Interpretable Machine Learning for the Structure Odour Relationship

Denis Kealy
School of Computing (Dublin City University)
MCM Practicum Project
denis.kealy2@mail.dcu.ie

***Abstract*—Highly predictive models are widespread in many disciplines but there exist a multiplicity of cases where the designer of the model cannot explain the decision made by the algorithm. Explainable models can give a justification relating to a given output, known as a local explanation; or give a representation of the whole model space, known as a global explanation. These explanations can range from text and images to interactive applications. This investigation compares different machine learning methodologies based on their performance and interpretability/explainability and we subsequently analyse the explanations of our system to better understand population-level human olfaction. We have trained models to perform a state of the art regression task of predicting the human-perceived odour of a molecule using only the molecular properties as descriptors. This problem is known as the Structure Odour Relationship and despite the efforts of many widely disciplined researchers there still exits no generalised solution or theory to this puzzling relationship. This approach of statistical analysis combined with machine explanations will be a very powerful tool for understanding biological or even other complex systems.**

***Index Terms*—eXplainable Artificial Intelligence (XAI), Machine Learning, Explainable Models, Olfaction, Digital Scent, Structure Odour Relationship (SOR)**

I. Introduction

The prevalence of machine learning methodologies in industry, academia, and the government has resulted in concerns about our ability to comprehend the decisions being made by these algorithms [1]. Legislation has been put in place, and research and debate have ensued as machine learning becomes ever more present in our daily lives. From personalised advertisements and news, to credit rating systems, policy quotation systems and now autonomous vehicles [2] - questions have been raised as to the morality and interpretability of these systems [3].

Simple models such as linear methods lend themselves to being interpretable [4] while more complex algorithms, such as ensemble or deep learning models, tend towards obfuscation. Measuring performance is well understood and is indeed central to the mechanics of machine learning itself; learning requires an objective measure of performance on a task, and optimisation of this performance is core to how an algorithm learns from data. Measuring interpretability or explainability, however, is more abstract. This is essentially a measurement of the successful transmission of a concept to a human mind. Images, textual information, and plain numbers can all relay their own subsets of information more reliably than one another and, as such, all three are discussed in this investigation. A recent paper [5] published by researchers at Google discusses the combination of dimensionality reduction techniques with interactive visualisations to interpret deep learning systems. The accompanying website [6] gives examples of such interactive explanations which are quite illuminating, along with open source code to see how it all works. Both Google and DARPA [7] emphasise the importance of Human-Computer Interaction (HCI) for machine learning and knowledge discovery/data mining and the above website conveys the usefulness of such systems elegantly, even to a layperson.

Why should we care about Explainable models? Legally, explainability is now mandated by law. Article 21 and Article 22, in particular, of the GDPR regulations [8] are aimed at providing EU citizens with more control over personalisation or decision-making algorithms which process their data. Article 22 allows the data owner to ask for a justification of any decisions made about them of a "legal or similar" matter. For some decision-making algorithms, it is not entirely possible to explain why a decision was made without explicit knowledge of the math/concepts involved. A statistician will undoubtedly understand your regression model but a member of the public may not; a machine learning expert may never gain insight into how a deep learning algorithm makes a decision just by examining the model properties. For these reasons, it is imperative that certain decision-making algorithms be explainable - and not just to an expert (this may seem obvious but it is a critical distinction). Arguments have been put forth by the medical community that this could greatly improve the trust in, and transparency of their work [9]. Reason codes are another existing use case for these explanations where credit checking and mortgage qualification companies have to provide a justification for a given rating.

In this investigation, we have applied and compared techniques of interpreting models mainly for the purpose of model recovery and discovering population-level relationships between molecular features and their perceived scent. A working theoretical model of the human olfactory system is a worthy, yet, an elusive long-term goal of many researchers in different disciplines.

So why should we care about scent? Our sense of smell is uniquely linked to certain parts of our brain; Smell is anatomically interlinked with higher centres of the brain in ways that no other sense are [10]. Unravelling the mystery of scent will undoubtedly bring advances in many areas of medicine; Smell has been closely linked to memory, and harnessing the power of scent could help people with memory issues such as dementia or Alzheimer's patients [11]. Some cancers such as prostate [12], ovarian [13], and lung cancer [14], as well as other diseases such as arthritis [15], have been shown to be cheaply and effectively diagnosed using an electronic nose.

