

UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Applications of Computation to Understand Chemosensory Processing

Permalink

<https://escholarship.org/uc/item/53g064dk>

Author

Kowalewski, Joel

Publication Date

2021

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Applications of Computation to Understand Chemosensory Processing

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Neuroscience

by

Joel Kowalewski

March 2021

Dissertation Committee:

Dr. Anandasankar Ray, Chairperson

Dr. Nicholas DiPatrizio

Dr. Chia-En Chang

Copyright by
Joel Kowalewski
2021

The Dissertation of Joel Kowalewski is approved:

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE DISSERTATION

Applications of Computation to Understand Chemosensory Processing

by

Joel Kowalewski

Doctor of Philosophy, Graduate Program in Neuroscience

University of California, Riverside, March 2021

Dr. Anandasankar Ray, Chairperson

Chemosensory processing encodes environmental information, relaying it to neural systems that regulate key behavioral responses. This broad definition implies the study of chemosensory processing is relevant across model organisms, leading to multiple practical applications. Of interest is chemosensory processing in agricultural pests and insect vectors, since volatile organic compounds and tastants determine behavior toward humans and agriculture. Some work has been done to uncover key pathways mediating behavioral attraction and aversion in the fruit fly, *Drosophila melanogaster*, as well as mosquito vectors. However, the limited number of pathways that can be experimentally manipulated suggests computational methods offer a complementary method. Machine learning has been applied to successfully predict ligands of insect chemosensory receptors. But these tools have not yet been applied across sensory encoding, identification of important neural pathways for attraction or aversion, and the discovery

of receptor ligands and chemical repellents. Such a comprehensive analysis pipeline is the aim of this work. Although emphasis is on insect repellent discovery, human as well as broader ecological toxicity remain highly relevant. This demands accurate in silico toxicity estimation in addition to cosmetic properties such as odor perceptual qualities that are a key consideration in designing topical formulations. Modeling of toxicological endpoints and human perceptual encoding by odorant receptors and the physicochemical features of odorants, are therefore discussed independently in detail, and later included into the repellent discovery pipeline. Ultimately, the discovery pipeline has helped identify numerous insect repellents that have desirable properties such as flavors and fragrances, has provided key insights into theories of chemosensory processing, and has been adapted to drug repurposing and discovery for COVID-19, with several top predicted compounds subsequently confirmed in vitro assays by others.

TABLE OF CONTENTS

Chapter 1: Introduction	1
Chapter 2: Sensory pathways that are predictive of behavioral valence in insects	
2.1 Introduction	4
2.2 Results	6
2.3 Discussion	11
2.4 Figures	13
2.5 Methods	18
Chapter 3: Discovery of physicochemical properties of ligands that act on repellent pathways	
3.1 Introduction	21
3.2 Results	24
3.3 Discussion	26
3.4 Figures	29
3.5 Methods	33
Chapter 4: Natural repellent and attractant activity of microbial metabolites on human skin	
4.1 Introduction	39
4.2 Results	42
4.3 Discussion	48
4.4 Figures	51
4.5 Tables	63
4.6 Methods	64
Chapter 5: Predicting human odor perception of odorants including repellents	
5.1 Introduction	68
5.2 Results	71
5.3 Discussion	78
5.4 Figures	81

5.5 Tables	98
5.6 Methods	103
 Chapter 6: Predicting human olfactory perception using odorant receptor activities	
6.1 Introduction	118
6.2 Results	121
6.3 Discussion	127
6.4 Figures	131
6.5 Tables	147
6.6 Methods	157
 Chapter 7: Adapting the computational pipeline to discovery of odorant-based Covid-19 drugs	
7.1 Introduction	164
7.2 Results	168
7.3 Discussion	174
7.4 Figures	177
7.5 Tables	190
7.6 Methods	199
 References	 207

LIST OF FIGURES

Chapter 2: Sensory pathways that are predictive of behavioral valence in insects

Figure 2.1	13
Figure 2.2	14
Figure 2.3	15
Figure 2.4	16

Chapter 3: Discovery of physicochemical properties of ligands that act on repellent pathways

Figure 3.1	29
Figure 3.2	30
Figure 3.3	31

Chapter 4: Natural repellent and attractant activity of microbial metabolites on human skin

Figure 4.1	51
Figure 4.2	52
Figure 4.3	54
Figure 4.4	56
Figure 4.5	57
Figure 4.6	59
Figure 4.7	60
Figure 4.8	62

Chapter 5: Predicting human odor perception of odorants including repellents

Figure 5.1	81
Figure 5.2	83
Figure 5.3	85
Figure 5.4	87
Figure 5.5	89
Figure 5.6	91
Figure 5.7	93

Figure 5.8	94
Figure 5.9	96

Chapter 6: Predicting human olfactory perception using odorant receptor activities

Figure 6.1	131
Figure 6.2	133
Figure 6.3	135
Figure 6.4	137
Figure 6.5	138
Figure 6.6	139
Figure 6.7	141
Figure 6.8	142
Figure 6.9	144
Figure 6.10	146

Chapter 7: Adapting the computational pipeline to discovery of odorant-based Covid-19 drugs

Figure 7.1	177
Figure 7.2	179
Figure 7.3	180
Figure 7.4	182
Figure 7.5	184
Figure 7.6	185
Figure 7.7	186
Figure 7.8	187
Figure 7.9	188

LIST OF TABLES

Chapter 4: Natural repellent and attractant activity of microbial metabolites on human skin

Table 4.1	63
-----------------	----

Chapter 5: Predicting human odor perception of odorants including repellents

Table 5.1	98
-----------------	----

Table 5.2	102
-----------------	-----

Chapter 6: Predicting human olfactory perception using odorant receptor activities

Table 6.1	147
-----------------	-----

Table 6.2	150
-----------------	-----

Table 6.3	152
-----------------	-----

Table 6.4	155
-----------------	-----

Table 6.5	156
-----------------	-----

Chapter 7: Adapting the computational pipeline to discovery of odorant-based Covid-19 drugs

Table 7.1	190
-----------------	-----

Table 7.2	193
-----------------	-----

Table 7.3	195
-----------------	-----

Table 7.4	196
-----------------	-----

Scientific Contributions

Acknowledgements

I would like to thank my coauthors and the supervision of my advisor, Anandasankar Ray. I am equally thankful for my parents and their persistent support of my academic and scientific research pursuits.

The text of this dissertation is reproduced in part with the permission of the licensor through RightsLink® as it appears in:

MacWilliam, D., Kowalewski, J., Kumar, A., Pontrello, C., & Ray, A. (2018). Signaling Mode of the Broad-Spectrum Conserved CO₂ Receptor is One of the Important Determinants of Odor Valence in *Drosophila*. *Neuron*, 97(5), 1153-1167.e4.
<https://doi.org/10.1016/j.neuron.2018.01.028>

Kowalewski, J., Huynh, A., & Ray, A. A systems level analysis of the olfactory percept space. (2021, in press). *Chemical Senses*, Oxford Univ. Press.
<https://doi.org/10.1093/chemse/bjab007>

The text of this dissertation is also reproduced in part in accordance with licensors who have given these rights without permissions to the author(s) to reuse published content in the dissertation as it appears in:

Kowalewski, J., & Ray, A. (2020). Predicting Human Olfactory Perception from Activities of Odorant Receptors. *iScience*, 23(8), 101361.
<https://doi.org/10.1016/j.isci.2020.101361>

Kowalewski, J., & Ray, A. (2020). Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space. *Heliyon*, e04639.
<https://doi.org/10.1016/j.heliyon.2020.e04639>

Chen ST, Kowalewski J, Ray A. (2021). Prolonged activation of carbon dioxide-sensitive neurons in mosquitoes. *Interface Focus* 11: 20200043.
<https://doi.org/10.1098/rsfs.2020.004>

Data and editorial synthesis of the dissertation

Behavioral and electrophysiological data used in analyses for Figures 2.1 – 2.4 are from Drs. Crystal Pontrello and Dyan MacWilliam, respectively. Similarly, electrophysiological recordings and behavior data in Figures 3.1 and 3.2 were performed by Dr. Stephanie Turner-Chen. This data was used to develop the computational models in Figure 3.3.

Chapter 1

Introduction

Mosquitoes including *Aedes aegypti* and *Anopheles gambiae*, which are notable insect disease vectors, as well as the fruit fly, *Drosophila melanogaster*, sense the chemical surroundings through Ionotropic Receptors (IRs), Gustatory receptors (Grs) and Odorant receptors (Ors). These chemosensory receptor classes are largely housed in structured called sensilla, with limited overlap between classes per sensillum. Each sensillum is identified by hair-like protrusions from the superficial epidermis (cutical). These hair-like structures are porous, enabling the transfer of chemicals into the sensillum and onto the sensory neuron dendrites, which house the receptors. The dendrites, extending down to cell bodies, are bathed in lymph. An electrochemical potential accumulates due to a difference in charge and ion concentration between the intracellular membrane surface and extracellular lymph. The sensory receptors at the dendrites subsequently mediate ion flow, altering the charge and concentration balance either toward or away from the activity threshold. Surpassing this activity threshold leads to an action potential or wave of intracellular depolarizing current relative to the extracellular sensillum lymph. The wave in turn propagates to second order neurons or glomeruli. These glomeruli represent a critical layer in which sensory information from antennal neurons is integrated and subsequently relayed to high order brain structures for memory and behavior regulation.

One key neural pathway in Diptera, particularly mosquitoes and flies, is associated with CO₂ sensing. While the specific receptors are distinct, dedicated CO₂ response pathways are well conserved across Diptera, including the experimentally

tractable and well-studied model organism *Drosophila melanogaster* or more broadly the fruit fly. Evolutionary conservation among diptera thus offers the possibility to develop general chemosensory processing models using *Drosophila*. Researchers have already characterized attractive and aversive behaviors in response to various environmental chemicals as well as some neural pathways. However, these studies alone cannot easily highlight putative interactions among numerous sensory receptor pathways or compare them according to their contribution to behavior.

An important next step is to clarify the most relevant receptor pathways. Here, alongside collaborators, I have outlined computational approaches to understand the chemosensory pathways underpinning simple attractive and aversive behaviors (Chapter 2); the subsequent chapter then applies these findings to accurately predict the activity of chemicals that target these pathways (Chapter 3). Accordingly, the need for a comprehensive map of chemicals on or potentially on human skin including the relationships between their structures and mosquito behavior is addressed in Chapter 4 as well as the development of a computational discovery pipeline for chemical repellents. Notably, since the repellents must equally be safe and suitable for human use, cosmetic and physical properties of a chemical are also important. Later chapters (Chapters 5-7) therefore include computational modeling of human perceptual neuroscience (Chapters 5 and 6) and toxicity (Chapter 7). While this work is centered around chemosensory neuroscience, the concluding chapter (Chapter 7) illustrates that the discovery pipeline for insect repellents can be generalized for COVID-19 treatments; repurposing approved drugs and discovering novel therapeutic compounds. Taken together, these studies

demonstrate the value of computational tools in basic biology and neuroscience research,
with an emphasis on understanding chemosensory processing from theory to application.

Chapter 2

Sensory pathways that are predictive of behavioral valence in insects

2.1. Introduction

Although several insect repellents are widely and consistently used, the precise mechanisms, particularly for the most prominent chemical repellent, N, N-Diethyl-methyltoluamide (DEET), remain unclear. DEET is the outcome of a large-scale government initiative in the 1950s to discover chemical structures that repel or possibly incapacitate insects. In subsequent decades, it has been thoroughly studied to replicate its efficacy, but these efforts have suggested numerous modes for its repellent activity. This complexity makes it challenging to use traditional chemical approaches, which are most successful when one or few protein targets are well defined. It is then plausible to design chemicals around the relevant physical constraints of the protein target(s) such as the electrostatic interactions that facilitate docking. However, in chemosensory science, even if the protein targets were well defined, there is a paucity of 3D structural data for sensory receptors. This presents, at least initially, a computational problem. This problem can be broken down into two steps: (1) Identification of the sensory receptors/pathways that are predictive of simple behaviors in mosquitoes and agricultural pests; (2) study the physicochemical attributes of chemicals that act on these receptors/pathways by identifying which of these attributes best predict the activity.

The work presented in this chapter canvasses the first of these two steps. Here, my colleagues and I observed a correlation between electrophysiological recordings from the CO₂ detecting neuron in *Drosophila* and odor valence (e.g. attraction or aversion to an

odorant). Odor valence was quantified using the T-maze assay, where flies navigate to a solvent treated control arm or a chemically treated experimental arm. Counting the flies in each arm and expressing as a ratio gives the Preference Index (PI), a coefficient from -1.00 to +1.00 that quantifies attractive (positive values) and aversive (negative values) behaviors. Though the activity of odorant receptor neurons (ORNs) is not measured during the T-maze assay, the fly's navigation is based on detecting odorant molecules via these neurons. It is therefore possible to record from these neurons independently of the assay, later using these activities to predict the Preference Index (PI).

