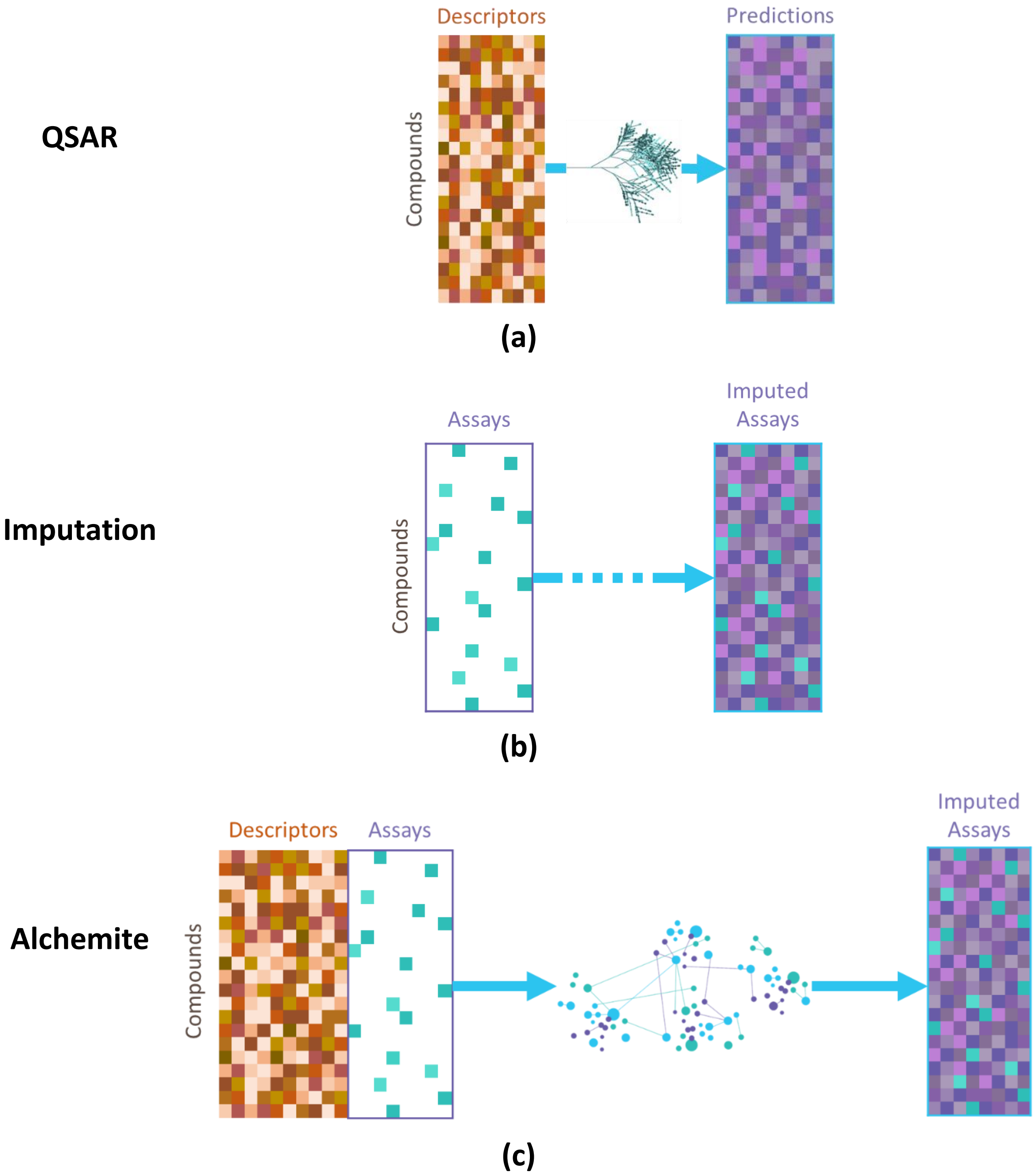


Fig. 1 *A comparison of QSAR regression, e.g. random forest models (a), with imputation (b) and the Alchemite deep imputation software (c). Compound descriptors are illustrated by orange squares in a complete matrix, experimental data are shown as green squares in a sparse matrix and predicted values as purple squares*

Methods

Alchemite method for deep learning imputation

The Alchemite method is a deep and iterative multiple imputation method that is a novel adaptation of a neural network in which all inputs are also outputs. The overall architecture is shown in **Error! Reference source not found.**.

The Alchemite algorithm starts with the input of the complete set of N_d descriptors and sparse set of N_e endpoints that include the available experimental results. Since the underlying machine learning architecture can use the endpoint values to help predict other endpoints, for any unknown properties there need to be estimates of them available for the neural network (which has a fixed architecture). Therefore, the second step is to fill in estimates of those absent values: here, we adopt the simplest possible a priori estimate, which is the mean of all other values for that property in the training set.

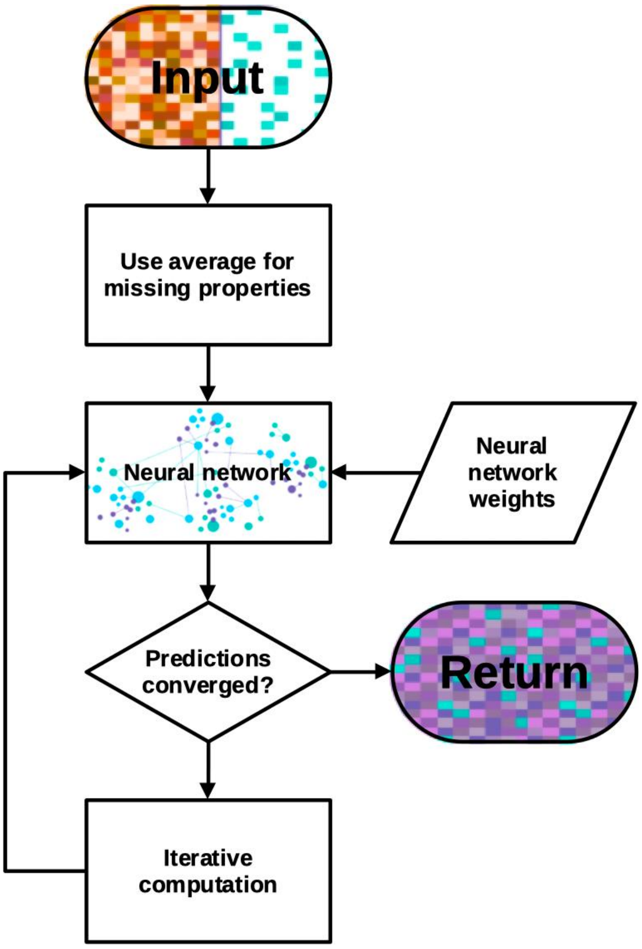


Fig. 2 Data imputation algorithm for starting from the descriptors and properties that have missing entries

The third step is to pass all of these values through to a standard neural network. The parameters of the neural network are considered as hyperparameters of the model. The neural network inputs all descriptors and property values and delivers predictions for all property values, taking care to not use a given property value to predict itself. This algorithm is able to automatically identify the descriptor-property correlations as well as property-property correlations to guide the extrapolation of the model. Furthermore, we adopt a standard bootstrapping approach and train multiple neural networks with different weights on each row of training data. The mean of the predictions from these neural networks is used as the output value of the network, with the standard deviation between predictions delivering an estimate of the uncertainty in the prediction.

The predictions from the neural network are used to impute the gaps in the originally sparse endpoint data. These predictions are used to augment the original descriptor and experimental data (in place of the original mean estimates) for another pass through a (different) neural network. The predictions at each iterative pass are softened by combining them with those from the previous pass, $z^{n+1} = \gamma z^{n} + (\gamma - 1)y^{n+1}$, where z^{n} denotes the prediction from the nth iteration, with y^{n+1} the prediction from the $(n+1)$th iteration before softening and the softening parameter γ being one of the hyperparameters of the model.