We currently understand how light is encoded in the human brain; The perceptual space for human colour vision has three dimensions and every colour sensation can be fully characterised by three numbers, namely the intensity of the primary colours that match it. We now know that colour-vision is based on three kinds of cone photoreceptors in the retina that differ in their sensitivity to the wavelength spectrum of light [16]. This understanding enables us to encode this as digital information and package it as images and videos in the form of 1's and 0's; No such understanding exists for olfactory information. Understanding how our brains process olfactory information is extremely difficult due to the number and variation of odorants, the number of receptor neurons, the nature by which they react with odorant molecules, the non-linear mappings of receptors to glomeruli, and the consequent flow of information to higher areas of the brain such as the amygdala and the hippocampus [10]. Advances in artificial olfaction [17], such as new sensors and models, have set up electronic noses for revolutionary environmental and medical uses [18].

Through interpreting a robust population prediction model our goal is to enable researchers to follow systematic approaches to encode smell at an objective level. Objective as in the opposite of subjective, ignoring the individualised differences in perception which is due to an underlying anatomical difference between individuals. This approach is much like how a regular RGB screen works well for most people (as most of our eyes function similarly) but doesn't work seamlessly for those with colour blindness; which, similarly to olfaction, is due to an underlying anatomical difference at the individual level. There already exists a system similar in goals to our proposal [19] which is specifically designed to tackle the SOR problem; whereas our methodology is not strictly tied to the problem being studied and is a more generalised approach that can be explored through a variety of applications.

This approach to modelling and interpreting population-level responses could be the shortest path to digitising smell and this forms the primary goal of our investigation - to successfully train and examine models that can predict population responses to smell. This feat has been recently accomplished by teams participating in a DREAM challenge [20], published in 2017 where teams were able to predict the smell of an unseen molecule with unprecedented accuracy. In our investigation, we leverage these previous efforts to train our own predictive models for the task. Finally, we interpret our predictive models using traditional and novel means. We compare the interpretation of linear models with that of XAI assisted random forest explanations. For these ensemble algorithms, we use Shapley values, originating from game theory, to measure the contribution of the individual features to the predicted output. This constitutes a novel and state of the art contribution to the body of work on the Structure Odour Relationship.

II. Background

A. Explainable Models

There has been an understandable race for performance when it comes to the field of machine learning but explaining why an algorithm works has now become a focus of research. In recent months, through a haze of ambiguity, definitions have begun to crystallise in the community [3]. Particularly, one paper from MIT where the authors have discussed the definitions of these terms in great detail while providing a review of all the recent work in this area [21]. They are critical of some definitions - particularly the way interpretability and explainability have been used interchangeably; the authors argue that there are important reasons to distinguish between them. Explainable models are interpretable by default, but the reverse is not always true - This definition, along with the author's points on interpretability vs completeness, help to solidify the concepts of explainable models that have risen in recent publications.

An explanation is complete when it fully describes the exact behaviour of the system. A measure of completeness would be how accurate the explanation is across all circumstances or possible inputs. For example, a complete explanation for linear regression or a neural network would be given by the all the mathematical operations and parameter weights; This would constitute a totally complete explanation. The trade-off between complete and interpretable explanations can be varied to gain even further insight into a black box system. The user should be notified of the level of completeness of an explanation, at any level, because of the real danger of creating a persuasive, rather than interpretable AI [22]. Life-critical systems, in particular, have many constraints placed upon where and how these algorithms can be used. A formal methods approach to designing such systems is preferred over an approximation based interpretability method; complete explanations are necessary for such domains.

There are many approaches to interpreting a machine learning model. Model-specific explanations are tailored to the type of algorithm in question and as such have good performance in general and are preferable to the alternative - model agnostic explanations. Model agnostic explanations tend to be approximation-based i.e where a simpler, more naturally interpretable, model is trained on the outputs of a more complex model [1].

Another salient distinction in the interpretation of models is local vs global interpretations. Locally we would try to explain a single prediction or a set of similar predictions e.g. the cases representing the boundaries or typical examples of our data i.e. why is this customer the riskiest in the test set or why does this molecule have the highest ratings for a 'Garlic' odour. Global explanations try to explain across the entire model input space what influences predictions the most. This type of explanation can help us figure out population-level trends in the odour sensing profiles of individuals, although not much of this data exists.

B. Olfaction - Our Sense of Smell

Our sense of smell is wonderfully complex and naturally more subjective than our other sense. This subjectivity of smell occurs, in part, due to the involvement of higher centres of the brain which consider learned behaviour and past experiences before arriving at our final conscious perception of an ”odorant” [23] [24]. Although the studies cited don’t investigate the human olfactory system, there are remarkable similarities between species regarding the neural process of sensing olfactory cues in our environments [25]. Vocabulary and culture also play a role in our personal sense of smell and some research shows that our understanding of smell may be linked to our language processing capabilities [26]. One example of this is the ”Jahai” language and culture, from the Malay Peninsula in Southeast Aisa, who dedicate a significant portion of their vocabulary to smells - a trait shared by the Maniq people of Thailand. These words for smells, in Jahai, also serve as the base of their vernacular of colour [27].