My colleagues and I started with activities from 24 odorant receptor neurons (ORNs) in the antennae. Interestingly, we failed to find a correlation between electrophysiological recordings from these neurons compared to ab1C, a unique sensory neuron, also housed in the antennae that expresses gustatory receptors rather than odorant receptors. Early developmental regulation typically leads to the expression of a unique odorant receptor in the odor-sensing neurons of the antennae; the neuron is therefore often abbreviated as the odorant receptor alone. The ab1C neuron, in contrast, expresses two receptors, Gr21 and Gr63a, which confer sensitivity to CO₂. This specialized neuron is highly evolutionarily conserved. Subsequently, the correlation between activity from this neuron to 54 odorants and the corresponding Preference Index (PI) values from the T-maze was surprising.

The relevance of ab1C activity to odor valence in *Drosophila* as well as evidence in mosquitoes suggested the hypothesis that CO₂ detecting neurons may be key in predicting mosquito and fruit fly behavior. My work focused on determining if a rigorous

computational analysis would support this, work which is outlined in this chapter for *Drosophila* (fruit flies).

2.2 Results

2.2.1 Predicting behavior with and without ab1C activity.

To assess the behavioral contribution of the activity recorded from the 24 ORs (olfactory sensory neurons, ORNs) and the Gr21a/Gr63a-expressing ab1C neuron, we performed a series of statistical and feature-selection approaches. These identify which receptor(s) optimally predict behavior (Figure 2.1A). Initially, a simple regression analysis using the known activities of the 24 ORs to 54 odorants failed to explain the variability in fly preference (Preference Index, PI) to these same odorants ($p > .05$). However, adding the activity of the ab1C neuron, improved the fit, explaining 63% of the variation in the T-maze behavior ($p = 0.03$). Interestingly, the activity of ab1C alone was also statistically significant ($p < 0.001$) and favored according to a measure that evaluates the quality of the model fit (BIC = 24.6) (Figure 2.1B, C).

We next identified the minimum number of receptors that could predict behavior, as was done previously for larval behavior (Kreher et al., 2008). The 25-predictor model (24 Ors and ab1C) was analyzed using stepwise regression, entailing the sequential removal of predictors until converging upon an optimal subset. Candidate models were screened using values of R squared and the Bayesian Information Criterion (BIC). Surprisingly, only a two-predictor model with ab1C and Or85f was retained when using the stepwise selection method alone as before (Kreher et al., 2008). In order to further control if this model was a byproduct of a few influential odorants affecting the

regression fit or spurious correlations with the Preference Index (PI), the odor space was sampled from with replacement, a procedure that is also referred to as bootstrapping. This resulted in thousands of random combinations of the 54 odorants, ensuring the reliability of the finding. Running the stepwise regression iteratively on 5000 combinations and recording the selection rate for each predictor in the final model suggested that a model including *ab1C*, *Or2a*, *Or67a*, *Or59b*, and *Or19a* generalized well across the different odorant combinations (Figure 2.1D). High selection rate across the combinations was for the most part consistent with the t statistic assigned to each predictor for the full linear regression model (e.g. all 25 predictors) to the 54 odors (Figure 2.1B, C). The linear regression model with the smaller subset of informative predictors resulted in the linear equation, Avg. PI = -0.23 - 0.09 *Or67a* + 0.02 *Or2a* - 0.04 *Or59b* - 0.03 *Or19a* - 0.14 *ab1C* (Figure 2.1E). Most of the predictors in the model are broadly tuned ((Hallem and Carlson, 2006), DoOr database (<http://neuro.uni-konstanz.de/DoOR/default.html>)), consistent with the expectation that, since they are activated by many odorants, they should remain predictive of the T-maze behavior across many different combinations of the 54 odorants. The importance of *ab1C* was the highest among the 25 sensory neuron activities, as determined by the number of times it was selected (out of 5000) for the final, “best” model, based on statistical criteria. This further emphasizes the role of *ab1C* in predictions of odor valence (Figure 2.1D).

Given further evidence that *ab1C* activity was more informative than Ors, we revisited the comparison between *ab1C* and the 24-Or model (shown in Figure 2.1B) but now using a cross validation procedure. This entails repeatedly fitting the regression

models on a smaller subset of odorants, then predicting the T-maze behavior for the odorants that are excluded. In this sense, it offers a true assessment of the average predictive power of the model and therefore helps evaluate its usefulness in predicting the T-maze behavior for any arbitrary odorant, rather than simply the 54 odorants studied here. To perform this analysis, the regression model with all Ors (e.g. excluding ab1C) was now fit using regularized regression (also called ridge regression). Because the larger 24 Or model is more complex than ab1C alone, it will also be less stable in its predictions. Such a scenario may give rise to poor prediction of T-maze behavior purely for statistical reasons. Regularization circumvents this by penalizing the larger, 24 Or model from being too complex, which means the coefficients for Ors that are not informative are shrunk toward 0; that is, they contribute little to the prediction. The results of this analysis indicated that ab1C explained 41% of the variability in T-maze behavior over the validation approach, as compared to 22% for the all Or model (Figure 2.1F).

It remained unclear from these analyses, however, to what extent odor valence was indeed a linear function of receptor/neuron activity in the antenna and if this was an unreasonable constraint. Recent studies have suggested the possibility of non-linear interactions in contribution of ORNs or glomerular activities to behavior behavior (Badel et al., 2016; Bell and Wilson, 2016). We therefore broadened the scope of our analysis using different machine learning algorithms that are more flexible and conducive to capturing non-linear relationships. Using these, we tried to determine (1) at what frequency would ab1C meaningfully improve predictions regardless of the algorithm

being used, (2) what are the consensus optimal predictor sets selected across these algorithms, and (3) which algorithms minimize prediction error after removing uninformative predictors (Methods). To compare the differing approaches and models, error rates were evaluated using bootstrap validation (1000 resamples) or 10-fold cross-validation, repeated 100 times (1000 folds). These techniques involve training each algorithm on a matrix of receptor activities to odorants, subsequently predicting the T-maze Preference Index (PI) for samples of odorants that were not used during the training. Across algorithms for identifying optimal predictors, ab1C was always ranked above the 24 Ors, followed by Or67a and Or22a. Intriguingly, Or22a, which displays a more complex relationship with behavior was high on every list but was nevertheless missed by the previous OLS regression and stepwise removal (Figure 2.2A, B). In general, models sensitive to non-linearity and interactions amongst the predictors resulted in slight improvement during validation, yet the major determinant was whether ab1C was in the model (Figure 2.2B, D; Figure 2.3A, B). Despite implementing many complex algorithms, any improvement approximated our control case, fitting a simple regression model with ab1C and Or67a ($R^2 = 0.45$) (Figure 2.3A, B). Larger odor samples will undoubtedly favor these sophisticated algorithms, but it remains surprising that ab1C was selected as one of the top predictors of valence for the T-maze behavior generated in this study.

It would be important to ask whether ab1C activity is also a significant predictor for other types of olfactory behavior in longer-term assays such as the wind-tunnel, walking assays, or traps. While large odor sets have not yet been tested in the wind-

tunnel, we were able to utilize a large behavioral preference data set generated using trap assays, which had substantial overlap with the odorants (47/110) that we tested in the T-maze (Knaden et al., 2012). The trap assay evaluates attraction to a “trap” or baited enclosure and is run for a longer duration than the T-maze (hours vs minutes). But the Preference Index (PI) (T-maze) and Attraction Index (AI) (trap assay) are otherwise conceptually similar. Interestingly, the behavioral preferences across the 2 assays differed for odorants common to the two studies ($r = 0.01$ $p = 0.9$; rank ordered correlation for the bottom ten scoring compounds in the T-maze assay $\rho = 0.44$, $p = 0.2$), which suggests the behavior is potentially occurring through different olfactory pathways. Applying the earlier computational approach to predict the trap assay behavior, we identified the top 7 optimal predictors (Ors) (Figure 2.4B). However, unlike with the T-maze, few predictors were individually informative; it was no longer evident that a simple rank ordering from the selection rate was useful. Instead, combinations of the top 7 predictors were reassessed using repeated 10-fold cross-validation, or repeatedly dividing the data into training and testing portions, as discussed earlier. The ordinary least squares (OLS) regression fit for the best model resulted in the linear equation, $\text{Avg. AI} = 0.19511 + 0.07894 \text{ Or}_{59b} - 0.09033 \text{ Or}_{49b} - 0.05763 \text{ Or}_{98a}$ on the original data (Figure 2.4D-F). These results suggest that the statistical approach we applied can nevertheless identify odorant receptors that predict the trap behavior ($R^2=0.4$), as was possible with T-maze. A more general approach excluding cross-validation and considering activities of all 24 Ors was not sufficient to predict behavior (Knaden et al., 2012). Surprisingly, however, ab1C was not a significant predictor of the trap behavior, suggesting that the behavioral

responses to these two olfactory assays are likely generated in a fundamentally different manner, using different receptors.

2.2. Discussion

An exhaustive statistical analysis to test whether a few selected ORN types could model the T-maze behavior in response to the tested odorants led to a linear model with 4 broadly-tuned Ors (Or2a, Or19a, Or59b and Or67a) and Gr21a/Gr63a. In fact, in every possible unbiased analyses we tried, both simple linear regression and based on sophisticated machine learning (altogether ~20 different methods), the activity of ab1C was consistently selected as the top performing descriptor for behavior predictions. The valence of several odorants therefore depends upon the *Gr21a/63a (ab1C)* pathway. However, narrowly tuned *Ors* detect odorants that elicit specialized behaviors such as oviposition, or act as pheromones, some that are species-specific (Knaden and Hansson, 2014). Our experiments also illustrate that the valence of ~16% odorants are lost and ~10% are altered in the *orco* mutant flies, suggesting the importance of the *Or* pathway (MacWilliam, Kowalewski, Kumar, Pontrello, & Ray, 2018). This is consistent with recent studies showing segregation of spatial inputs for attractive and aversive odorants in the Lateral Horn brain region of the second order projection neurons connected to *Or*-neurons (Strutz et al., 2014).

Although simple regression approaches were suitable to optimally predict behavior, more sophisticated algorithms with sensitivity to non-linearity led to incremental improvement. Consistent with Bell and Wilson (2016), some *Or* activities

relate to behavior non-linearly, and therefore it is expected that algorithms that can capture all relationships (linear and non-linear) will ultimately be optimal. However, it is likely that the number of odorants studied here was not large enough to result in a substantive performance difference. The predictive success shown here suggests computation may be used to merge separate behavior and electrophysiological experiments, gaining new insight into insect control.