This procedure is repeated until convergence is reached, with the termination criterion one of the hyperparameters of the model. The final model predictions are then the output of the final neural network. The result is a complete set of experimental endpoints in which the missing values have been imputed, as illustrated in Fig. 1(c). A further description of the underlying algorithm is given by Verpoort *et al.* [38] and for applications to drug discovery by Whitehead *et al.* [33] and Irwin *et al.* [34] [35].

There are two application scenarios for which Alchemite models can be built, depending on the availability of experimental data for the compounds of interest. As illustrated in Fig. 3(a), an Alchemite imputation model is built using both compound descriptors and sparse experimental data as input (as described above) and is trained to expect both compound descriptors and sparse experimental data for new compounds with which to impute missing values. An alternative to this is a 'virtual' model that is optimized to apply to compounds for which no experimental data are available, such as for compounds that have not yet been synthesized, *i.e.* virtual compounds. As illustrated in Fig. 3(b), the virtual model is built in a similar manner to the Imputation model, using compound descriptors and experimental data; however, it is trained to expect only compound descriptors as input. To achieve this, during hyperparameter optimization, the experimental data of compounds in the validation set of each cross-validation fold are withheld and only used to test the predictions of the model. Similarly, when testing the virtual model, no experimental data are used as input.

For both the Imputation and Virtual models, the hyperparameters of the model were optimized using a five-fold cross-validation within the training set data only. The tree-structured Parzen estimator [39] from the python library hyperopt [40] was used. 1,000 networks are trained in parallel for each iterative stage, using different weights for each row of data to generate an ensemble of predictions for each missing value in the dataset.

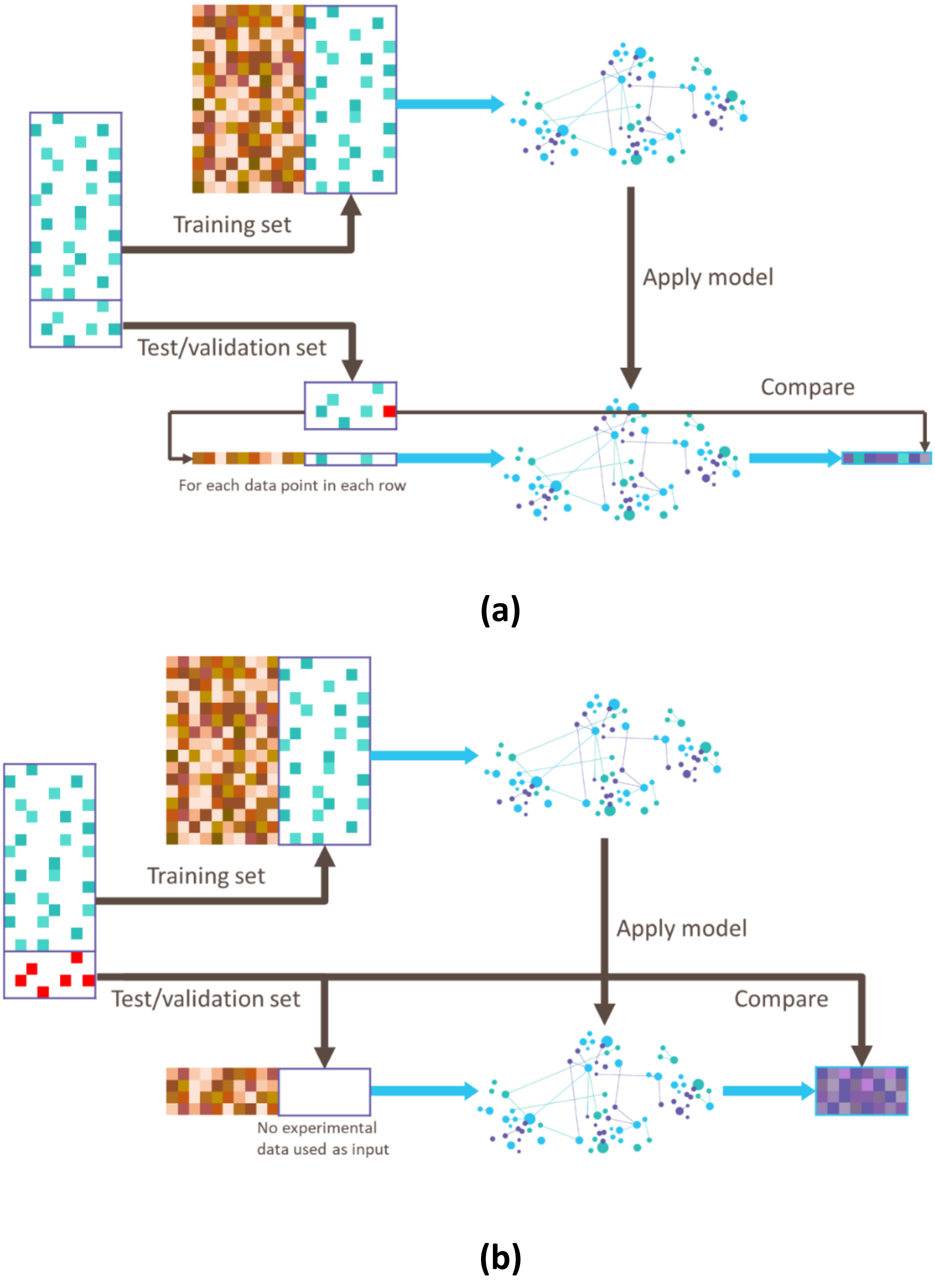


Fig. 3 *The processes to build and validate imputation (a) and virtual (b) models. The training and test sets consist of rows representing different compounds. For both models, all of the experimental data for the training set (green squares) are used to train a model. For the test or validation set of an imputation model, a leave-one-out approach is used to test the accuracy of the imputation of each experimental endpoint. For a given row, one data point (red square) is omitted, and the remaining experimental data (green squares) are used as input to the model. The predicted value (purple square) is compared with the held-out data point. This is repeated for each data point in each row of the test set. For the test set of the virtual model, no experimental data are used as input, and the predicted values (purple squares) are compared with the experimental values in the test or validation set (red squares)*

Other machine learning methods used for comparison

QSAR Methods

We compare Alchemite models described herein with conventional QSAR models built for each property using the Auto-Modeller module of the StarDrop software [41]. For each endpoint, we applied four widely-applied QSAR methods: Partial least squares (PLS) [21], which describes the target property as a linear combination of latent variables; Radial basis functions (RBF) [24]; Random forests (RF) [22]; and Gaussian process (GP) with fixed hyperparameters [25].

The same chemical descriptors were used as input for the QSAR and Alchemite models. For details, please see below.

Graph Convolutional Neural Network

In addition to conventional QSAR models, we compared the results with those generated using a multi-task GCNN, Chemprop [42]. Chemprop is a directed-message passing neural network, which learns directly from the graph structure of the molecule. The model constructs molecular encodings by using convolutions centred on bonds, which represent feature vectors. K. Yang et al. [42] provide a detailed description of the Chemprop method. To enable direct comparison with descriptor-based methods, Chemprop multi-task models were also built using the same chemical descriptors used as input for the QSAR and Alchemite models. Hereafter, the results for the application of graph-based Chemprop are referred to as 'Chemprop (graph)' and those relating to the descriptor-based application as 'Chemprop (descriptors)'.

Data set

The property data were collected for 1094 single volatile organic compounds from IFF's in-house database. Most of these compounds are individual ingredients used in a variety of fragrance formulations. They include chemicals from different molecular classes and olfactory groups, as illustrated in Fig. 4. Seven properties were selected for modelling: two physicochemical properties, vapor pressure (VP) and water solubility (WS), and five sensory properties, the odor detection threshold (ODT) and four odor intensity values collected along the odor intensity dose response curve. An odor intensity dose response curve is a collection of odor intensity evaluations performed by a panel of testers and arranged on a semi-log scale [43]. The odor intensity assessments (I1 through I4) are carried out for a series of dilutions of an odorant in a solvent commonly used in the fragrance industry, such as diethyl phthalate. The odor detection threshold is the lowest concentration that is detectable 50% of the time; the widest used method involves the three alternate forced choice test [44]. Together, these characteristics are important in measuring the performance of fragrance compositions. All experimental measurements were performed internally at IFF according to proprietary testing protocols. The acquired data for VP, WS and ODT are usually positively skewed and were log-transformed during pre-processing. An illustrative example of the structure-property data set is shown in Fig. 5.