An ”odorant” is a molecule which has been shown to produce a perception of smell when it interacts with our odorant receptors (OR); In other words, humans can smell and identify this molecule in isolation. Since the work of Axel & Buck in 1991, the prevailing theory of Olfaction has been that the structure of an odorant has a significant bearing on the type of smell we perceive, referred to since in the literature as the shape theory of olfaction [28]. Axel & Buck discovered that ORs in mammals are a large subfamily of G-protein-coupled receptors (GPCR). These types of receptors are seven-transmembrane receptors which have been studied and shown to activate through ligand/receptor binding. This interaction involves a lock and key mechanism where the shape of the molecule determines the receptor it binds with. Most odorant molecules also undergo protein-ligand binding in the nose for the purpose of becoming water soluble before entering the mucus [10] where they can bind with our odorant receptors. These bindings are only but the start to a long, convoluted chain of neural processing where eventually a perception is formed; perhaps only after context, acquired knowledge, and other senses are considered [29]. Currently, the literature has no answer for where and how our final conscious perception of smell is derived.

Although the ligand/receptor binding seems to explain and give a basis for the mapping of this relationship there are some edge cases and inconsistencies that cannot seemingly be explained by shape alone. Some molecules are structurally similar and have different percepts [30]. Some odorant molecules are structurally very different but have a similar perceived scent. These quandaries sparked debate for decades and the matter is still unsettled. Recent work from the biological side has shifted to gene expression experiments, new imaging techniques for in-vivo experimentation [31], gene sequencing and in-silica modelling of olfactory systems [32].

C. Olfaction - The data

The original goal of this project was to collect and assimilate multiple SOR datasets for the purposes of robustly testing our models across data sources which have different data collection methods and data representations. Although classification of molecules based on their perceived odour was the goal set out in the initial literature review, the availability of datasets and the disparity between data representations and collection methods was deemed too great. As such our modelling efforts are based on the most recent and comprehensive SOR dataset available at present [33].

This dataset consists of 55 subject responses to 480 molecules at 2 dilutions. 20 of such molecules solutions were double sampled to measure the within-individual variability of responses. This gives 2 * 55 * (480+20) = 55,000 observations. The perceptual ratings assigned to these molecules are in the form of an odour panel. An odour panel is a list of set descriptors, usually denoting a category or group of smells. The perceptual dimensions of smell accounted for in this data are given in Table 1.

Our independent variables here are the molecular descriptors which describe the structure of the molecules that were tested in 2014. The dependant variables here are the perceptual ratings (from 0-100) that the subjects of the above experiment could provide such as ”Fruit”, ”Fish”, ”Garlic” & ”Spices”.

Intensity	Pleasantness	Edible	Bakery
Sweet	Fruit	Fish	Garlic
Spices	Cold	Sour	Burnt
Acid	Warm	Musky	Sweaty
Urinous	Decayed	Wood	Grass
Flower	Chemical	Familiarity	-

Table 1. Available Odour Ratings (from 0-100)

The substantial noise in the data, due to individual subjectivity, is only compounded by the differences in how we personally use language and the inherent ambiguity/subjectivity of human languages in general. The within-individual variation combined with the subjective noise place a theoretical upper limit on the predictions possible from this data [20]. The quality of the best-performing models varied greatly across attributes and as such no team performed well with respect to all targets.

III. Methods

A. Data Selection & Collection

In the initial stages of the project, we scraped data from online public databases using python scripts. One dataset was obtained from Flavornet, an online database containing odour profiles of individual molecules. This dataset was analysed and findings from a previous study were verified [34]. Network Analysis helped to identify the differences in the structure of these smell networks and to demonstrate their significant difference from a random network [35]. Another factor influencing the choice of the dataset was the difference in targets - categorical and continuous. The online databases I initially worked with were organised categorically whereby a molecule is assigned a number of perceptual categories or tags, this being suited to multi-label classification; similar to multi-label topic modelling in document processing. The most recent SOR dataset, however, included continuous ratings from 0-100 for each perceptual category or target, this data obviously being more suited to regression.