2.4. Figures

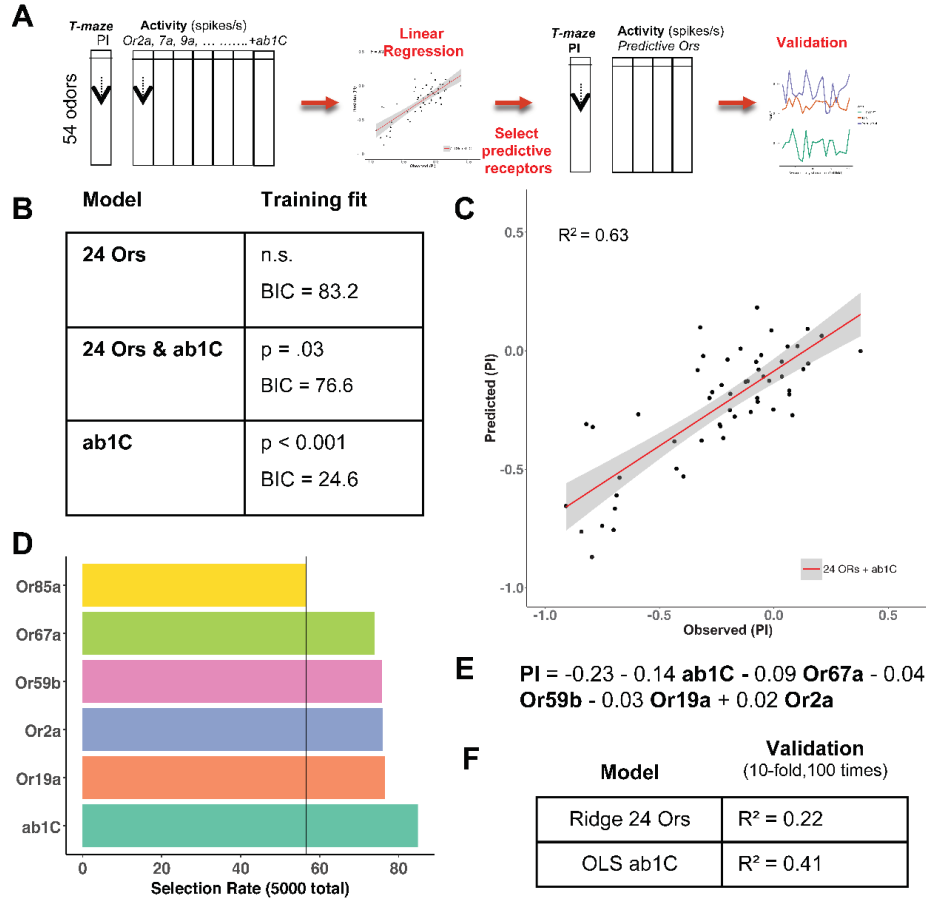


Figure 2.1. CO₂ receptor neuron activity required for prediction of odor valence from responses. **A)** Sample workflow of the modeling approach. The T-maze preference index for 54 odor x 24 Or-response matrix was used to predict the PI; this Or-only model was initially fit using OLS regression and was then retested for fit after adding ab1C activity for the 54 odorants. Uninformative predictors were removed and the reduced model was validated. **B)** Tabulated measures of fit are shown for the labeled model on the original data. **C)** Predicted PI was plotted as a function of the observed PI for the 24OR+ ab1C model; the red line depicts the linear trend while the overlaying gray band is the standard error for the fit. **D)** Predictors that are selected most frequently and their selection rates, across 5000 iterations of stepwise regression, resampling the 54-odorant set on each run. The black vertical line is the empirically determined threshold for consistent selection out of 5000 iterations. **E)** Linear equation of the optimal predictors. Units for the coefficients reflect the Z transformed spikes/s. **F)** Average performance on 1000 cross-validation test folds is shown for two models. To ensure optimal performance and stability of the larger Or-only model, the test average is shown for ridge regression and compared to ab1C alone using OLS regression. Abbreviations: OLS, Ordinary Least Squares; BIC, Bayesian Information Criterion.

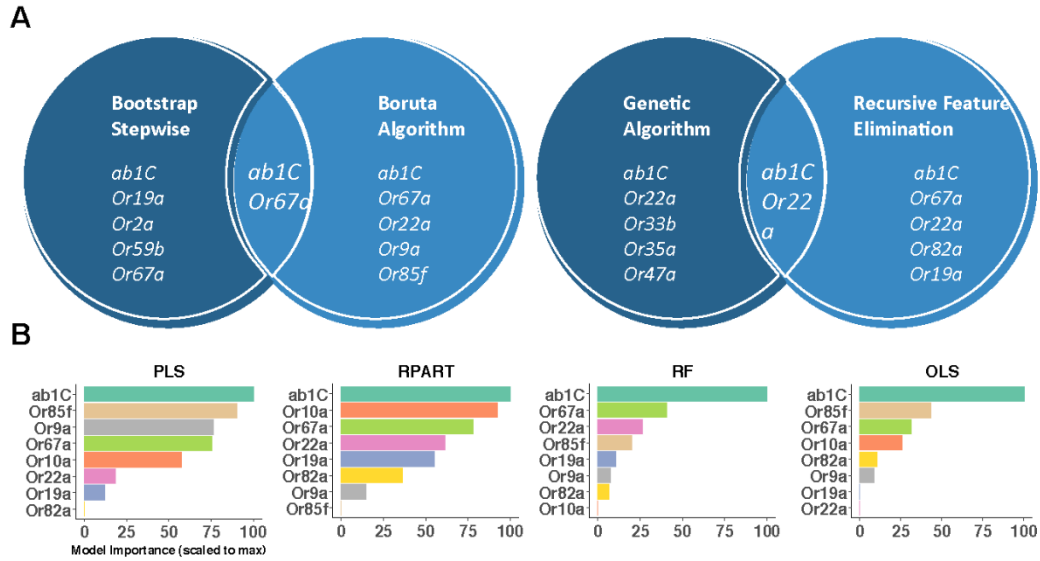


Figure 2.2. Identifying the optimal predictors using multiple approaches. **A)** Several model selection algorithms arrive at a consensus minimal set of optimal predictors and all include ab1C. **B)** Additional algorithm-specific variable importance measures scaled to the maximum. Ten predictors appearing in the lists shown in (A) were tested on the 54 odorant set using recursive partitioning (RPART), partial least squares (PLS), random forest (RF) and ordinary least squares (OLS) regression. Except for random forest the scaled importance metrics are derived from the original fit and are independent of resampling or cross validation. Of the optimal Ors selected for reducing prediction error on 1000s of resampled odor sets, some are not important on the original data; ab1C does however display consistency in this context.

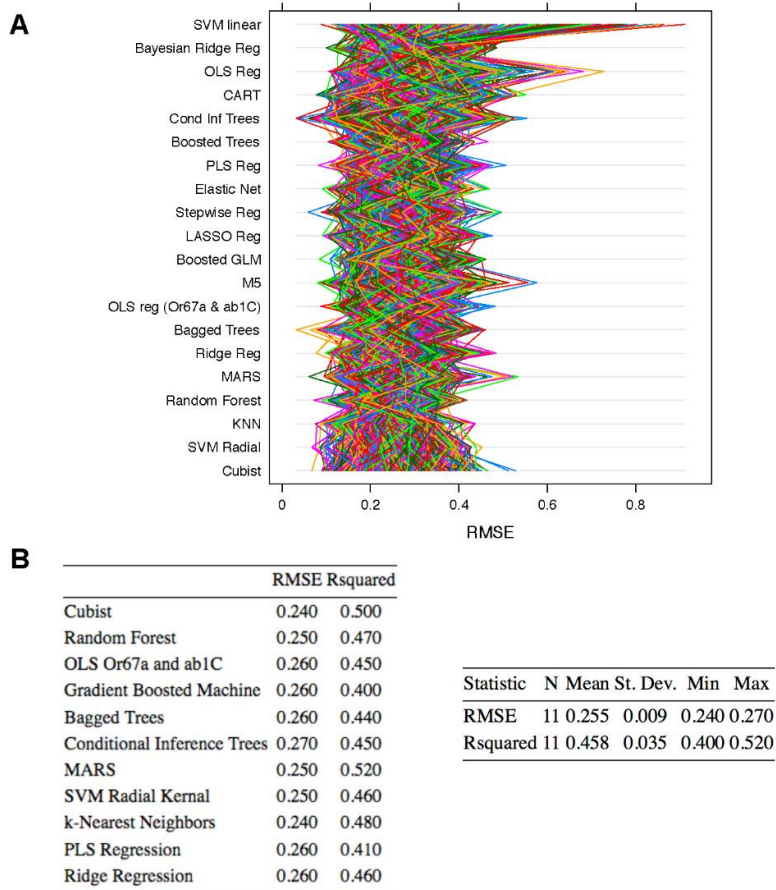


Figure 2.3. Identifying the optimal methods for predicting behavior. **A)** Multiple machine learning algorithms are compared for the 10 best predictors, using the two predictor model of ab1C and Or67a as a control case. Performance is evaluated on different portions of data “hidden” from the original fit. The performance metric is the square root of the average difference (error) between the predicted and observed Preference Index (PI) for the T-maze (RMSE). Each algorithm that is labeled along the vertical has been evaluated 1000 times according to this metric; each time is the prediction of the Preference Index on a different set of “hidden” or test odorants. Because the odorants in these test sets are the same for all algorithms, it is possible to compare. Therefore, the colored vertical lines represent all the test performances from one algorithm to the next. The objective then is to identify the lowest RMSE (error) values and those that are less scattered; the latter highlights the algorithms that predicted the behavior with less variability. The plot illustrates that the simple regression model (OLS) with Or67a and ab1C predict with error rates comparable to many sophisticated algorithms. Algorithms like the linear support vector machine (SVM Linear) predict less accurately, with high variability. **B)** Left, the averaged performance metrics, including the R squared, confirm no approach warrants selection over the two-predictor model fit using OLS regression, given a diverse but not exhaustive set of 54 odorants. Right, the summary statistics for the tabulated performance shown on the left.

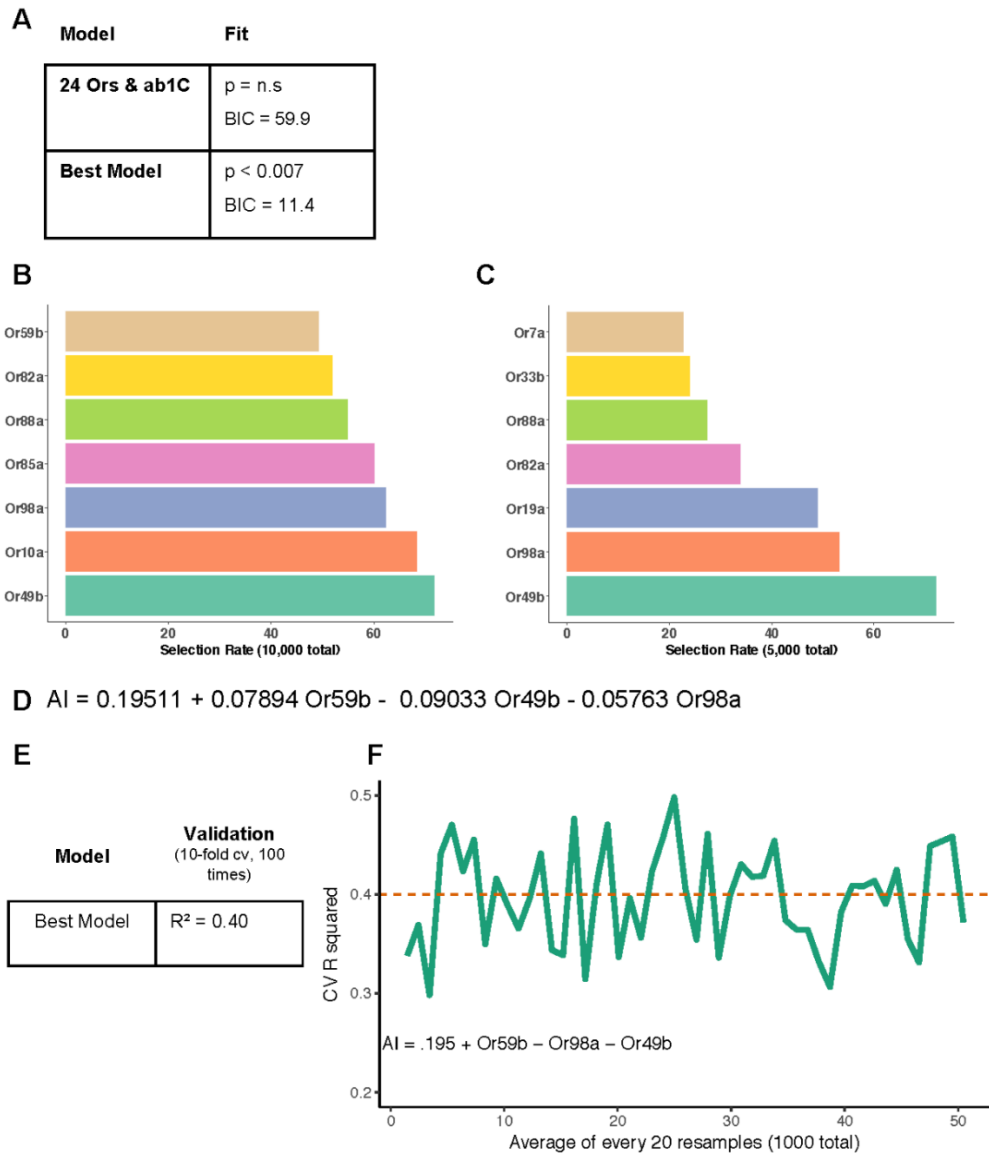


Figure 2.4

Figure 2.4. Optimizing the prediction of odor valence from the trap assay. **A)** The 24 Or model including ab1C is compared to a three predictor model, performing optimally during cross-validation. Statistics comparing these models are based on the OLS regression fit, predicting the attraction index (AI) for all 47 odorants overlapping with available ab1C activities. **B)** Following filtering for high correlations, potentially predictive models were pre-screened based on the rate of predictor selection according to stepwise regression using BIC minimization and backward elimination on 10,000 resamples of the odor space. Unlike in the T-maze, few predictors were individually informative; it was no longer evident that a simple rank ordering from the selection rate was useful. Instead, combinations of the top 7 predictors were assessed using repeated 10-fold cross-validation. **C)** Top 7 predictors based on the selection rate for the 110 odorant set. **D)** The best performing model as determined by cross-validation. Estimates for coefficients in the linear equation are representative of the standardized activities (Z-transform) for each of the predictors on the 47 odorant set and are subsequently on the same scale. **E)** The average variability accounted for in the attraction index across 100 iterations of 10-fold cross-validation using the model in D and the 47 odorant set. **F)** The cross-validation performance is collapsed into 50 bins representing the variability in performance (solid line), along with the overall average (dashed line). Abbreviations: BIC, Bayesian Information Criterion; OLS, ordinary least squares.