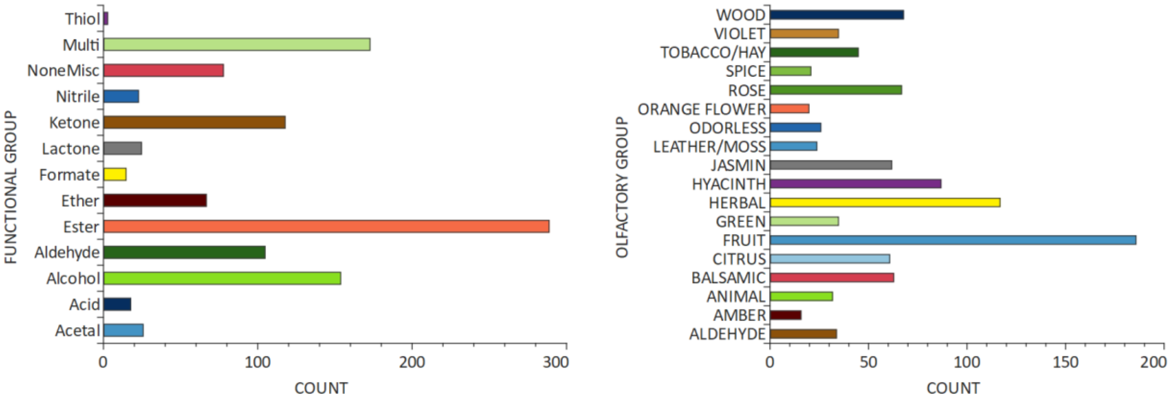


__Fig. 4__ Distribution of data set compounds across chemical classes and odor categories

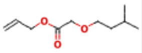
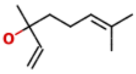
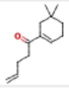
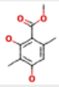
	🟩NAME	🟩STRUCTURE	🟪ODOR GROUP	🟦VP	🟩WS	🟩ODT	🟩I1	🟦I2	🟥I3	🟨I4
1	ALLYL AMYL GLYCOLATE		FRUIT	0.67	-2.316	1.47	2	8	21	27
2	LINALOOL		HERBAL	1.22	-1.971	1.23	2	5	15	23
3	GALBASCONE		FRUIT	0.05	-3.34	-1.73	3	7	14	18
4	VERAMOSS		LEATHER/MOSS	-2.63	-4.04	-2.1	4	7	13	18

__Fig. 5__ A sample from the structure-property data set. VP: vapor pressure (log transformed). WS: water solubility (log transformed). ODT: odor detection threshold (log transformed). I1-I4: odor intensity measurements from the odor intensity dose-response curve

The sparsity of the property data, the percentage of missing values, varies from 15% for LogVP to 56% for LogODT. There is a significant degree of intercorrelation between four odor intensity measurements with the Pearson pair correlation coefficient ranging from 0.5 to 0.91. This is expected because the four odor intensity records were taken along the dose-response curves, which follow a similar sigmoidal shape for many fragrance compounds. The correlations were less pronounced between the physicochemical properties (VP and WS) and sensory properties, varying from 0.32 to 0.68. The odor detection threshold (LogODT), as measured by the internal IFF protocol, was the property least correlated with the other properties in the set: it shows a moderate positive correlation of 0.52 with vapor pressure (LogVP) and weak negative correlations, -0.23 and -0.17, with low odor intensity measurements.

The data set was divided into a training subset of 931 molecules (~ 85% of the total data set) and an independent test subset of 163 molecules (~ 15%), which were used to train and independently assess models' performance respectively. The primary objective of an imputation model, for which the objective is to 'fill in' missing data for compounds that have been synthesized and tested in a limited number of experiments. For this reason, the chemicals in the held-out test set were selected to approximately maintain the observed distributions in both chemical and property spaces as seen in the training sets, as shown in Fig. 6, 7, and 8. The additional requirement was to have a sufficient number of chemicals with measured odor-detection threshold (ODT) in the test set, the property with the fewest available experimental data, to permit practical comparison between the considered models. This approach contrasts with cluster- or time-based training/test set splits often used to assess QSAR models, for which the goal is often to guide the design of new (virtual) compounds, which may differ substantially from the training set of the model.

A further subset of 20% of the compounds in the training set was selected for internal validation. The remaining 80% subset was used to optimize the hyperparameters and build models to assess the performance on the internal validation set. All of the training set (931 compounds) was used to train the final model, using the optimal hyperparameters, which was applied to the independent test set.

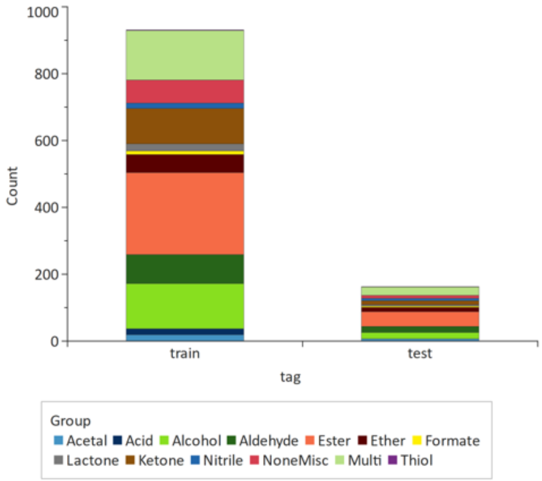


Fig. 6 *Chemical composition of the training (931 compounds) and test (163 compounds) sets*

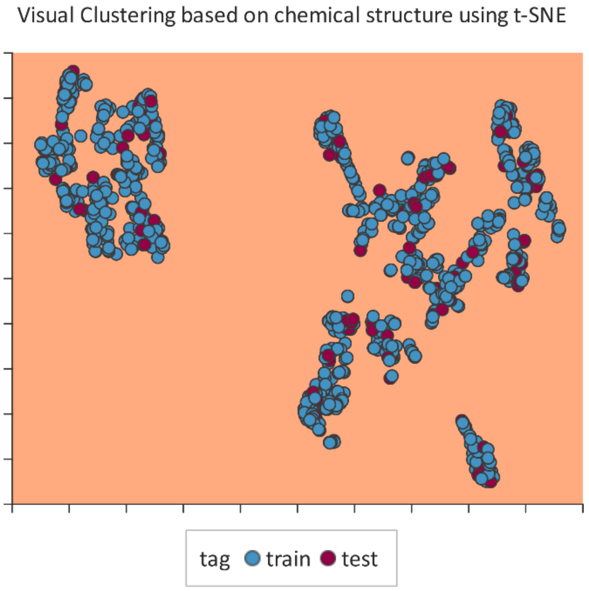


Fig. 7 *t-SNE 2D projection based on chemical structure. Blue: molecules from the training set. Maroon: molecules from the test set*