Compounding these incompatibilities are the differences between the layman perception of smell and that of a trained expert. The online databases are compiled by experts in the field of fragrance discovery and perfumery and as such an entirely different vocabulary forms with experts using such descriptors as 'green', 'blue', and 'ether' - whereas laymen tend to use objects and nouns over conceptual notions such as colours. The proposed transformation of either dataset would be an obvious way to introduce bias into our dataset and we would either lose or distort the information transforming one representation to another. For scope and complexity purposes we have chosen to use only the most comprehensive SOR data available for training and testing our models.

The dataset chosen for final modelling and interpretation was the Keller & Voshall dataset collected in 2014 which was used as a basis for the 2017 DREAM Olfaction Machine Learning Challenge. This gives us an exact target for our regression modelling, which we can use to determine if our models are worth interpreting for the purposes of investigating olfaction.

B. Data Exploration

Data exploration and analysis were performed in early stages to get an understanding for the data at hand. Using the methods of previous researchers we explored different strategies for handling missing values in both the perceptual and the molecular descriptors. Masking NaNs, median, mean & zero imputation were explored and their effects on the overall statistics of the resulting dataset were investigated. Median imputation based on the chemical ID and dilution of the observation was the most successful imputation method in the literature and in my own experiments - however masking NaNs and calculating the mean observations across subjects was the most effective strategy for predicting population perception - this was the approach taken by the winning team of the population sub-challenge.

The distributions of the molecular descriptors were also investigated. Many machine learning models cannot handle outliers well and the molecular descriptors (after standardisation) contained many outliers which hinder learning performance. As such a transform such as cube root or log has been shown to improve predictions on this task. The effects of these transforms were studied in the data exploration phase and tested in the machine learning evaluation stage (available in source code).

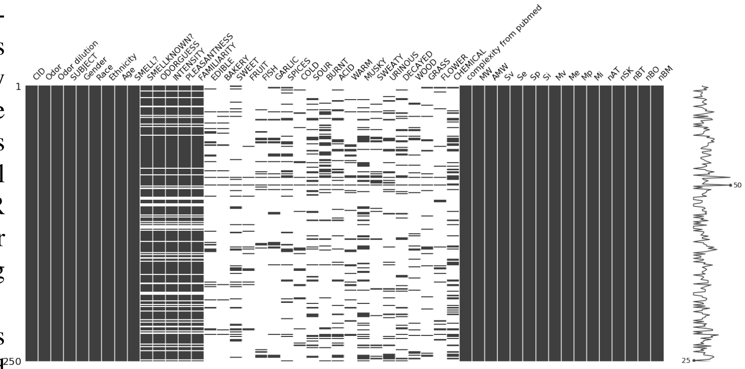

Fig. 1. Missing data distribution across our descriptive (rightmost) and target (middle) features; leftmost columns are information on the subjects.

Finally, an initial investigation into the dimensionality of this problem was conducted and compared with previous literature. Many such attempts have been made to study the dimensional properties of this perceptual space. This area has been heavily investigated and a review in 2013, admittedly before this DREAM challenge took place, criticised some of the methods used, conclusions drawn thus far in the literature [36]. Disentangled representations such as the variables produced by PCA/NMF may be naturally less interpretable than the original variables but their inclusion in the machine learning process need not complicate the resulting interpretation. However, both of the winning teams in each subchallenge (individual & population perception) used the raw (transformed and scaled) feature values without any form of dimensionality reduction. For these reasons, I have focussed on other areas for improving the predictions of our models. It should be noted that some teams aggregated the outputs of models trained on principal components with models trained on the raw features with varying success.

C. Preprocessing

Three files were used to construct the training and test sets. These consisted of two files containing the information on molecules, our X values, and subject responses, our Y values, and another file for optimally splitting the training and test set. The molecular descriptors were obtained for 476 molecules of the 480 using the DRAGON software. These were provided to the competitors and were obtained for this experiment from a public repository [20]. These are our descriptive features and apart from a column denoting 'Odor Dilution' this is the only information our models can use for prediction. These features were standardised and scaled, and subsequently visualised and evaluated using different techniques.

The best results were achieved using the mean responses of subjects across chemical ID and "Odor dilution" to predict population mean responses while masking missing values in the calculation. This reduced our observations from nearly 55,000 to less than 1,000 observations of mean perception which significantly reduced training times while retaining predictive accuracy.