2.5. Methods

2.5.1. Drosophila T-maze assays

T-maze behavioral testing using *Drosophila* was performed as described previously (Turner and Ray, 2009), with minor modifications. Twenty males and 20 females, 3-7 days old, wet starved ~20-25 hrs were used in each trial in a T-maze without airflow, placed inside a 30 cm³ white card perimeter. Odorants were of the highest available purity (Sigma-Aldrich). Chemicals were diluted in water or paraffin oil. For most odorants, tubes contained 10 ul of odorant solution or solvent, were sealed with Parafilm and allowed to volatilize for ~10 min prior to the start of each 1 min trial.

PI = (flies in test arm – flies in control arm)/total number of flies in arms of T-maze.

2.5.2. Computational modeling of behavior

Regression analyses were conducted in R version 3.3 (R Core Team, 2016) using the step() and lm() functions. After fitting the full model, predictors were assessed in smaller subsets using an exhaustive search algorithm, applying multiple parameters for the quality of the fit. Models that reduced complexity while optimizing the R squared, Mallows' C_p and BIC statistics were cross referenced with the solution from stepwise regression, which employs an automated search for optimal predictors; the full model was fit with successive removal of predictors (backward selection) based on BIC minimization (BIC: Bayesian Information Criterion). The BIC is a probability measure that is used to identify the model that is best supported by the data. If a model has many

predictors (e.g. many parameters estimated from the data), the BIC attempts to justify the choice of this large, complex model relative to its explanatory power.

To control for overfitting, or the tendency to overemphasize the noise in the data, the model including the optimal predictors was tested by applying repeated 10-fold cross-validation (1000 folds) or the bootstrap (1000 resamples), unless stated otherwise. Also, since the selection of predictors on training cases is not always representative, a cross-validation approach was taken to confirm and possibly identify other predictors that explained variability in the T-maze behavior (PI) on resamples of the odor space. Machine learning algorithms applied in support of these and other variable selection approaches were based on customized scripts in the R programming environment, along with support from the classification and regression training (caret) package (Kuhn, 2008), the kernlab (Karatzoglou et al., 2004) and e1071 (<https://cran.r-project.org/web/packages/e1071/e1071.pdf>) packages. Optimal predictor selection with the Boruta algorithm was similarly carried out using the implementation available in R (Kursa, et al., 2010). In cases where algorithms could be tuned, particularly for regularization, optimal values were identified by searching the space of available parameters and using the combinations that maximized predictive performance on data withheld during training.

Bootstrapping the stepwise regression addresses mild correlations amongst predictors (e.g. Ors). This affects the selection of an optimal model, since the choice of one predictor over another in the presence of correlations is arbitrary; namely, these correlated predictors could be substituted for each other without affecting the model fit.

But in many cases the correlations become too severe and more rigorous procedures are necessary to corroborate which predictors and models are indeed optimal. Correlated predictors (multicollinearity) can be addressed through partial least squares regression (PLS) or principal component regression (PCR), but these approaches are at the expense of detail on the best predictors. Model-specific variable importance measures are available to determine how much certain variables contribute to the best predictive equation; however, the coefficients of this model nevertheless lack interpretability. These data were ultimately excluded from the primary text. As a complement, models were also fit using regularized regression, such as ridge regression, elastic net and lasso (least absolute shrinkage and selection operator); the latter two offer alternative, built-in methods for model selection given correlated predictors by shrinking the standardized predictor coefficients toward zero, if they are too high or unstable. These regression approaches also retained ab1C. But these optimal models failed to significantly improve performance beyond similarly sized OLS (Ordinary Least Squares) regression models. In the interest of thoroughness, specialized predictor selection algorithms, genetic, Boruta and recursive feature elimination, were applied in conjunction with random forest regression to generate lists of optimal predictors. These do not assume a linear relationship between the Preference Index (PI) and responding unit (sensory neuron), so they offer a potentially more robust interpretation.

Chapter 3

Discovery of physicochemical properties of ligands that act on repellent pathways

3.1. Introduction

The analysis of odor processing in mosquitoes and flies identified a few odorant receptor neurons in addition to the CO₂ detecting neuron strongly contributed to predictions of simple behaviors (e.g. attraction/repulsion). The ab1C neuron, which confers CO₂ sensitivity, in the fruit fly provided the most significant contribution to behavior prediction, leading to a preliminary model where evolutionary conserved sensory pathways, such as for CO₂ detection, may play a more important role in determining aversive and attractive behaviors. One implication is then that these conserved pathways may provide key insight into repellency and the discovery of novel chemical repellents. To that end, my colleagues and I characterized the neural activity of the CO₂ -detecting cpA neuron in mosquitoes, with my work centering on the physicochemical basis of the activity.

Carbon dioxide (CO₂) serves as a long-distance orientation and host-seeking cue for most mosquito species. Human beings generate CO₂ odor plumes through exhaled breath, causing fluctuation in CO₂ between background (0.04%) and expired levels (4%). This intermittency in CO₂ concentration is thought to increase host-seeking behavior in mosquitoes, causing them to fly upwind toward the odor source (Cardé & Willis, 2008; Dekker, Geier, & Cardé, 2005). Once the mosquito has followed the CO₂ plume toward its source, it is thought that the insect will then detect other sensory cues such as skin

odors and heat (Takken & Knols, 1999). Not surprisingly, mosquito species such as the ornithophilic *Culex quinquefasciatus* and the anthropophilic *Anopheles gambiae* and *Aedes aegypti*, are differentially attracted to host odors such as those from avian, and human sources, respectively.

However, CO₂ is an odor common to all hosts as it signifies the presence of a vertebrate's exhaled air. When presented in an optimal fashion, CO₂ can readily attract mosquitoes in the field and in the laboratory (Cooperband & Cardé, 2006; Dekker, Takken, & Braks, 2001; A. J. Grant, Aghajanian, O'Connell, & Wigton, 1995; Xue, Doyle, & Kline, 2008), as well as increase the sensitivity of mosquitoes to other human odors (Dekker et al., 2005). Since CO₂ is highly influential in host-seeking behavior of many mosquito species, the majority of mosquito traps employ CO₂ as the primary lure. The maxillary palp is the CO₂ detecting organ, where of the three neurons housed in the club-shaped capitulate peg (cp) sensilla, the cpA neuron expresses the CO₂ receptor Gr1, Gr2 and Gr3 (also called Gr22, Gr23, and Gr24) which belong to the gustatory receptor family (Lu et al., 2007; Syed & Leal, 2007). These proteins are closely related to the CO₂ receptor of *Drosophila melanogaster*, Gr21a and Gr63a which are required for response to CO₂ (Jones, Cayirlioglu, Grunwald Kadow, & Vosshall, 2007; Robertson & Kent, 2009).

Apart from CO₂, this receptor is also activated and inhibited by an array of volatile odorants that can be grouped into multiple structural categories (Coutinho-Abreu, Sharma, Cui, Yan, & Ray, 2019; MacWilliam, Kowalewski, Kumar, Pontrello, & Ray, 2018; Tauxe, Macwilliam, Boyle, Guda, & Ray, 2013; Turner et al., 2011; Turner & Ray,

2009). Each of the proteins in the receptor have a 7-transmembrane structure and while Gr2 and Gr3 constitute the core receptor, Gr1 increases sensitivity to CO₂ and to inhibitory odorants (Kumar et al., 2020). It has been previously shown that inhibition of the CO₂ response by volatile odorants corresponds to complete loss of innate CO₂ avoidance behavior in *Drosophila* (Turner & Ray, 2009). Given the reversal of behavior to CO₂ in the presence of the inhibitory odorants, and that mosquito CO₂ receptors have high amino acid identity with the *Drosophila* ortholog Gr63a and Gr21a (Hill et al., 2002; Kent, Walden, & Robertson, 2008; Lu et al., 2007; Robertson & Kent, 2009), we tested and identified similar odorants that could have a similar effect on CO₂-mediated host-seeking behavior in mosquitoes (Coutinho-Abreu, Sharma, Cui, Yan, & Ray, 2019; MacWilliam, Kowalewski, Kumar, Pontrello, & Ray, 2018; Tauxe, Macwilliam, Boyle, Guda, & Ray, 2013; Turner et al., 2011; Turner & Ray, 2009). The identified volatile odorants included: odors that inhibit the CO₂-sensitive neuron and are candidates for use in disruption of host-seeking behavior, odors that activate the neuron and can be a substitute for CO₂ as a lure in trapping devices, and odors that cause strong and prolonged activation of the CO₂ neuron which blocks the ability to detect changes in CO₂ concentration and therefore offers a novel approach for disruption of host-seeking. These compounds could be used as tools for mosquito control as they modify peripheral olfactory responses to one of the most important host-seeking cues. These odor-based strategies once developed could potentially lower the incidence of human-mosquito contact, and hence lower the spread of vector-borne diseases.

3.2. Results

In the past we have used single-sensillum electrophysiology to screen a large number of odorants for their effect on the activity of the CO₂-sensitive neuron in the peg sensilla of the maxillary palp of female *A. gambiae*, *A. aegypti*, and *C. quinquefasciatus*. The cpA neuronal response to CO₂ is nearly identical in all three species and it can be unambiguously identified since it has a much larger spike amplitude than the other two neurons in the same sensillum. When looking for activator and inhibitory odorants, we also found that the responses showed significant conservation (Coutinho-Abreu, Sharma, Cui, Yan, & Ray, 2019; MacWilliam, Kowalewski, Kumar, Pontrello, & Ray, 2018; Tauxe, Macwilliam, Boyle, Guda, & Ray, 2013; Turner et al., 2011; Turner & Ray, 2009). One of the interesting questions has been how volatile components of malodorous body odor might be interacting with the mosquito CO₂ receptor. Many of the malodorous compounds are due to bacterial breakdown of lipids, such as butyric acid. When performing the electrophysiological recording odor screens, we observed that butyric acid caused an initial phasic activation followed by inhibition of the CO₂ response (Figure 1). However, following this brief phasic excitation, the odorant induced a ‘prolonged’ tonic activation of the cpA neuron.

In previous studies, a prolonged tonic activity has been shown to mask the activation caused by subsequent exposures to CO₂ such as 2,3 butanedione, (E)-2-methylbut-2-enal, 3-Methyl-2-Butenal, 3-Methylbutanal (Tauxe et al., 2013; Turner et al., 2011). This type of effect has also been observed in other odorant receptor neurons with odorants like Methyl 2-propenoate and Methyl propionate (Boyle, McInally,

Tharadra, & Ray, 2016). To investigate if prolonged activation by butyric acid could also cause a reduced response to subsequent CO₂, *A. gambiae* and *A. aegypti* mosquitoes were exposed to a 3-sec application of the odorant followed by repeated 1-sec stimulus of 0.15% CO₂ applied every 30-sec for a period of approximately 5 minutes. When comparing spike rate in both mosquito species, there is an increase in baseline activity of the cpA neuron (Figure 3.1 and 3.2). However, the brief exposure to butyric acid significantly reduced CO₂ response for as long as 5.5 min in *A. gambiae* (Figure 3.2, right), while the CO₂ response in *A. aegypti* was completely abolished (Figure 3.2, left). These results suggest that the prolonged tonic response can substantially impair the ability to sense other ligands like CO₂ for minutes.

To investigate the structural basis of the different activities, we first compared simple enriched substructures or cores among activators, prolonged activators, and inhibitors of the cpA neuron (Figure 3.3A). Interestingly, the correspondence between enriched substructures and activity was unclear. We next computed additional physicochemical features, incorporating information about 3D geometries, the distribution of charge across a molecule and other atomic-level properties describing bonds and bonding potential. As it is not feasible to manually search numerous features, we applied machine learning to identify sets of features that were particularly different amongst prolonged activators (Figure 3.3B) and all other cpA activities. This approach involved iteratively training a support vector machine (SVM) on a portion of data, followed by predicting the remaining ‘left out’ portion (Methods). Consistent with the overlapping enriched substructures (Figure 3.3A, B), the features that were predictive of

prolonged activators often described 3D geometries (Figure 3.3C). We next tested whether SVMs trained on these important features could successfully discriminate prolonged activators from the other cpA activities.