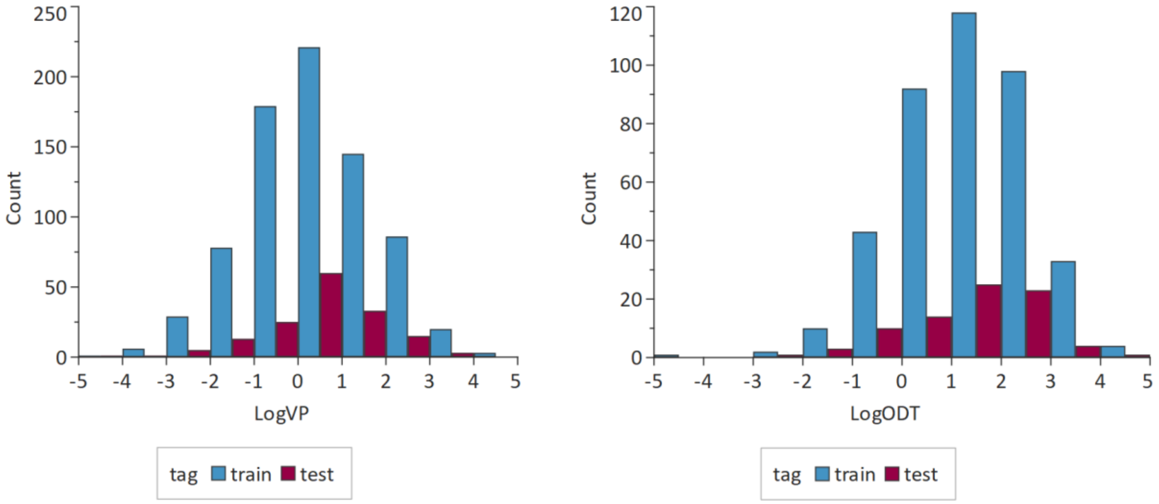


Fig. 8 *Distributions of physicochemical (vapor pressure) and olfactory (odor detection threshold) properties for the training and test sets*

We note that the density of data in the test set is higher than that in the training or total data sets. Given the relatively small data set size, it was necessary to maximize the number of experimental observations in the test set in order to more reliably evaluate models' performance in the 'virtual screening' scenario when experimental properties are not used for model training and prediction. The evaluation of Alchemite 'imputation' models, where experimental data are used to train the models, in addition to chemical molecular descriptors, was also conducted using the same independent test set. It is possible that the dense structure of the data in the test set with fewer missing experimental values may benefit imputation predictions for strongly correlated properties, such as for the close points on the odor intensity dose-response curve. However, we believe the predictions for the key three property categories, physicochemical VP and/or WS, odor detection threshold, and single point intensity, will remain largely uninfluenced because the three properties are only moderately correlated.

Chemical Descriptors

The number of molecular descriptors used as inputs was $N_d = 330$. The descriptors used included whole-molecule properties such as molecular weight, lipophilicity, and polar surface area; and structural fragments defined by SMARTS strings. These descriptors were calculated with the Auto-Modeller™ module of the StarDrop™ software using SMILES strings defining the structure of each compound.

Model Evaluation

The models were evaluated on the independent test set using two statistics: the coefficient of determination (R2), defined as

$$R^2 = 1 - \frac{\sum_i (y_i^{pred} - y_i^{obs})^2}{\sum_i (y_i^{obs} - \overline{y^{obs}})^2},$$

and the Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^{pred} - y_i^{obs})^2},$$

where N is a number of compounds in the set, y_i^{pred} is the predicted value and y_i^{obs} is the experimentally observed value for data point i. The RMSE is expressed in the same units as the observed property values.

Results

Internal Validation

Fig. 9 illustrates the performance profile for the seven properties on the internal validation set for the Alchemite Imputation, Alchemite Virtual, Chemprop and the average of four QSAR models. The R^2 values for each model and property are also summarised in Table 1.

Table 1. Performance of the Alchemite Imputation, Alchemite Virtual, graph- and descriptor-based Chemprop and QSAR models on the internal validation set. For the QSAR models, the average performance is shown, with the standard deviation of the results for these models.

	Coefficient of Determination (R^2)						
	VP	**WS**	**ODT**	**I1**	**I2**	**I3**	**I4**
Alchemite Imputation	0.85	0.75	0.60	0.79	0.90	0.92	0.8
Alchemite Virtual	0.83	0.75	0.36	0.29	0.49	0.57	0.54
Chemprop (graph)	0.86	0.71	0.26	0.34	0.45	0.51	0.51
Chemprop (descriptors)	0.83	0.69	0.31	0.31	0.29	0.38	0.45
4 QSAR models (average)	0.84±0.05	0.73±0.02	0.33±0.08	0.32±0.08	0.40±0.10	0.46±0.10	0.50±0.06

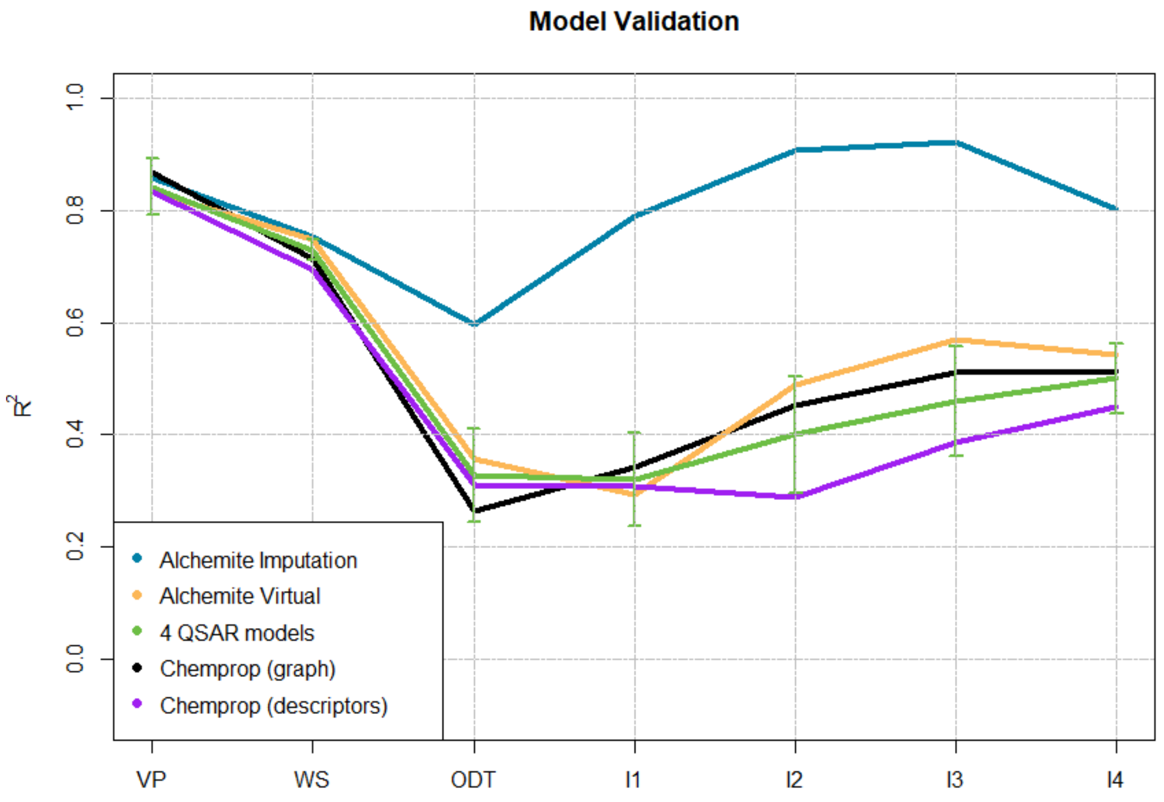


***Fig.* 9** *Coefficient of determination (R^2) profile of the seven properties on the internal validation set for Alchemite Imputation (blue), Alchemite Virtual (yellow), graph-based Chemprop (black), descriptor-based Chemprop (purple) and*

QSAR models (green). For the QSAR models, the average performance is shown, with the standard deviation of the results for these models

Independent Test

Fig. 10 shows the performance profile for the seven properties on the independent test set for the Alchemite Imputation model, Alchemite Virtual model, Chemprop and the four QSAR models. The R^2 values for each model are also summarised in Table 2.

Table 2. Performances of the Alchemite Imputation, Alchemite Virtual, graph- and descriptor-based Chemprop and QSAR models on the independent test set. For the QSAR models, the average performance is shown, with the standard deviation of the results for these models.