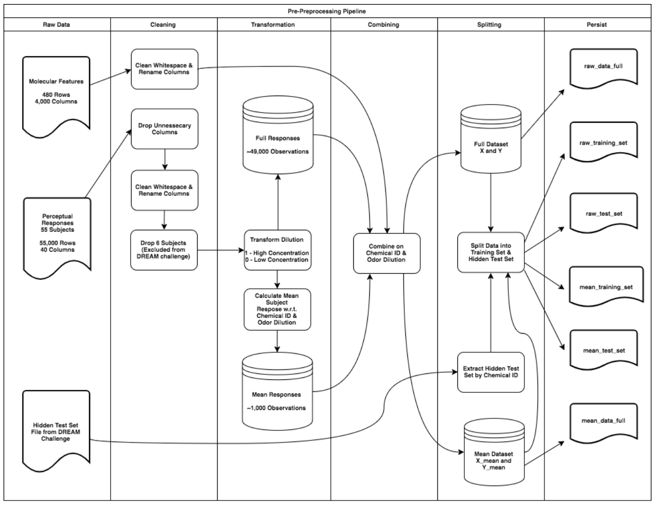

Fig. 2. Cleaning data, transforming dilution column, calculating mean response with masking, combining datasets, splitting hidden test set, persist data. - for more information see "preprocessing_pipeline" notebook in source.

A third file, obtained from the DREAM challenge repository contained the test set used in the challenge. This test set was constructed using a linear model trained using log-transformed principal components as features to partition the validation, training and test sets. The training data was chosen such that the baseline model had good predictive accuracy for this partition. Then the test set and validation set were split such that the median correlation for molecules for each partition was above zero. For our purposes, this train/test split allows us to directly compare our models with these previously published models.

D. Training Models

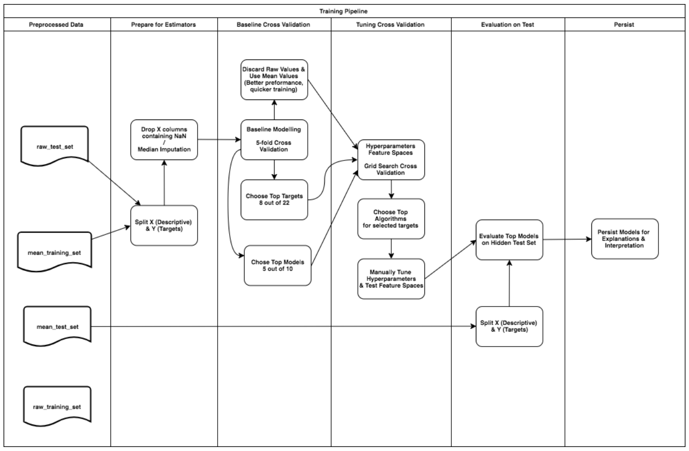

Fig. 3. Prepare data for estimators, baseline training, reduce targets, reduces algorithms, search for optimal hyperparmeters, manually tune final models, persist. - see notebooks in source

For baseline modelling we trained 10 different machine learning algorithms using default hyperparameter settings; the only preprocessing done was standardising to mean zero unit variance. Baseline learners were assessed through 5-fold cross-validation using the sci-kit learn function cross_val_predict() whereby the models were shown a random eighty per cent of the dataset and tested on the remaining twenty per cent. These predictions were compared to the actual responses and the correlation between predictions and actual values were used to rank our base learners. The 5 top performing algorithms were then chosen for further tuning.

Some targets were harder to predict than others. After baseline modelling and consulting the literature we decided to select the most promising 8 targets. Multiple feature spaces were constructed for the purposes of further tuning our 5 chosen algorithms by reducing and transforming our features. We tested performance using the top twenty discriminative features as published by a contributing team to the DREAM challenge (winner of the individual prediction sub-challenge) [37]. We then performed a randomised grid search, with 3-fold cross-validation to find the optimal hyperparameters for our models. The effect of applying different transformations was explored with the most successful method being chosen for final modelling.

Final modelling yielded the best results with random forest and regularised linear regression being chosen as the final two methods. As our final models were comparatively as predictive as the winning teams of the challenge these models were persisted to disk to be studied and interpreted.

IV. Results

A. Predictive Models

The first results of note in this investigation are the predictive accuracies of our final models. Regularised linear models and random forest produced the best predictions overall in my experimentation and this aligns with previous efforts. Table 2 shows the Pearson correlation for our final models which consisted of two sets of random forest models and one set of lasso regularised models to predict across 8 targets.

Individual models were trained for each target. Both sets of random forest models were trained with 1000 trees, one using all features and one using subsets of those features depending upon the target in question (top twenty per target). The training time of our reduced feature random forest model was significantly improved over using all of the features, with only a relatively small difference between the results of both random forest models.

Lasso linear regression trained much faster, compared to the random forest with a significant number of trees, and using the full feature set. It performs well in this task due to the efficient feature selection L1 regularisation.