ROC analysis is a method for evaluating successful discrimination (Methods). The machine learning model (SVM) predicts chemicals that were not in the training data. Predictions for these new chemicals are then compared to the ground truth (e.g. “prolonged activator” / “not prolonged activator”). Success is defined by high positive (sensitivity) and low false positive (1-specificity) rates. Subsequently, an ROC plot shows the relationship between these two rates. The best possible performance is an area under the curve (AUC) of 1.0 (Methods). When we evaluated the model using this method, the high AUC suggested prolonged activators are physicochemically distinct (Figure 3.3D) (avg AUC = 0.958, Shuffled Activities, avg AUC = 0.592). But this is particularly true when considering physicochemical properties (e.g. 3D geometries) other than enriched 2D substructures or motifs, as indicated by the clear overlap in Figure 3.3A.

3.2. Discussion

Interestingly, butyric acid is a component of human sweat (Cork & Park, 1996), which has been shown to activate as well as inhibit several sensilla trichodae in *A. gambiae* (Meijerink & Van Loon, 1999; Van Den Broek & Den Otter, 1999). Although human sweat is highly attractive to anthropophilic mosquitoes (Braks & Takken, 1999; Healy, Copland, Cork, Przyborowska, & Halket, 2002), it is not clear what role carboxylic acids

play in the attractiveness of this host-odor blend. For example, there are several conflicting studies as to the attractiveness of carboxylic acids to mosquitoes where in some cases carboxylic acids are actually unattractive (Healy et al., 2002; Mboera, Knols, Takken, & Della Torre, 1997; Smallegange, Qiu, van Loon, & Takken, 2005). The varied attractiveness to human skin odors could be attributed to intraspecific preferences for certain human hosts as their emanations differ from individual to individual (Acree, Turner, Gouck, Beroza, & Smith, 1968; Besansky, Hill, & Costantini, 2004; Dekker et al., 2001; Qiu et al., 2004; Takken & Knols, 1999). No study, to our knowledge, has looked at the attractiveness of carboxylic acids (or human odors) as it pertains to activation or inhibition of neurons in the maxillary palp. It is unclear from these and other studies if behavioral responses observed result from a direct repellent effect or another mechanism whereby the insects are failing to respond to normally attractive cues such as CO₂. Perhaps levels of butyric acid from person to person can contribute to host preference in the mosquito as a means of CO₂ response augmentation. Future behavioral assays will be required to test this hypothesis.

Although the substructure that was enriched among the prolonged activators differed subtly from cpA activators and inhibitor, more rigorous 3D analyses indicated the presence of distinct physicochemical attributes for each. When incorporating these features into a machine learning model, we observed high success rates for classifying prolonged activators from other cpA activities. The degree of success implies cpA prolonged activation is indeed related to a set of physicochemical attributes, and machine learning could therefore play an important role in identifying new ligands. The prolonged

activator represents an interesting class of ligand, though there are currently few examples. Machine learning pipelines could predict new prolonged activators and help resolve even finer distinctions from cpA activators and inhibitors. This would subsequently have long-term implications for mosquito vector control strategies.

3.3. Figures

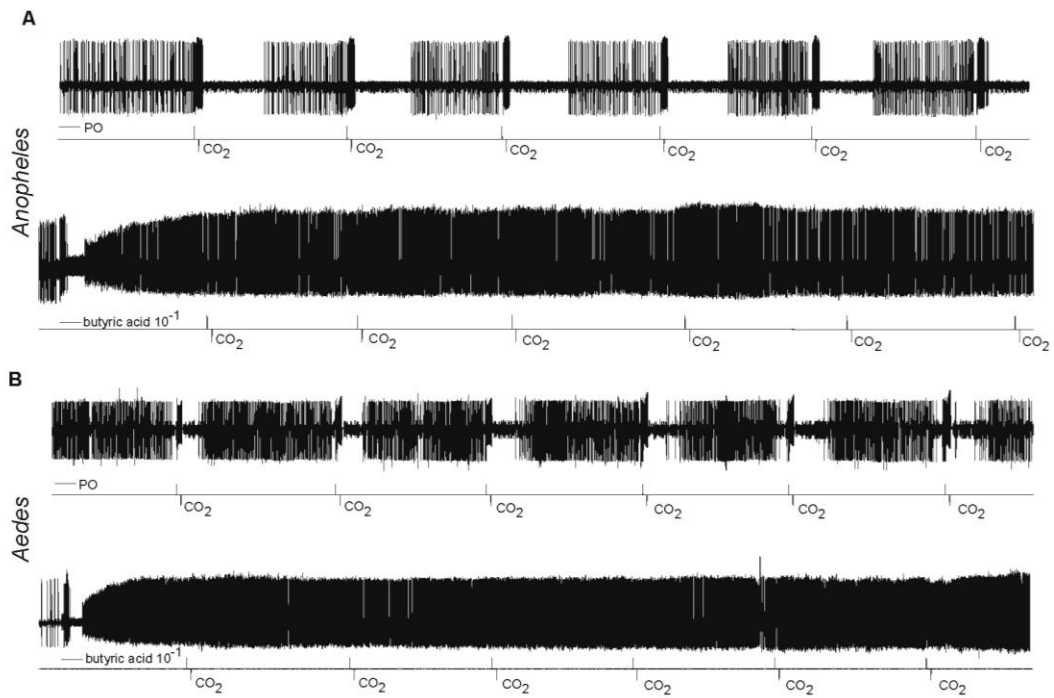


Figure 3.1. Butyric acid is an ultra-prolonged activator of the CO₂ sensitive neuron in *A. gambiae* and *A. aegypti*. A,B, Long-term traces from the cpA neuron of *A. gambiae* and *A. aegypti*, respectively. A 3-sec stimulus paraffin oil top or butyric acid (4ac) bottom is given followed by 1-sec pulses of 0.15% CO₂ every 30-sec. Odor diluted 10⁻¹.

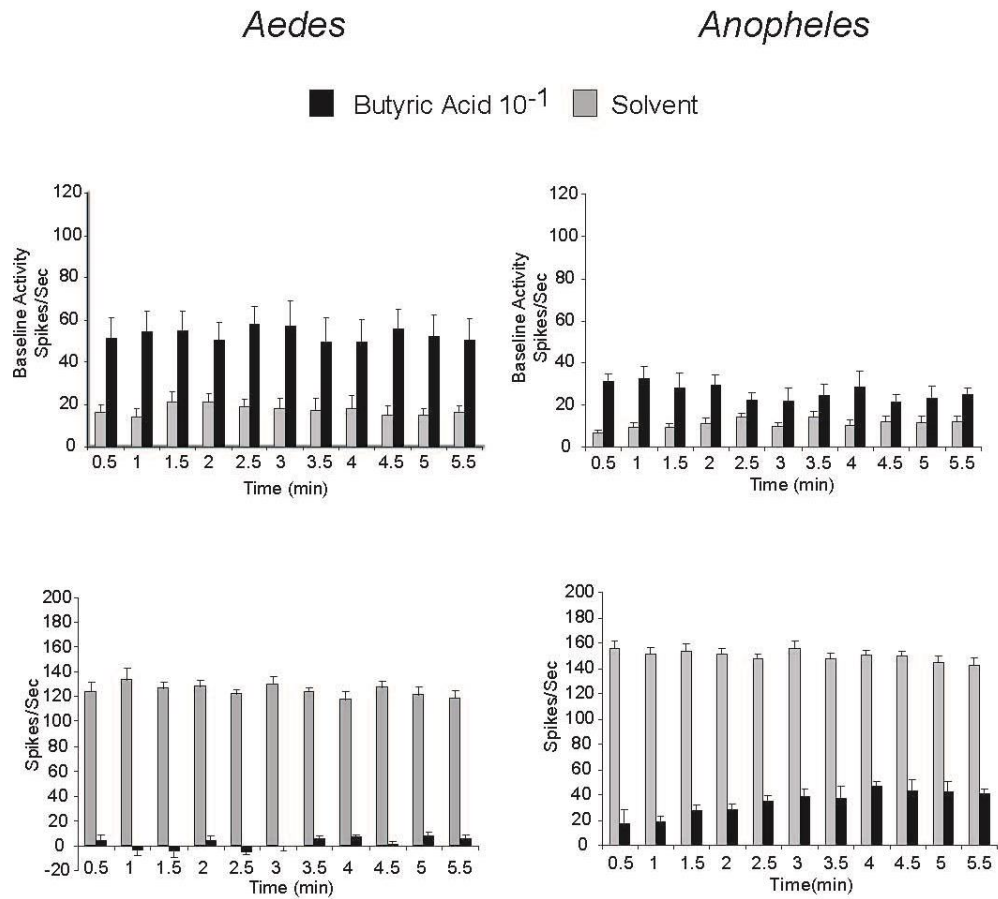


Figure 3.2. Butyric acid is an ultra-prolonged activator of the CO₂ sensitive neuron in *A. gambiae* and *A. aegypti*. **A**, Mean baseline activity of the cpA neuron counted every 30-sec interval after pre-exposure to a 3-sec stimulus of butyric acid (10^{-1}) or paraffin oil (PO) solvent. **B**, Mean change in frequency of response of the cpA neuron to stimulus of 1-sec 0.15% CO₂ applied approx. every 30-sec, following a 3-sec pre-exposure to butyric acid (10^{-1}) or paraffin oil (PO) solvent. n=5, error bars=s.e.m.

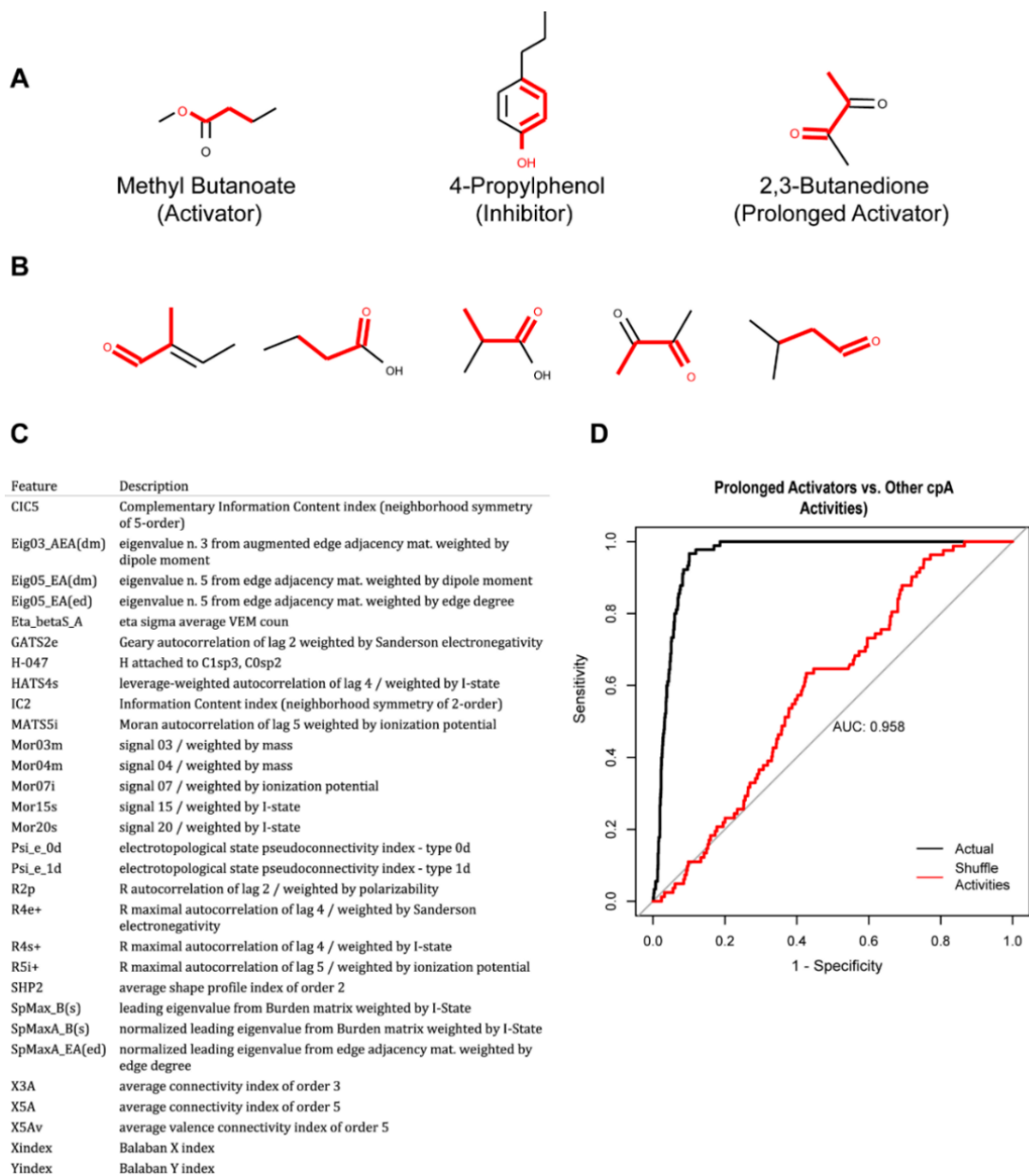


Figure 3.3

Figure 3.3. Ultraprolonged activators of the CO₂ neuron have a shared substructure and can be modelled computationally. **A**, cpA activators, inhibitors, and prolonged activators; enriched substructures in red. The activators, inhibitors, and prolonged activators have similar enriched substructures or simple 2D structural features. **B**, Additional prolonged activators; enriched substructure in red. **C**, Table of top 2D/3D chemical features to discriminate prolonged activators of cpA from the activators and inhibitors. **D**, Support vector machines (SVM) are iteratively fit on a portion of chemicals, “training,” and then predictions are made for the chemicals excluded from training; the quality of the predictions is assessed using ROC analysis. The plot shows the performance across 3 SVM models trained with slightly different chemical feature combinations (black colored curve). Random or chance level performance is estimated by training these SVM models on shuffled activity labels (red colored curve). Diagonal line is the theoretical random performance (AUC = 0.50). The y-axis (Sensitivity) is the true positive rate whereas the x-axis is the false positive rate (1-Specificity). Each point along the curve is from computing these rates at different probability score cutoffs; the probability scores (0-1.0) are assigned by the SVM model to new chemicals. These scores are the predictions that a chemical is a super activator of cpA. It is expected that high scores are assigned to super activators and low scores to the other cpA activities. The ROC plot tests this expectation. Additional details included in the methods.