	Coefficient of Determination R^2						
	VP	**WS**	**ODT**	**I1**	**I2**	**I3**	**I4**
Alchemite Imputation	0.79	0.71	0.56	0.63	0.92	0.94	0.89
Alchemite Virtual	0.80	0.71	0.46	0.19	0.47	0.56	0.55
Chemprop (graph)	0.62	0.65	0.23	0.05	0.35	0.40	0.46
Chemprop (descriptors)	0.83	0.71	0.34	0.19	0.40	0.37	0.53
4 QSAR models (average)	0.80±0.07	0.70±0.02	0.29±0.04	0.2±0.03	0.32±0.08	0.52±0.02	0.53±0.02

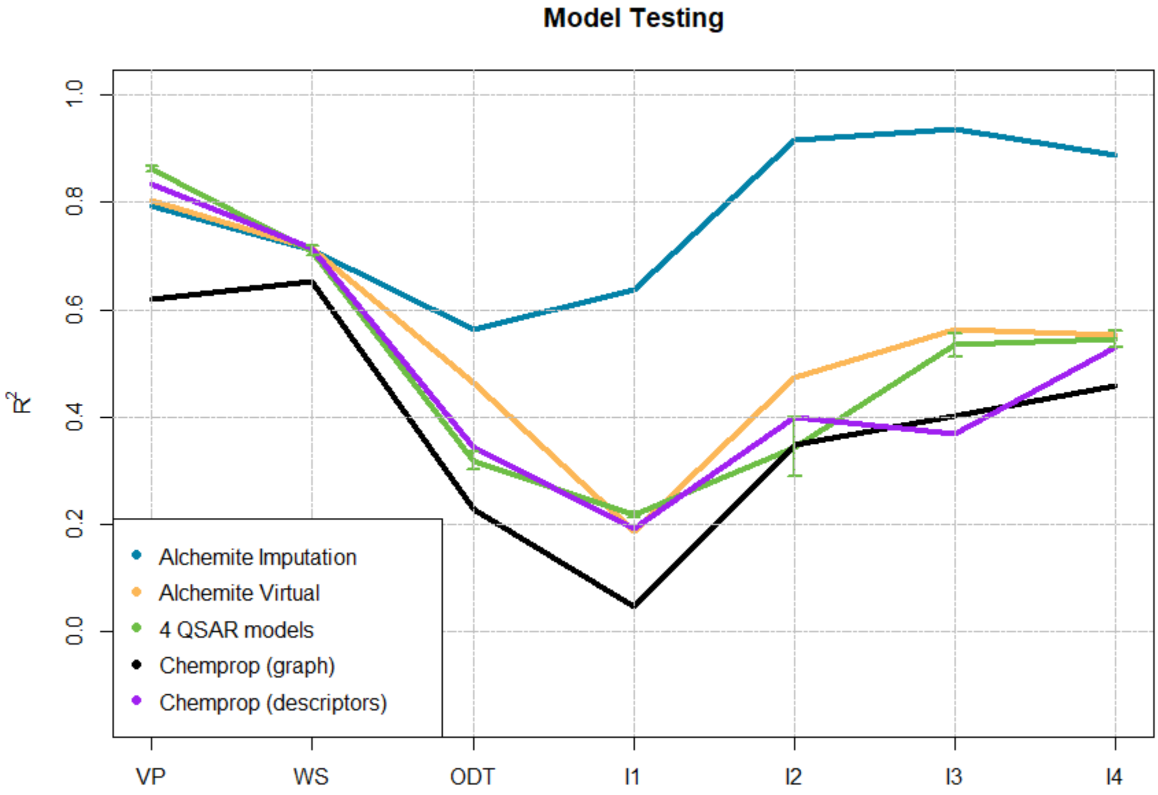


Fig. 10 *Coefficient of determination (R^2) profile of the seven properties on the external test set for Alchemite Imputation (blue), Alchemite Virtual (yellow), graph-based Chemprop (black), descriptor-based Chemprop (purple) and the best QSAR models (green). For the QSAR models, the average performance is shown, with the standard deviation of the results for these models*

Discussion

The results from the independent test are consistent with those from the internal validation and show that the Alchemite Imputation model significantly outperforms the Alchemite Virtual model, Chemprop and QSAR models for ODT and odor intensity assessments. On the internal validation set, the Alchemite Imputation model achieved excellent R^2 values of 0.79 or higher on all of the odor intensity endpoints. This was maintained for the independent test set for all of these endpoints, except I1, which dropped to 0.63. The improvement in R^2 for the Alchemite Imputation model over the next best model for each sensory endpoint was between 0.26 and 0.45.

As noted above, ODT is a particularly challenging property to predict, as demonstrated by an R^2 under 0.4 for the QSAR and Chemprop models for both validation and test sets. On the test set, the Virtual model outperformed the QSAR and Chemprop models in predicting ODT, achieving an R^2 of 0.46. A significant improvement is achieved by Alchemite Imputation, with an R^2 of 0.6 and 0.56 for the validation and test sets, respectively. This improvement in accuracy can also be seen in Fig. 11, where the observed values of ODT in the test set are plotted against the predicted values for the best QSAR and the Alchemite Imputation models. This shows that predictions made by the best QSAR model for this property are more scattered around the identity line when compared to the predictions made by the Alchemite Imputation model, as reflected in the difference in R^2.

Considering the models based only on chemical structure, the performance of the different modelling methods was closer. The Alchemite Virtual model slightly outperformed the other methods on the sensory properties, achieving an average R^2 across these properties of 0.45 on the validation set, compared with 0.41, 0.40 and 0.35 for the Chemprop (graph), Chemprop (descriptors) and averaged QSAR models, respectively. On the independent test set, the performance of the Alchemite Virtual was greater on ODT but similar to the averaged QSAR models on the other properties. However, the performance of the Chemprop (graph) model decreased between the validation and test sets, from an average R^2 of 0.41 to 0.30.

Notably, the R^2 for property I1 is lower for all models in the independent test set when compared with the internal validation set. This may be because I1 represents a test at the lowest intensity and, hence, may be subject to greater inter-individual variability between test subjects. Nonetheless, the Alchemite Imputation model retains a significant improvement in predicting I1, achieving an R^2 of 0.63, compared with less than 0.2 for the other models.

All models accurately predict the physicochemical properties, VP and WS, with R^2 in the ranges 0.69 – 0.86 and 0.70 – 0.83 respectively on the validation and test sets. The exception was the performance of Chemprop (graph) on the test set, for which the performance dropped to 0.62 and 0.65, respectively. It is interesting to note that Alchemite Imputation does not benefit the prediction of physicochemical properties to the same extent as ODT and odor intensity assessments. In part, this is because the physicochemical properties are already well-predicted by models based only on structural features. It also illustrates the benefits that imputation gains from leveraging correlations between experimental endpoints, even when the available data are sparse. The largest benefits are gained where there are strong correlations between experimental endpoints, for example, between physicochemical properties and sensory properties. The data generated in earlier, less expensive assays can substantially improve the predictions for more expensive and hence more sparsely sampled experimental endpoints.

The poorer transferability of the Chemprop (graph) model relative to the Alchemite and QSAR models is notable for both the sensory and physicochemical properties. The QSAR, Chemprop (descriptors) and Alchemite models are based on the same chemical descriptors as the Chemprop model; therefore, the most likely explanation is that the graph features identified by the GCNN from the training set do not adequately capture the SAR of the test compounds.

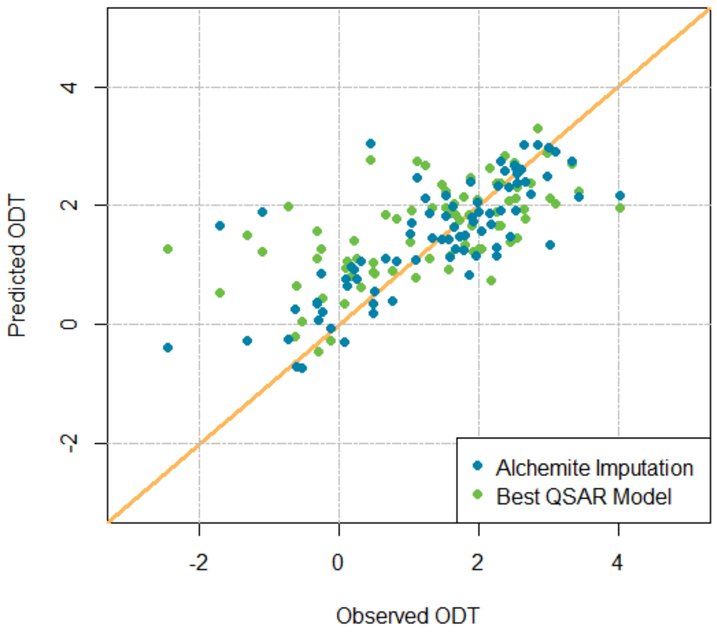


Fig. 11 Predicted values plotted against observed values for ODT for the independent test set, using the Alchemite Imputation model (blue) and the best QSAR model (green)

Analysis of Uncertainties

The Alchemite Imputation model shows exceptionally high performance on the odor intensity assessments, I2, I3 and I4, with R^2 values of approximately 0.9 for these properties on the independent test set. Moreover, these properties are predicted with high confidence by this model, as shown by the small error bars in Fig. 12.