Target	Lasso	RandomForest(20)	RandomForest
Intensity	0.71	0.68	0.74
Pleasantness	0.59	0.62	0.64
Fruit	0.34	0.23	0.29
Garlic	0.19	0.16	0.25
Sweet	0.30	0.33	0.34
Fish	0.22	0.25	0.28
Spices	-0.08	0.09	0.004
Burnt	0.16	0.13	0.16

Table 2. Pearson Correlation: predicted vs actual values.

Molecular features were cube root transformed and then scaled to mean 0 & unit standard deviation for final modelling. The distribution of the molecular descriptors was much more uniform after this set of transformations. This improved predictions significantly more than hyperparameter tuning or feature selection.

The most accurate models submitted for the DREAM challenge had Pearson correlation of 0.78 for intensity and 0.71 for pleasantness. The semantic descriptors were much more difficult to predict for teams in the challenge and we also found this to be the case. Our final models approach this level of predictive accuracy and are well above most teams who submitted models to the challenge. We were happy to investigate and interpret these models given their performance on the hidden test set.

”Spices” was reported to have been predicted well but in my own experiments, and in the final modelling, this descriptor was very difficult to predict.

B. Interpreting Models

The examination of our L1 regularised linear models is straightforward due to the natural interpretability of these algorithms. We can examine the coefficients to see which have a dominant effect and which features have dropped to zero. Our Lasso model to predict intensity clearly uses ”odor dilution” as it's main descriptive feature, as can be seen when examining the coefficients of the model. This explains the high correlation for intensity predictions. For pleasantness our lasso model also uses ”odor dilution” to predict this target but to a lesser degree, relying upon other features in the dataset without heavily biasing towards one descriptor as with the intensity model. For pleasantness we see a negative non-zero coefficient; We interpret this to mean that the higher concentration samples are perceived as ”less pleasant” across molecules.

The interpretation of our random forest regressor is conducted using novel methods to extract post-hoc explanations. The method we will discuss uses Shapley values to measure individual feature contribution to a prediction or across predictions. Shapley values originated in game theory, which is a subset of the study of complex systems. The implementation used is written in python and is called SHAP (SHapley values Additive exPlanations) [38]. This is a model agnostic explainer that provides local and global explanations. This work has coalesced multiple interpretability projects including LIME (Local Interpretable Model-Agnostic Explanations), the original library we had proposed to use.

The figure below shows the Shapley values for each descriptive features for a particular observation - this is a local explanation of our random forest model trained to predict pleasantness. For these chemicals, we can see various weights are assigned to the features pushing the value up and down for this prediction. We can see that for pleasantness the ”odor dilution = 0” pushed the value higher while ”odor dilution = 1” pushes the value lower. The way we can interpret this is that the lower dilution (dilution = 0) tends to be perceived as more pleasant than the higher dilution of a given chemical. This aligns with our interpretation of predictions from our regularised linear model and it is well established that odour dilution has a significant effect on the perceived scent [39].

The other feature descriptions are available from the DRAGON manual. To interpret these features are beyond our capabilities as computer scientists/students. A professional with a degree in chemistry, biology, or some flavour in that combination, would be the intended end user of this type of explanatory system.

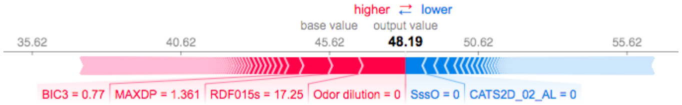

Fig. 4. Odor dilution = 0 is the strongest contributor to this prediction for pleasantness - pushing the prediction higher.

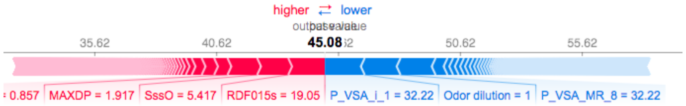

Fig. 5. For this observation the opposite is true - Odor dilution =1 is the second strongest contributor in the other direction for predicting pleasantness.

V. Conclusions

Interpretable Machine Learning will soon integrate into machine learning solutions such as scikit-learn and XGBoost. Fitting an interpreter to the end of a machine learning pipeline and being able to try different methods in parallel and compare them seems to be a natural integration into the existing workflow and architecture. For tensorflow or other deep learning frameworks, the task of integration is not more difficult architecturally but rather conceptually. Complex feature spaces and deep learning networks, and particularly a combination of both, seem to be the final bastion of truly black box methods. Interpreting deep learning models seems to be where most of the future work lies. XNNs (eXplainable Neural Networks) [40] learn interpretable feature representations as a part of their learning process. This approach of integrating explanations directly into the algorithm, rather than post-hoc explanations, seems to be more promising moving forward.