3.4. Methods

3.4.1. Mosquitoes

Ae. aegypti wild-type (Orlando strain) and *Anopheles gambiae* (recently renamed *Anopheles coluzzii*) were maintained using standard protocols in an insectary at ~27C, ~70-80% humidity on a 14:10 hr (Light: Dark) photoperiod.

3.4.2. Electrophysiology

Extracellular single-unit recordings were performed as described previously (Turner & Ray, 2009) with few modifications. Chemicals were of the highest purity available, typically >99% (Sigma-Aldrich). Odorants were diluted in paraffin oil at indicated concentration. Unless indicated 50 ul of diluted odorant is applied/cartridge, and each cartridge used for 3 stimuli. A controlled volume of air 5ml/sec was puffed through the odour cartridge containing vapors, and was delivered into a constant humidified airstream of 10ml/sec that flowed over the fly antenna. The odorant vapor present in the cartridge was thus diluted ~3-fold before being passed over the fly (each delivery cartridge was used no more than 3 times; 10^{-1} stimulus = ~0.43 ug equivalent from cartridge/application; 10^{-2} stimulus = ~0.043 ug equivalent from cartridge/application). CO₂ stimulus was pulsed through a separate delivery system that delivered controlled pulses using a PSM 8000 microinjector (variable 2.5ml/sec – 6.5ml/ sec) into the same humidified airstream, from either a 1% or 5% tank of CO₂ (Airgas) . The baseline constant humidified airstream (10ml/sec) was generated from a purified air tank (Airgas) and mixed with a constant controlled volume (5ml/sec) of filtered room air (~0.035%

CO₂). For delivery of binary mixtures of CO₂ with another odorant, we ensured a steady concentration of CO₂ to the fly preparation as described in detail in (Turner & Ray, 2009). Unless mentioned, responses were quantified by subtraction of baseline activity immediately preceding stimulus application from activity during the stimulus. For each odorant that had a long-term effect on CO₂ response, each recording was obtained from a naive insect.

3.4.3. Chemical informatics

Chemicals were analysed for maximum common substructures using RDKit (Python) (Landrum, 2006). The algorithm performs an exhaustive search for enriched structural patterns over a set of chemicals. For larger, more diverse sets of chemicals specifying a threshold value can help the algorithm converge on more substantive structural patterns. Here, we set the threshold at .5, which ensures that half of the chemical set should contain the pattern. This algorithm was run separately for activators, inhibitors, and prolonged activators of cpA. The distinction between the 3 (activators, inhibitors and prolonged activators) was based on the spikes/sec calculation, where inhibitors reduce activity below the baseline firing rate and activators increase activity above this rate; the super activators significantly above.

Chemical structures were converted into 3D optimized geometries using RDKit (Python) (Landrum, 2006). The 3D chemical information was then supplied to alvaDesc, which computes ~5,300 physicochemical features. We later removed the features with low variance, high correlations ($r=0.85$) and imputed missing values using the median.

3.4.3.1 Selecting Important Chemical Features

The reduced feature set was then run through the recursive feature elimination algorithm (RFE) over 300 train/test partitions (e.g. 10-fold cross validation, repeated 30 times). Here, the algorithm involves iteratively fitting a support vector machine (Radial Basis Function (RBF) kernel) with different chemical feature sets on the training portion, predicting what remains. Subsequently, the average performance across these different feature sets provides an estimate of the number of features that are needed for successful predictions. This analysis suggested between 20-50 features. The importance of each feature is from the AUC achieved independently. A feature rank is assigned at the end of the cross-validation iterations. Machine learning algorithms for feature selection are from the caret (Kuhn, 2008) and kernlab (Karatzoglou et al., 2004) packages in the R programming language and similar to the way it has been used for ligand prediction of human odorant neurons (Kowalewski & Ray, 2020a).

3.4.4. Machine Learning

After selecting the physicochemical features that are important for the task, models are trained using these features, and predictions are made for chemicals that are not in the training set to evaluate whether learning has indeed occurred. Here, three support vector machine models are fit, sampling different physicochemical features. The individual predictions (probability scores) are then averaged. Each support vector machine learns a decision boundary from the physicochemical features at training. To validate, new chemicals are repeatedly projected into this space. The location of this new chemical

relative to the decision boundary provides the prediction, which is compared to an observed value or label (e.g. ground truth).

In machine learning terminology, cross validation refers to training models iteratively on subsets of data and then predicting new chemicals with each of the trained models. Dividing the data into 10 different training and testing subsets refers to 10-fold cross validation. Here, we repeated that procedure 5 times (e.g. 10-fold cross validation, repeated 5 times). By using more than one model, it is possible to diversify the training, gaining more coverage of the data, getting better estimates of the error, and ultimately, in most cases, producing more generalizable predictions. Implementations of the machine learning algorithms are from the caret (Kuhn, 2008) and kernlab (Karatzoglou et al., 2004) packages in the R programming language

3.4.4.1. Support Vector Machine (SVM)

The support vector machine (SVM) algorithm uses kernels to facilitate the learning of complex, non-linear decision boundaries. The kernel is a function that projects the chemical data into a new space where non-obvious boundaries among chemicals of different classes are increasingly identifiable. The support vector machines implemented here used the gaussian or radial basis function kernel. This kernel is adjusted during the training phase through the sigma parameter, which determines the influence of chemicals or data points that are far from the decision boundary. This affects the prediction of new chemicals and therefore the proper value is set by removing and predicting a small subset of chemicals while training. An additional parameter, C, defines the cost associated with

incorrect prediction performance. As the cost increases, the boundary adapts to improve performance. However, setting the cost value too high produces irregular boundaries that fail to generalize to new chemicals or data points. The proper cost value is therefore set alongside sigma using the approach discussed above.

3.4.5. ROC Analysis

Receiver operating characteristic (ROC) analysis graphically represents classification success and/or failure by comparing the true positive (y-axis: Sensitivity) and false positive rates (x-axis: 1-Specificity). In this study, it is analysing the success or failure of a machine learning model to classify “super activators” versus other activities on the cpA neuron. The trained machine learning model takes the chemical features of a new chemical (e.g. not in the training data) as input. Then it assigns a probability score to this new chemical based on its similarity to the super activators and other activities on cpA from the training data. Subsequently, the ROC analysis defines cutoffs or thresholds for these probability scores. For example, if the score is above .50, then these chemicals are labelled as super activators or simply positive/active cases. The labels are compared to the observed cpA activity, yielding a tally of true positives and false positives that are converted into rates. In the ROC plot, this information is a single point (x, y). Continuing the above process for multiple cutoffs results in a curve. The success is evaluated as the area under the curve (AUC = 1.0; perfect success).

Typically, the curve is compared to a theoretical random classifier (AUC = .50), and this is shown as a diagonal that bisects the plot area. Because chance-level

performance depends on the classification problem, it may be higher or lower than $AUC = .50$. Some classification problems are, for instance, trivial, particularly if there are few positive and negative examples. The chance performance could match the performance of the actual machine learning model. To address this, we trained the models using shuffled data, while keeping other parameters constant. This showed that the success of the actual model(s) was not attributable to chance.

Chapter 4

Natural repellent and attractant activity of microbial metabolites on human skin

4.1. Introduction

The finding that a microbial metabolite, 2,3-butanedione, which is also detected in human sweat, masked detection of CO₂ by the cpA neuron suggested that other microbial metabolites may meaningfully affect mosquito behavior. This seems particularly true when considering the numerous sensory neural pathways beyond cpA-CO₂ and that mosquitoes display heightened attraction to some humans according to unique chemical signatures on skin. Studies have shown, for instance, that certain humans are highly attractive to *Anopheles Gambiae*, the mosquito vector for the malaria parasite. The authors suggested a small number of attractive chemical classes including esters may in part explain the strong difference in attraction (Verhulst et al., 2011). Notably, the opposite observation—innate differences in repellency among humans—has been observed as well (Logan et al., 2008), but the interpretation is complex, as this could result from increased repellent compounds, reduced attractive compounds, an increase in compounds that mask attraction, or a combination of each possibility. A strategy based on recreating natural or innate repellency has significant utility in vector control.

Demand for new chemical repellents has been steadily increasing. From the years 2004-2016, the CDC reported cases of vector-borne illnesses in the United States and territories approaching 1 million, with many others going unreported (Rosenberg et al., 2018). This nevertheless is a fraction of the global incidence. Annually, *Aedes aegypti*

mosquitos account for 100s of millions of vector-borne illness cases worldwide, according to WHO statistics. Insect vectors are currently managed with insecticides and repellents. However, genetic drift and overuse has increased resistance. Pyrethroids, synthetic derivatives of floral extracts with insecticidal and repellent activity, are becoming less effective against mosquito vectors. The mutations that confer insecticide resistance have also been shown to affect insect responses to well established chemical repellents such as N, N-Diethyl-meta-toluamide (DEET).

While a bio-inspired approach based on studying the chemicals on human skin that naturally repel insects, may be lead to the best long-term outcomes in terms of human health and safety and resistance, this data could be too limited currently to meet the demand for new repellents. Additional sources of safe chemicals are those approved for use as flavors and fragrances, both natural and synthetic. But consideration should also be given to chemical libraries far exceeding the number of known flavors and fragrances.

Chapters 2 and 3 outlined an approach to study repellency through electrophysiological data. The recorded activity helped uncover neural pathways that are strongly associated with mosquito and fruit fly behavior, alongside computational modeling. If considering the complex chemical mixtures on human skin that are responsible for differing degrees of mosquito attraction, in addition to the likely numerous pathways these chemicals target, computational modeling plays a critical role. Specifically, in analyzing the human skin microbiome and relevant microbial metabolites and known skin volatiles for repellents.

In this chapter, I focus on the microbes, microbial pathways, and metabolites that are potentially relevant. These metabolites are divided into 2 categories: (1) metabolites that are broadly sourced to a microbe abundant on human skin and (2) metabolites that are volatile organic compounds detected on human skin. Further, I analyze the human gut microbiome for repellent metabolites due to the generally desirable safety profiles of endogenous chemicals, and that the skin microbiome data is incomplete. Therefore, initially any commensal microbe may be of relevance. This study lays the groundwork for the future when appropriate microbes may be used for production of the compounds in bioreactors or used as part of a skin microbiome transplant that confers repellency. However, since it is also possible that the current data is too limited to meet immediate demands, I conclude with a proof of concept that machine learning can be applied to successfully predict the odor profiles of repellent chemicals. Then this is followed by prediction of 10+ million purchasable chemicals for repellency, applying machine learning models to filter for toxicity and unpleasant odor profiles. Such large-scale prediction rapidly expands the space of possible chemical repellents, immediately aiding in the analysis of physicochemical features that might be associated with insect behavior.