In Fig. 13, error bars generated by the Alchemite Imputation model for ODT reflect the estimated uncertainty (one standard deviation) in each predicted value. An accurate estimate of the uncertainty in each prediction enables us to focus on the most accurate results by disregarding those with the highest uncertainty. We would expect the remaining, more confident values to have higher accuracy. Fig. 14 shows the impact of discarding the predictions in increasing order of confidence (i.e. the predictions with the largest error bars are discarded first). The RMSE is plotted on the y-axis of the graph, such that lower values indicate more accurate predictions. The blue line shows, subject to some statistical fluctuations, that as the least confident predictions are removed, the RMSE falls sharply, confirming the expected behaviour. The black points show the average Alchemite error bar for the remaining predictions, demonstrating a good agreement between the estimated uncertainties and actual errors in prediction.

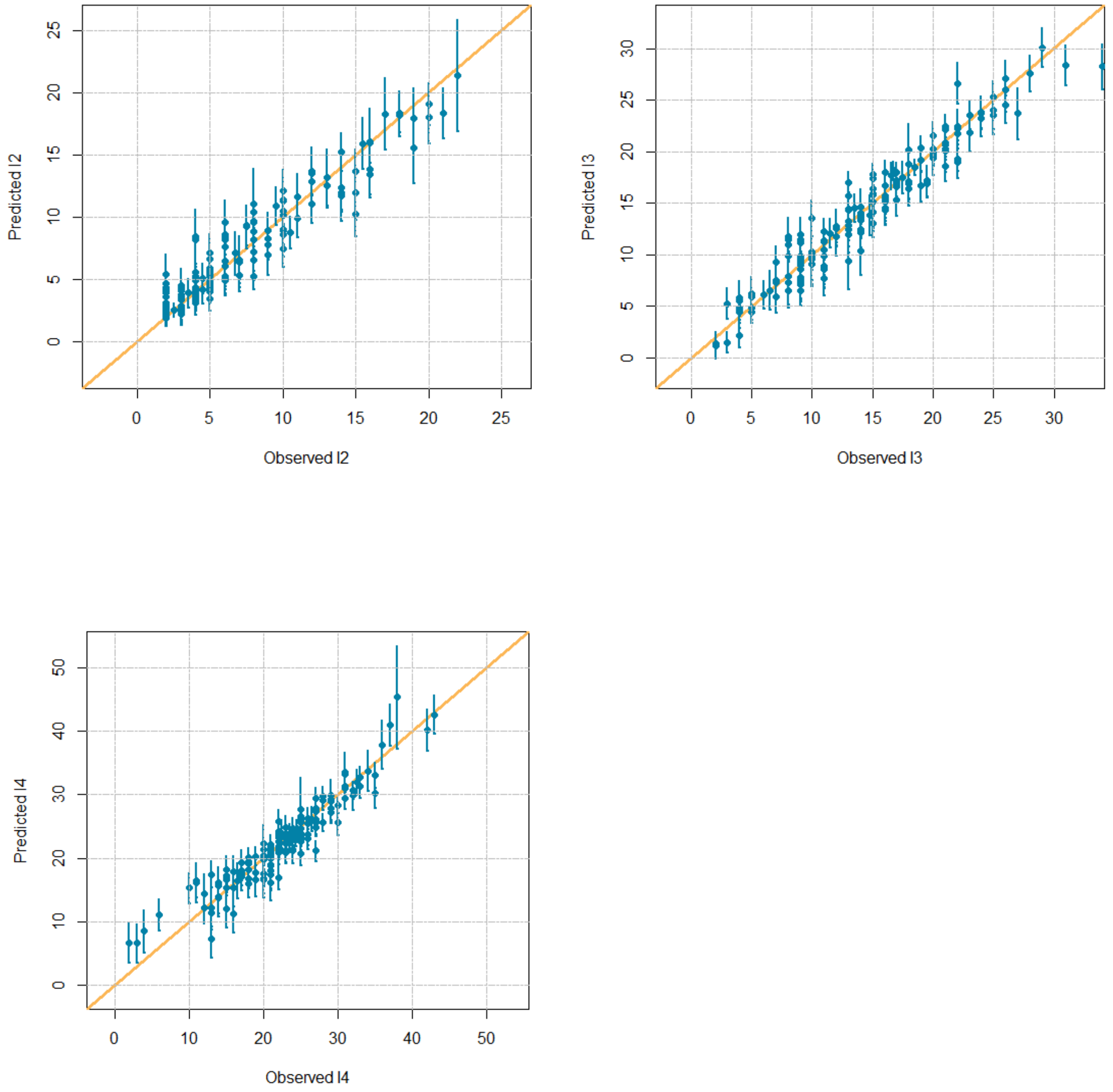


***Fig. 12** Observed values plotted against imputed values for* odor intensity properties, I2, I3 and I4*, with error bars on predictions*

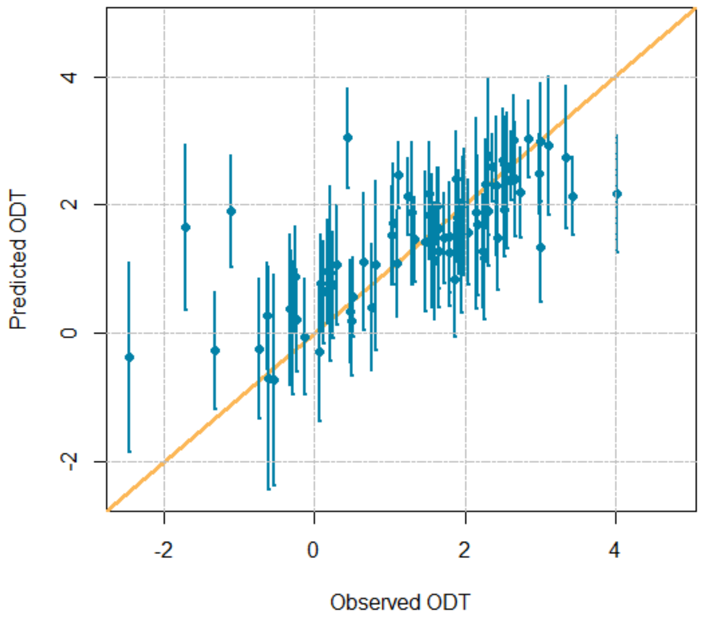


Fig. 13 *Predicted values plotted against observed values for ODT for the independent test set, using the Alchemite Imputation model. Error bars on predictions represent the estimated uncertainty (one standard deviation) in each prediction*