Another approach to studying the olfactory system is to investigate the laws that govern complex systems - such as cells, animals, brain, and economic systems; all are examples of systems that contain many entangled and unintuitive interactions - many moving parts combining in complicated ways. Random matrix theory [41] is one such example of attempting to understand complex systems by studying the system as a whole rather than the individual parts.

Other structure-dependant interactions of molecules with complicated systems (biological/ mechanical/ computer systems) could be elucidated using an Explainable AI approach. For instance, if we trained our Explainable AI on a training set of mappings of "structure-drug effects" we could use our model to predict the effects of new, unseen (or yet to be developed) drugs on the human body. This type of modelling could drive the creation of new fragrances, drugs and other compounds which interact with complex systems including but not limited to the human body.

This system can produce meaningful predictions but importantly it does not account for relative quantities of the molecules or the combinatorial effects of multiple odorant compounds in the air, both of which have an effect on the final perception of smell in real-world scenarios [42].

In the context of the SOR problem, this approach can help in the statistical analysis of the multivariate interactions between chemical compounds and human olfactory systems. In the broader context, this investigation outlines a generalised solution for the AI-assisted analysis any structure-dependant chemical interaction with a complex system.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016.

[2] J. Kim and J. F. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," *CoRR*, vol. abs/1703.10631, 2017.

[3] Z. C. Lipton, "The mythos of model interpretability," *CoRR*, vol. abs/1606.03490, 2016.

[4] S. R. Searle and M. H. Gruber, *Linear models*. John Wiley & Sons, 2016.

[5] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," *Distill*, 2018. https://distill.pub/2018/building-blocks.

[6] C. Olah, "The building blocks of interpretability," March 2018.

[7] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[8] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," *Official Journal of the European Union*, vol. L119, pp. 1–88, May 2016.

[9] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?," *CoRR*, vol. abs/1712.09923, 2017.

[10] G. Shepherd, *New Perspectives on Olfactory Processing and Human Smell*. No. Chapter 16, https://www.ncbi.nlm.nih.gov/books/NBK55977/: CRC Press/Taylor & Francis, 2010.

[11] D. Devanand, S. Lee, J. Manly, H. Andrews, N. Schupf, R. L. Doty, Y. Stern, L. B. Zahodne, E. D. Louis, and R. Mayeux, "Olfactory deficits predict cognitive decline and alzheimer dementia in an urban community," *Neurology*, vol. 84, no. 2, pp. 182–189, 2015.

[12] A. Roine, E. Veskimäe, A. Tuokko, P. Kumpulainen, J. Koskimäki, T. A. Keinänen, M. R. Häkkinen, J. Vepsäläinen, T. Paavonen, J. Lekkala, T. Lehtimäki, T. L. Tammela, and N. K. J. Oksala, "Detection of prostate cancer by an electronic nose: A proof of principle study," *The Journal of Urology*, vol. 192, pp. 230–235, 2018/02/25.

[13] N. Kahn, O. Lavie, M. Paz, Y. Segev, and H. Haick, "Dynamic nanoparticle-based flexible sensors: Diagnosis of ovarian carcinoma from exhaled breath," *Nano Letters*, vol. 15, no. 10, pp. 7023–7028, 2015. PMID: 26352191.

[14] M. TirzÄte, M. Bukovskis, G. Strazda, N. Jurka, and I. Taivans, "Detection of lung cancer in exhaled breath with an electronic nose using support vector machine analysis," *Journal of Breath Research*, vol. 11, no. 3, p. 036009, 2017.

[15] M. P. Brekelmans, N. Fens, P. Brinkman, L. D. Bos, P. J. Sterk, P. P. Tak, and D. M. Gerlag, "Smelling the diagnosis: The electronic nose as diagnostic tool in inflammatory arthritis. a case-reference study," *PLOS ONE*, vol. 11, pp. 1–10, 03 2016.

[16] M. Meister, "On the dimensionality of odor space," *eLife*, vol. 4, p. e07865, jul 2015.

[17] J. Gutiérrez and M. Horrillo, "Advances in artificial olfaction: Sensors and applications," *Talanta*, vol. 124, pp. 95 – 105, 2014.

[18] N. Katta, "Robust odorant recognition in biological and artificial olfaction," 2017.

[19] G. Bosc, "h(odor): Interactive discovery of hypotheses on the structure-odor relationship in neuroscience," September 2016.