4.1.1 The human microbiome as a source of novel insect repellents and attractants

DEET and other effective repellent chemicals act on complex neural circuitry and efforts to isolate receptors or specific pathways have proven difficult. This has particularly slowed progress in developing synergistic and ecologically safe chemical mixtures, as discovery for synergistic combinations fundamentally depends on

knowledge of the pathways and protein targets that drive repellency. One way around this fundamental knowledge gap is to look for and study naturally occurring chemical mixtures that are already known to modify insect behavior. For instance, complex microbial and metabolite compositions on skin are meaningfully related to insect behavior (Verhulst et al., 2010, 2011). But unraveling these relationships also presents obvious experimental challenges. It is for one unclear which of the thousands of chemical and microbial possibilities might be affecting mosquito behavior; the choice of analyzing some possibilities in-depth rather than others is arbitrary. The question then is if computational methods could be used to annotate skin microbial metabolites, identifying which ones are of known or potential relevance to insect behavior.

4.2. Results

Our understanding of the chemical and microbial compositions on human skin is still emerging, but my colleagues and I reasoned that we could develop a theoretical space that identifies possible repellents as well as the enriched microbes and pathways as a map to guide future research. We started with the 10,000 chemicals in the KEGG databases, which include microbe, metabolite, and pathway annotation and next studied the skin microbiome literature to identify microbes that may contribute to insect repellency (Figure 4.1A). An in-depth analysis of the top predicted microbial metabolites (Figure 4.1B, C), suggests many can be found in *Cornyebacterium*. Based on the Euclidean distance between physicochemical features of known repellents, the top metabolites closely resemble anthranilates and (+)-nootkatone (Figure 4.1C). Some

metabolites such as jasmonates were also ranked highly but have known repellency. The biochemical pathways for the top metabolites are diverse, including tryptophan and carotenoid metabolism pathways as well as quinone and terpenoid-quinone biosynthesis (Figure 4.1C).

The chemicals identified in this analysis represent a diverse set of metabolites from species of skin microbes enriched on human skin. Importantly, these metabolites may not all be detectable on the surface of human skin. Due to the diversity, the metabolites may have high or low vapor pressure and the mode of potential repellent activity would then vary. Metabolites with high volatility would be expected to act more spatially, targeting olfactory system whereas those of lower volatility would act primarily through taste or contact, targeting ionotropic receptors (Irs) or gustatory receptors (Grs). It is therefore of interest to (1) clarify the metabolites that have been detected on human skin and (2) identify the putative receptor pathways these chemicals may be acting on.

The collection of volatile molecules detected on human skin (De Lacy Costello et al., 2014) provided some promising leads; high ranking predictions included synthetics likely originating from cosmetics or topicals such as the paraben isopropyl 4-hydroxybenzoate (PubChem CID: 20161) and also fragrances such as hex-3-enyl 2-hydroxybenzoate (PubChem CID: 103379) (Figure 4.2A). When filtering down to molecules also sourced to human skin microbes, top candidates included vanillate and 4-hydroxybenzoate (Figure 4.2B), although, in general, these microbially sourced molecules had lower predicted repellency compared to some of the synthetics appearing on skin in Figure 4.2A. Interestingly, these molecules are linked to several species of

cornyebacterium, which are particularly well suited for laboratory culture and genetic engineering.

Chapter 3 reported on a subset of small molecules detected on human skin that led to prolonged activation of cpA, the neuron in mosquitoes that detects CO₂ through a complex of three gustatory receptors (Grs). Due to the potential for prolonged activation to mask CO₂ detection and effect attraction behavior, it is therefore important to identify the molecules on human skin with physicochemical properties that may act on the CO₂ -detection pathway. To improve the mapping of molecules sourced to human skin and mosquito behavior, the cpA activity prediction model (Chapter 3) was used to screen the ~1000 volatiles reported in the literature as detectable on human skin (De Lacy Costello et al., 2014). Known prolonged activators were assigned higher probability scores, with several structural derivatives also scoring highly. The highly scoring compound acetoin (PubChem CID: 179), for instance, substitutes one of the ketones in the known prolonged activator 2,3-butanedione with a hydroxide group. Lower scoring chemicals such as 2-pentanone (PubChem: 7895) are known to simply activate cpA, and is a prospective trap lure. Thus compounds with significant structural overlap do still show score differences between prolonged activators and simple activators of the cpA (Figure 4.3A). However, the machine learning models also help categorize the metabolites according to broader activity on the CO₂-detection pathway.

In order to visualize the chemical space of volatiles detected on human skin, they were clustered using ~300 2D and 3D physicochemical attributes. The volatiles are organized into 4 broad groups. The known cpA prolonged activators and best predicted

candidates fall closer in this space, roughly in cluster 4, with some in the bordering cluster (cluster 3) (Figure 4.3B). This is consistent with the observation that activity on the cpA neuron is biased towards certain chemical features. Interestingly, many chemicals predicted to be repellents appear structurally diverse and indeed the cpA neuron is simply one of many relevant pathways in mosquito behavior (Figure 4.3B).

To investigate additional pathways, the volatiles detected on human skin were next analyzed relative to acid sensing. Acid sensing is typically mediated through ionotropic receptors (Irs). However, these receptors are lesser characterized in mosquitoes, so the known ligands of the acid sensing Ir64a/8a pathway from the highly conserved *Drosophila melanogaster*, was used for finding training set compounds instead. The structure-activity data for several odorants is available. We created a model for acid-ligands for insects and the chemical features that were selected to optimally predict the activity are in Table 4.1. Computational validation shows a strong relationship between physiochemical features and activity on the pathway, as evidenced by successful classification of chemicals excluded from training (AUC = 0.99, Sensitivity = 0.97, Specificity = 0.85) (Figure 4.4).

Using this model to predict Ir64a/8a ligands from skin volatiles led to finding several predicted actives (Figure 4.5A). Since mosquitoes are known to use acidic odorants from skin as attraction cues involved in landing, the predicted hits give us an opportunity to identify attractive compounds. When mapping the hits onto the chemical space of human odorants for the known and prospective Ir64a/8a activities, prolonged activators of the cpA neuron and behavioral repellents, it was evident that these

molecules of behavioral relevance, cluster close together (Figure 4.5B). They are nevertheless structurally distinct, with low overlap for the top 5 scoring repellents of skin microbial origin compared to the cpA prolonged activators and Ir64a/8a ligands (Figure 4.5B, color dots). Although not expected to directly impact mosquito host-seeking and differential attraction across humans, the human gut microbiome offers an additional library of natural metabolites for future consideration (Figure 4.6). This opens the possibility for developing even more comprehensive structure-behavior maps to recreate natural chemical mixtures that repel mosquitoes and pests, a task that is greatly accelerated through machine learning models.

4.2.1. Prioritizing candidate repellents by modeling their odor qualities

Microbial metabolites may have unpleasant odors and therefore are less value as topically applied repellents. Modeling approaches to predict human odor perception from chemical structure, physicochemical properties or in vitro human odorant receptor activities have proven successful (Kowalewski & Ray, 2020a) (details to follow in Chapters 5-6), particularly on flavor and fragrance databases. As topically applied chemical repellents must be further characterized by cosmetic descriptions such as odor, I developed a set of machine learning models that provide a proof of concept for prediction of human odor perceptual qualities (odor descriptors). By predicting 146 odor perceptual descriptors, including “Fruity other than citrus”, “Lemon”, “Orange”, “Cinnamon” and unpleasant ones such as “Sickening”, “Rancid”, “Animal”, and “Dirty Linen,” it was possible to

assign a complex odor profile to chemical repellents, and then compare this to the human assigned labels.

While odor qualities remain poorly characterized for most repellent chemicals, among the predicted and tested repellents from our analyses, some have been evaluated by humans. The prediction performance could therefore be assessed using ROC analysis. The perceptual models predicted the observed descriptors for most repellents with a high success rate (Avg. AUC = 0.77) (Figure 4.7A, B). Given this success, it was evident that such a filter could be incorporated into the repellent discovery pipeline to better prioritize confirmed and putative repellents. This result further confirmed the odor perception prediction methods detailed in later chapters (5 and 6).

4.2.2. Mining massive commercially available chemical spaces for novel insect repellents

In our earlier machine learning applications, we emphasized the prediction and verification of chemical repellents from natural sources, as if they are not already approved for human use, obtaining approval is less challenging than synthetics. But the size and diversity of a chemical library as well as the cost and availability of chemicals is fundamentally limiting. Subsequently, we scaled-up the analysis to a 10+ million commercially available chemical space (ZINC 15), canvassing more structural diversity than the metabolites we previously screened. We built a new training set including the chemical repellents we experimentally verified. Then, by applying the filters that we progressively incorporated into the pipeline (e.g. odor perceptual qualities) we identified

a smaller set of candidates (Figure 4.8A). The top predictions were enriched with many anthranilate-like compounds, which formally are aminobenzoates comprised of benzene and amino and ester groups. Some chemicals notably differ in that they contain a carbonyl carbon bonded to nitrogen, resembling amide repellents like diethyltoluamide (DEET) (Figure 4.8B).

4.3. Discussion

The analysis presented here was guided by the observation that some humans are especially attractive to mosquitoes whereas others are not. Although previous work has been done on its chemical basis by studying volatile emanations from humans of differing attractiveness to mosquitoes, these have not yielded bio-inspired solutions. One key issue is the combinatorial complexity of the problem that is best suited for computational modeling. Here, I developed for the first time a machine learning-based study to map connections between skin microbial metabolites, volatile organic compounds on human skin and prospective and known activities on insect sensory receptors/pathways. The analysis suggested that few chemicals potentially on human skin are candidate repellents. This also implies that mosquito host seeking behavior preference could be a complex chemical puzzle, rather than be due to one or few volatiles. Therefore, efforts to develop an even more comprehensive mapping of human skin chemicals is essential, with a key role for computational modeling going forward. This work will help drive research into safe, biological repellent strategies and provides a

template for more advanced study as more data emerges on both skin microbes and volatiles on human skin.

Importantly, the work is limited to the availability of data. The interpretation therefore depends on understating this uncertainty. It is clear with respect volatiles on human skin and microbial species on human skin that there are more chemicals and species of note than included in this study. Similarly, machine learning models do not overcome the uncertainties of experimental studies. For receptor activation studies (cpA and Ir64a), smaller training sets will lead to uncertainty in computational models. The computational validation of the machine learning models supports the conclusion of accurate prediction however this should be experimentally determined. Nevertheless, the successful application of the computational method here still suggests the plausibility of using these methods to advance repellency research, particularly the notable differences in mosquito behavior toward some humans. The emphasis on metabolites sourced to skin and gut microbiota raises the intriguing possibility of genetically engineering microbes to efficiently produce repellent mixtures, offsetting many additional costs that arise with chemical synthesis in a laboratory.

In general, this study demonstrates the successful development and application of a machine learning pipeline that accelerates research into insect repellency and its physicochemical basis. The methods and data will help identify additional novel repellent chemicals as well as bio-inspired repellent strategies. Here, we emphasized screening candidate chemicals that are most likely to fit multiple requirements rather than one (e.g. repellency). As these requirements steadily increase, it is obvious that massive,

unexplored chemical spaces offer the most promising leads. Our prediction of 10+ million purchasable chemicals further illustrates the essential role of machine learning in future repellency studies and efforts to identify safe, effective repellent chemicals.