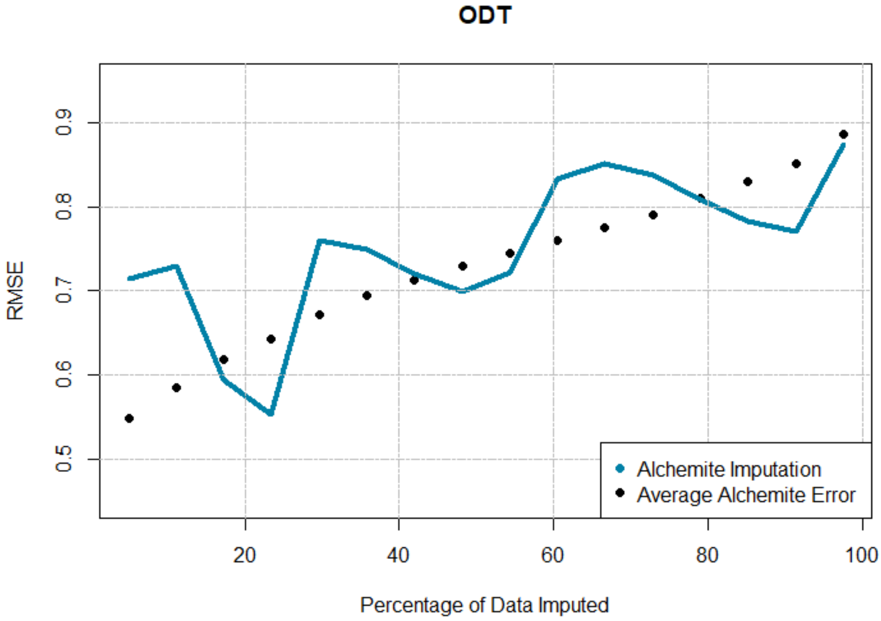


Fig. 14 *Percentage of imputed values for ODT after removing the predicted values with greatest uncertainty against Root Mean Square Error (RMSE) (blue) and the average error bars from the Alchemite Imputation model for the remaining predictions are shown as black dots*

Analysis of Chemical Similarity Versus Accuracy

We assessed the similarity of each test compound and the training set by calculating the correlation between the normalized StarDrop descriptors of the test compound and each compound in the training set. The five-nearest-neighbour (5NN) similarity was calculated by taking the average correlation between the five training compounds with the highest similarity to the test compound. The range of 5NN similarity for the compounds in the test set is 0.69 to 0.99.

In Fig. 15, we compare the absolute error in ODT predictions made by the Alchemite imputation and QSAR models for the 20% of the compounds in the test set that are least similar to those in the training set (the lowest 5NN similarity). This shows that the least-similar test compounds were predicted more accurately by Alchemite imputation than the best QSAR model for this property.

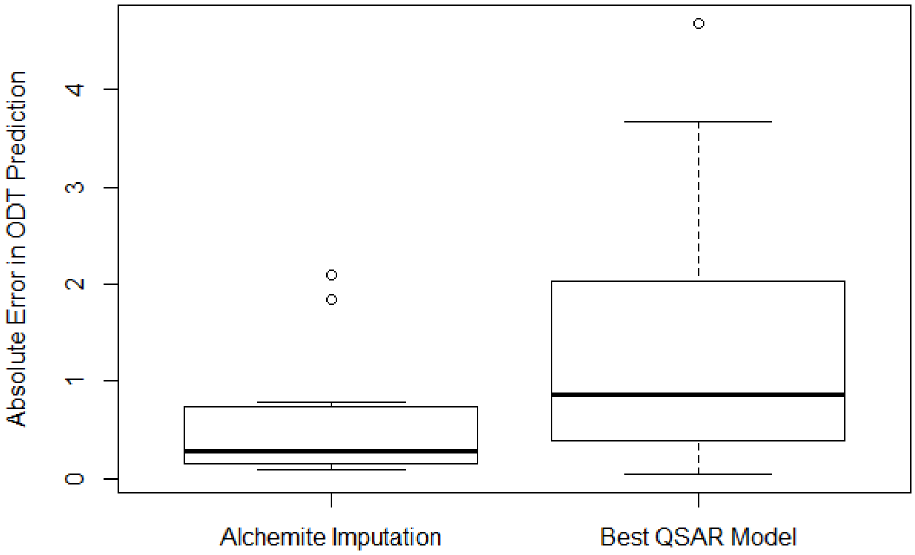
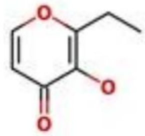
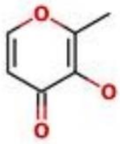
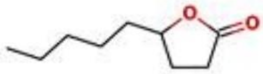
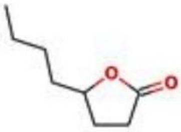
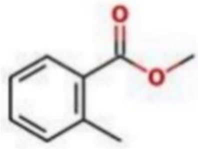
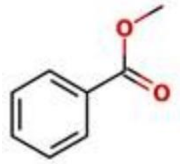


Fig. 15 *The absolute error in prediction of ODT by the Alchemite Imputation model and the best QSAR model on the 20% of the test compounds that are most dissimilar to the training set*

Activity cliffs are compound pairs that have high structural similarity but show significantly different activity or property values. Table 1 shows examples of activity cliffs for the observed ODT between compounds in the training and test sets. In these examples, we can see that the Alchemite Imputation model predicted the ODT values of the activity cliffs more accurately than the best QSAR model for this property.

These results show an advantage of imputation over prediction based only on chemical structure. The Alchemite Imputation model goes beyond a nearest-neighbour approach in descriptor space by exploiting correlations between experimental endpoints that are carried across into endpoint-descriptor space.

Table 3. Examples of activity cliffs in observed ODT between training and test set compounds. For each compound, the observed ODT values are shown with the test compound predictions made by the Alchemite Imputation model and the best QSAR model.

Structure of Nearest Neighbour in Training Set	Observed ODT of Nearest Neighbour in Training Set	Test Compound	Observed ODT of Test Compound	Predicted ODT by the Alchemite Imputation model	Predicted ODT by the best QSAR model
	0.93		-0.6	-0.7	1.43
	1.93		-0.12	-0.05	0.98
	0.52		1.99	1.89	0.89

Conclusions

Sensory property data are noisy due to inter-individual variability between test subjects and present a particular challenge for predictive modelling. In this paper, we have demonstrated an advantage provided by Alchemite™ deep learning imputation over conventional QSAR and multi-target GCNN models in this challenging application. We observed a considerable advantage when using sparse experimental data to 'fill in' missing values. ODT is an important property for the development of flavors and fragrances and poses a particular challenge for predictive modelling. However, the use of physicochemical property and odor intensity assessment data, even when incomplete, dramatically improves the accuracy of prediction, enabling us to use data from less expensive, early experiments to make a better selection of compounds for more costly ODT studies, saving experimental time and cost. The analogy in pharma would be the use of early high-throughput assays requiring small samples of novel compounds to make better selections for downstream, more expensive experiments.

In addition, it is notable that deep learning in this context provides a valuable improvement in predictive accuracy, even with a relatively small data set. Typically, deep learning methods require tens of thousands of data points to offer a significant improvement over conventional machine learning methods for QSAR.

For virtual compounds with no pre-existing experimental data, Alchemite™ offered a smaller but still valuable improvement over the other methods using only chemical descriptors as input.

We have also highlighted the ability of Alchemite™ to focus attention on the most accurately predicted values by using a robust estimate of the uncertainty in each individual prediction, in contrast with other machine learning methods, which struggle to estimate errors reliably [37]. It is essential to take into consideration uncertainties in both predicted and experimental data when making decisions about the selection of compounds for progression. This can help to focus synthetic and experimental resources where the probability of success is highest and avoid missed opportunities caused by inappropriately discarding compounds due to inaccurate predictions [36].