[20] A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, l. schietgat, K. De Grave, R. Norel, , G. Stolovitzky, G. A. Cecchi, L. B. Vosshall, and p. meyer, "Predicting human olfactory perception from chemical features of odor molecules," *Science*, vol. 355, no. 6327, pp. 820–826, 2017.

[21] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning," *ArXiv e-prints*, May 2018.

[22] B. Herman, "The promise and peril of human evaluation for model interpretability," *CoRR*, vol. abs/1711.07414, 2017.

[23] N. Miyasaka, K. Morimoto, T. Tsubokawa, S.-i. Higashijima, H. Okamoto, and Y. Yoshihara, "From the olfactory bulb to higher brain centers: Genetic visualization of secondary olfactory pathways in zebrafish," *Journal of Neuroscience*, vol. 29, no. 15, pp. 4756–4767, 2009.

[24] Y. Seki, H. K. Dweck, J. Rybak, D. Wicher, S. Sachse, and B. S. Hansson, "Olfactory coding from the periphery to higher brain centers in the drosophila brain," *BMC biology*, vol. 15, no. 1, p. 56, 2017.

[25] B. W. Ache and J. M. Young, "Olfaction: Diverse species, conserved principles," *Neuron*, vol. 48, no. 3, pp. 417 – 430, 2005.

[26] E. Wnuk and A. Majid, "Revisiting the limits of language: The odor lexicon of maniq," *Cognition*, vol. 131, no. 1, pp. 125 – 138, 2014.

[27] A. Majid, "Cultural factors shape olfactory language," *Trends in cognitive sciences*, vol. 19, no. 11, pp. 629–630, 2015.

[28] L. Buck, "A novel multigene family may encode odorant receptors: A molecular basis for odor recognition," *Cell Press*, vol. 65, pp. 175–187, April 1991.

[29] J. B. Castro, A. Ramanathan, and C. S. Chennubhotla, "Categorical dimensions of human odor descriptor space revealed by non-negative matrix factorization," *PLOS ONE*, vol. 8, pp. 1–16, 09 2013.

[30] D. Laing, P. Legha, A. Jinks, and I. Hutchinson, "Relationship between molecular structure, concentration and odor qualities of oxygenated aliphatic molecules," *Chemical Senses*, vol. 28, no. 1, pp. 57–69, 2003.

[31] G. M. Lerman, J. V. Gill, D. Rinberg, and S. Shoham, "Two photon holographic stimulation system for cellular-resolution interrogation of olfactory coding," in *Optics in the Life Sciences Congress*, p. BrM3B.5, Optical Society of America, 2017.

[32] F. Peng and L. Chittka, "A simple computational model of the bee mushroom body can explain seemingly complex forms of olfactory learning and memory," vol. 27, 12 2016.

[33] A. Keller and L. B. Vosshall, "Olfactory perception of chemically diverse molecules," *BMC neuroscience*, vol. 17, no. 1, p. 55, 2016.

[34] R. Kumar, R. Kaur, B. Auffarth, and A. P. Bhondekar, "Understanding the odour spaces: A step towards solving olfactory stimulus-percept problem," *PLOS ONE*, vol. 10, pp. 1–19, 10 2015.

[35] T. G. Lewis, *Network science: Theory and applications*. John Wiley & Sons, 2011.

[36] K. Kaeppler and F. Mueller, “Odor classification: A review of factors influencing perception-based odor arrangements,” *Chemical Senses*, vol. 38, no. 3, pp. 189–209, 2013.

[37] H. Li, B. Panwar, G. S. Omenn, and Y. Guan, “Accurate prediction of personalized olfactory perception from large-scale chemoinformatic features,” *GigaScience*, vol. 7, no. 2, p. gix127, 2018.

[38] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[39] C. M. McGinley, M. A. McGinley, and D. L. McGinley, “Odor basics, understanding and using odor testing,” in *The 22nd Annual Hawaii Water Environment Association Conference*, pp. 6–7, 2000.

[40] J. Vaughan, A. Sudjianto, E. Brahimi, J. Chen, and V. N. Nair, “Explainable neural networks based on additive index models,” *arXiv preprint arXiv:1806.01933*, 2018.

[41] T. Conlon, H. J. Ruskin, and M. Crane, “Random matrix theory and fund of funds portfolio optimisation,” *Physica A: Statistical Mechanics and its applications*, vol. 382, no. 2, pp. 565–576, 2007.

[42] B. Malnic, J. Hirono, T. Sato, and L. B. Buck, “Combinatorial receptor codes for odors,” *Cell*, vol. 96, no. 5, pp. 713–723, 1999.