References

[1] M. Kass, M. Rosenthal, J. Pottackal and J. McGann (2013) Fear Learning Enhances Neural Responses to Threat-Predictive Sensory Stimuli. Science 342:1389-1392

[2] E. Block, (2018) Molecular Basis of Mammalian Odor Discrimination: A Status Report. J. Agric. Food Chem. 66:13346-13366

[3] J. McGann, (2017) Poor Human Olfaction is a Nineteenth Century Myth. Science 356:

[4] M. Genva, T. Kemene, M. Deleu, L. Lins and M. Fauconnier (2019) Is It Possible to Predict the Odor of a Molecule on the Basis of its Structure?. Int. J. Mol. Sci. 20:3018

[5] L. Buck, (2000) The Molecular Architecture of Odor and Pheromone Sensing in Mammals. Cell 100:611-618

[6] K. Nara, L. Saraiva, X. Ye and L. Buck (2011) A Large-Scale Analysis of Odor Coding in the Olfactory Epithelium. J. Neurosci. 31:9179-9191

[7] R. Araneda, A. Kini and S. Firestein (2000) The molecular receptive range of an odorant receptor. Nature Neurosci. 3:1248-1255

[8] Y. Yeshurun and N. Sobel (2010) An odor is not worth a thousand words: from multidimensional odors to unidimensional odor objects. Annu. Rev. Psychol. 61:219-41

[9] F. Zufall and T. Leinders-Zufall (2000) The Cellular and Molecular Basis of Odor Adaptation. Chem. Senses 25:473-481

[10] P. Kraft, (2018) The Odor Value Concept in the Formal Analysis of Olfactory Art. Helvetica 102:e1800185

[11] A. Dunkel, M. Steinhaus, M. Kotthoff, B. Nowak, D. Krautwurst, P. Schieberie and T. Hoffmann (2014) Nature's Chemical Signatures in Human Olfaction: A Foodborne Perspective for Future Biotechnology. Angew. Chem. Int. Ed. 53:7124-7143

[12] K. Rossiter, (1996) Structure-Odor Relationships. Chem. Rev. 96:3201-3240

[13] P. Kraft, J. Bajgrowicz, C. Denis and G. Frater (2000) Odds and Trends: Recent Developments in the Chemistry of Odorants. Angew. Chem. Int. Ed. 39:2980-3010

[14] P. Kraft, V. Di Cristofaro and A. Jordi (2014) From Cassyrane to Cashmeran – The Molecular Parameters of Odorants. Chem. & Biodiv. 11:1567-1596

[15] W. Zhan, F. Doro and M. Teixeira (2019) A rapid approach to optimize the design of fragrances for fabric care products. Flavor Frag. J. 35:167-173

[16] C. Trimmer, A. Keller, N. Murphy, L. Snyder, J. Willer, M. Nagai, N. Katsanis, L. Vosshall, H. Matsunami and J. Mainland (2019) Genetic variation across the human olfactory receptor repertoire alters odor perception. PNAS 116:9575-9480

[17] M. Teixeria, L. Barrault, O. Rodriguez, C. Carvalho and A. Rodrigues (2014) Perfumery Radar 2.0: A Step toward Fragrance Design and Classification. Ind. Eng. Chem. Res. 53:8890-8912

[18] L. Ruddigkeit, M. Awale and J. Reymond (2014) Expanding the fragrance chemical space for virtual screening. J. Cheminformatics 6:27

[19] J. Medino-Franco, K. Martinez-Mayorga, T. Peppard and A. Del Rio (2012) Chemoinformatic Analysis of GRAS (Generally Recognized as Safe) Flavor Chemicals and Natural Products. PLOS One 7:e50798

[20] E. Brenna, C. Fuganti and S. Serra (2003) Enantioselective perception of chiral odorants. Tetrahedron Asymmetry 14:1-42

[21] S. Wold, M. Sjöström and L. Eriksson, "Partial Least Squares Projections to Latent Structures (PLS) in Chemistry," in *Encyclopedia of Computational Chemistry*, P. Schleyer, N. Allinger, T. Clark, J. Gasteiger, P. Kollman, H. Schaefer III and P. Schreiner, Eds., Chichester, UK., Wiley, 1998, 2006-2022.

[22] L. Breiman, (2001) Random Forests. Machine Learning 45:5-32

[23] N. Cristianini and J. Shawe-Taylor, An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge, U.K.: Cambridge University Press, 2000.

[24] P. Hunt, L. Hosseini-Gerami, T. Chrien, J. Plante, D. Ponting and M. Segall (2020) Predicting pKa Using a Combination of Semi-Empirical Quantum Mechanics and Radial Basis Function Methods. J. Chem. Inf. Model. 60:2989-2997

[25] O. Obrezanova, G. Csanyi, J. Gola and M. Segall (2007) Gaussian Processes: A Method for Automatic QSAR Modelling of ADME Properties. J. Chem. Inf. Model. 47:1847-1857

[26] N. Sadawi, I. Olier, J. Vanschoren, R. van Rijn, J. Besnard, R. Bickerton, C. Grosan, L. Soldatova and R. King (2019) Multi-task learning with a natural metric for quantitative structure activity relationship learning. J. Cheminform. 11:68

[27] E. Feinberg, D. Sur, Z. Wu, B. Husic, H. Mai, Y. Li, S. Sun, J. Yang, B. Ramsundar and V. Pande (2018) PotentialNet for Molecular Property Prediction. ACS Cent. Sci. 4:1520-1530

[28] Y. Nozaki and T. Nakamoto (2018) Predictive modeling for odor character of a chemical using machine learning combined with natural language processing. PLOS ONE 13:e0198475

[29] T. Gunaratne, C. Gonzalez Viejo, N. Gunaratne, D. Torrico, F. Dunshea and S. Fuentes (2019) Chocolate Quality Assessment Based on Chemical Fingerprinting Using Near Infra-red and Machine Learning Modeling. Foods 8:426

[30] A. Dagan-Wiener, I. Nissim, N. Ben Abu, G. Borgonovo, A. Bassoli and M. Niv (2017) Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. Sci. Rep. 7:12074

[31] L. Shang, C. Liu, Y. Tomiura and K. Hayashi (2017) Machine-Learning-Based Olfactometer: Prediction of Odor Perception from Physicochemical Features of Odorant Molecules. Anal. Chem. 89:11999-12005

[32] B. Irwin, S. Mahmoud, T. Whitehead, G. Conduit and M. Segall (2020) Imputation versus prediction: applications in machine learning for drug discovery. Fut. Drug Discov. 2:

[33] T. Whitehead, B. Irwin, P. S. M. Hunt and G. Conduit (2019) Imputation of Assay Bioactivity Data Using Deep Learning. 59:1197-1204

[34] B. Irwin, J. Levell, T. Whitehead, M. Segall and G. Conduit (2020) Practical Applications of Deep Learning to Impute Heterogeneous Drug Discovery Data. J. Chem. Inf. Model. 60:2848-2857

[35] B. Irwin, T. Whitehead, S. Rowland, S. Mahmoud, G. Conduit and M. Segall (2021) Deep Imputation on Large-Scale Drug Discovery Data. App. AI Lett. 2:e31

[36] M. Segall and E. Champness (2015) The challenges of making decisions using uncertain data. J. Comp.-Aided Mol. Des. 29:809-816

[37] L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay and C. Coley (2020) Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. J. Chem. Inf. Model. 60:3770-3780

[38] P. C. Verpoort, P. MacDonald and G. J. Conduit (2018) Materials Data Validation and Imputation with an Artificial Neural Network. Comput. Mater. Sci. 147:176-185

[39] J. Bergstra, R. Bardenet, Y. Bengio and B. Kégl, "Algorithms for Hyper-Parameter Optimization," in *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing*, Red Hook, NY, 2011.

[40] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins and D. D. Cox (2015) Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization. Comput. Sci. Discov. 8:014008

[41] Optibrium Ltd., “StarDrop,” [Online]. Available: https://www.optibrium.com/stardrop. [Accessed 27 September 2021].

[42] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen and R. Barzilay (2019) Analyzing Learned Molecular Representations for Property Prediction. J. Chem. Inf. Model. 59:3370-3388

[43] G. Green, P. Dalton, B. Cowart, G. Shaffer, K. Rankin and J. Higgins (1996) Evaluating the 'Labeled Magnitude Scale' for measuring Sensations of Taset and Smell. Chemical Senses 21:323-334

[44] ASTM International, "ASTM E679-19, Standard Practice for Determination of Odor and Taste Thresholds by a Forced-Choice Ascending Concetration Series Method of Limits," ASTM International, West Conshohocken, PA, 2019.

[45] G. McLachlan and T. Krishnan, The EM Algorithm and Extensions, 2nd Edition, Noboken, New Jersey: Wiley, 2008.

[46] D. Jones, (2001) A Taxonomy of Global Optimization Methods Based on Response Surfaces. J. Global Opt. 21:345-383