4.4. Figures

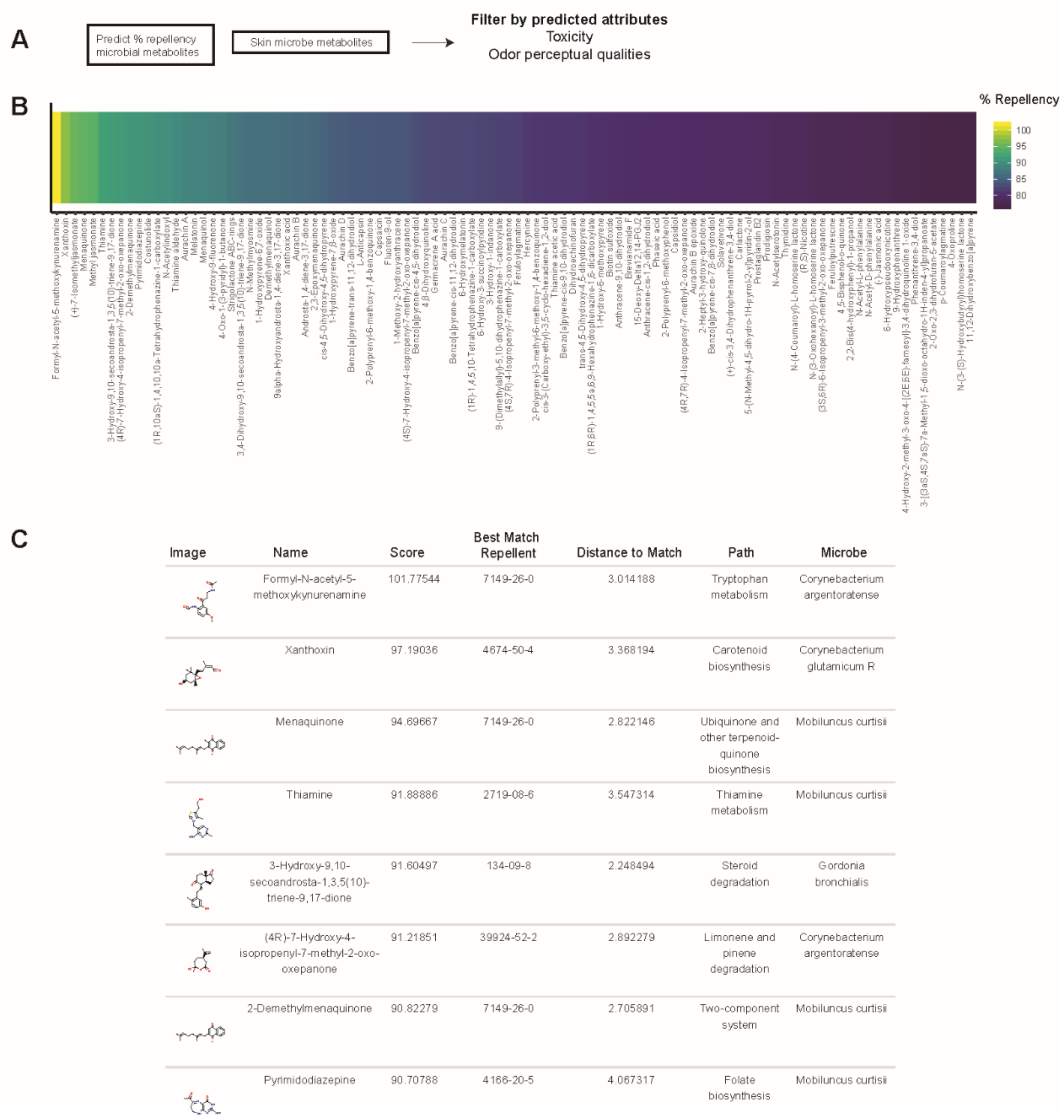


Figure 4.1. Metabolites of skin microbiota include many candidate repellents. **A**, Overview of the approach to predict repellency of microbial metabolites from skin microbiota, filtering the predictions into priority sets according to odor perceptual qualities (Kowalewski & Ray, 2020a) and toxicity (Kowalewski & Ray, 2020b). **B**, heatmap with the predicted % repellency, filtered to the top values. **C**, tabulated top predicted metabolites among skin microbiota, displaying the structures alongside the closest matching known repellent and the Euclidean distance, microbe species/strain and the pathway for the metabolite. Note some pathways and metabolites are also identified in microbial species other than the one listed.

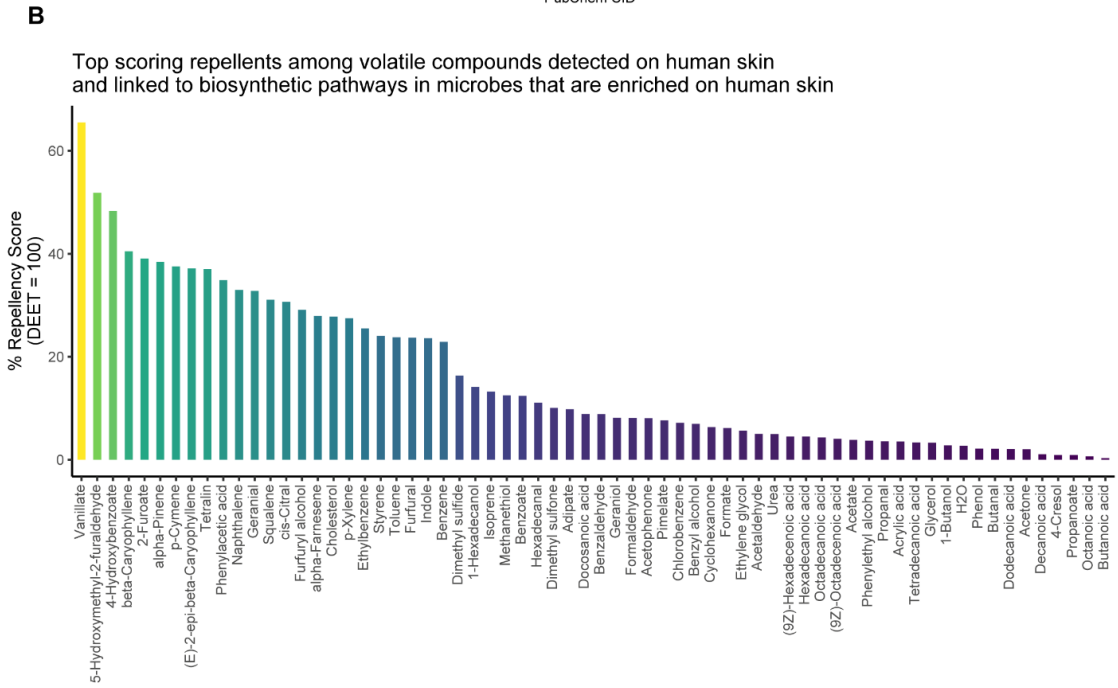
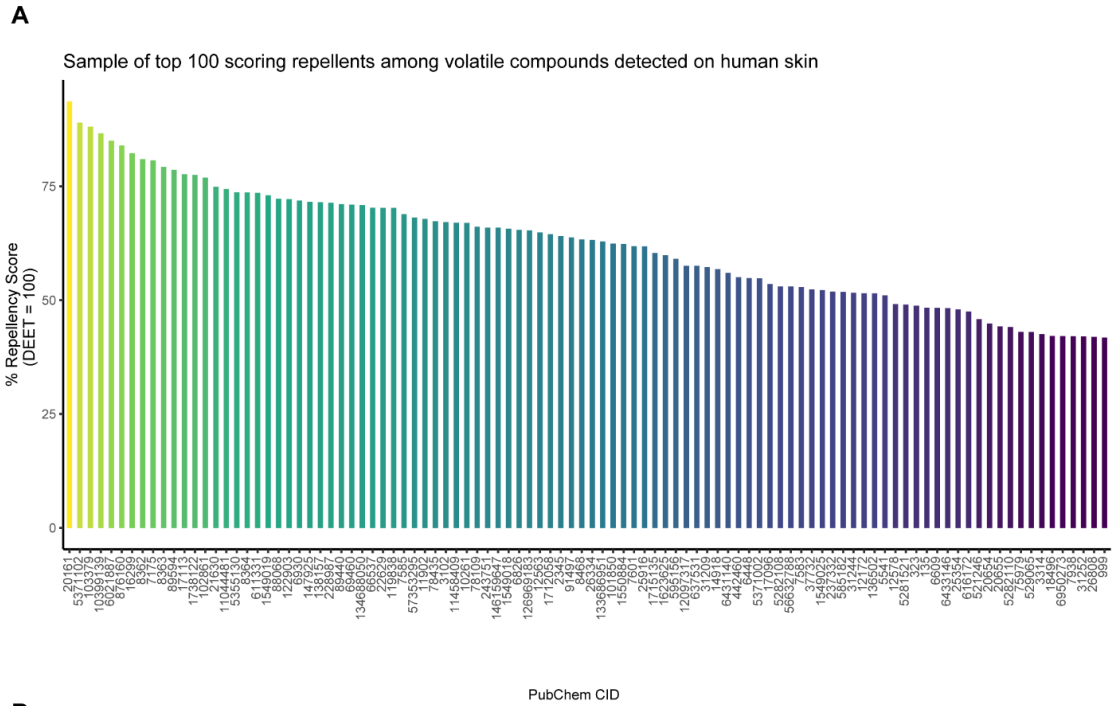


Figure 4.2

Figure 4.2. Volatiles on human skin from biological and synthetic sources are known and prospective repellents. **A**, Repellency, defined as a percentage relative to DEET, is estimated using machine learning models as in Figure 1 but for small molecules categorized as volatile organic compounds (VOCs). These molecules therefore likely act spatially, targeting sensing by odorant receptors. Metabolites in Figure 1 may affect both taste and/or odor sensing. Many top scoring repellents in the plot are synthetic (e.g. sourced to cosmetics). Heat colors emphasize the difference in repellency scores. **B**, Volatiles detected on human skin that are linked to biosynthetic pathways in microbial species enriched on skin. As before, the repellency is the percent relative to DEET; heat colors emphasize the difference in scores.

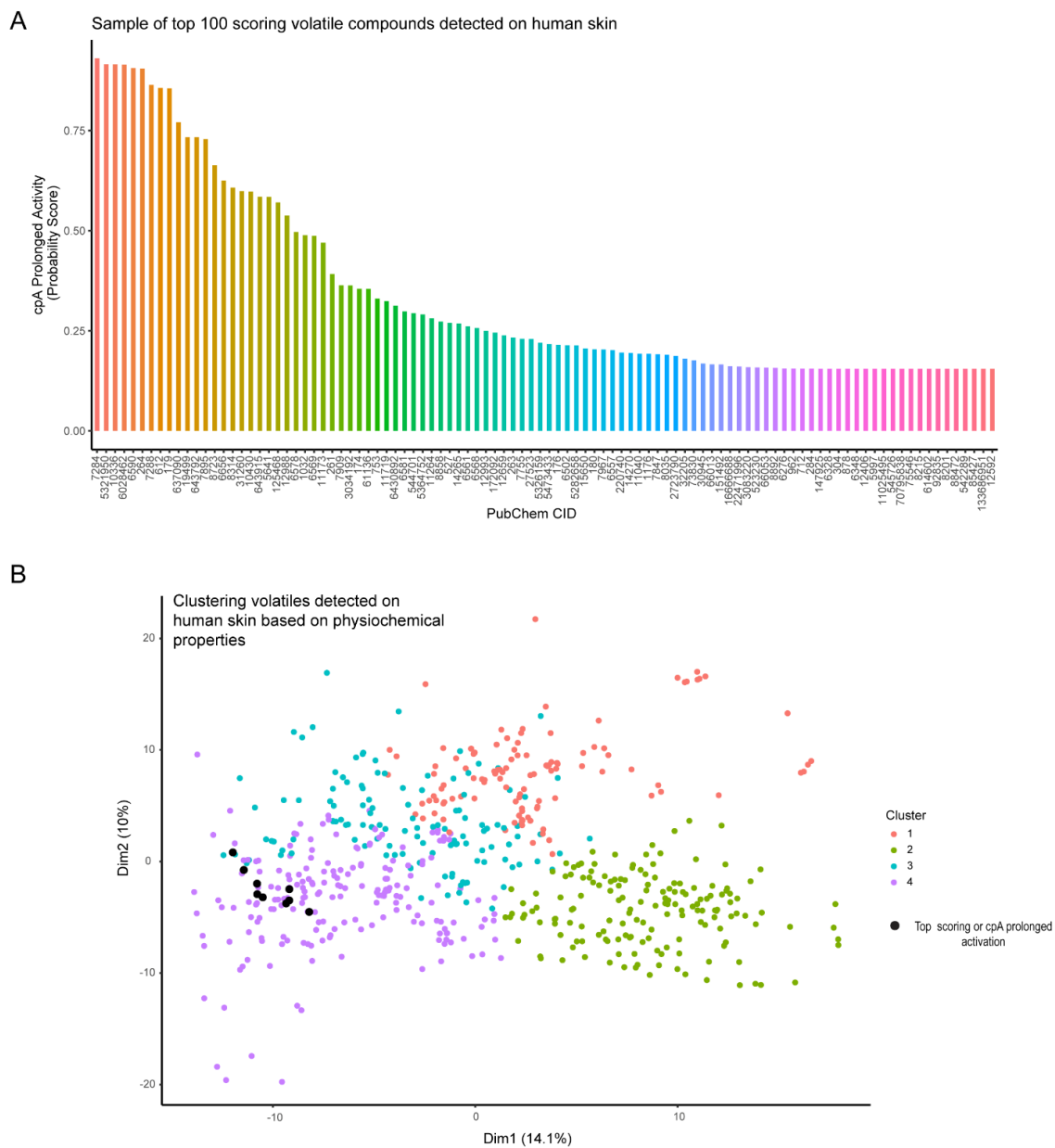


Figure 4.3

Figure 4.3. Microbial volatile metabolites are known and candidate prolonged activators of the CO_2 -detecting neuron in mosquitoes. **A**, The approach described in Chapter 3 to successfully predict cpA neuron activity is applied to further annotate the volatiles on human skin. The cpA activity is an aggregate machine learning score, scaled 0-1.0, with 1.0 indicating high likelihood of prolonged cpA activation. Colors emphasize different scores. The models ranked known prolonged activators with the highest scores. High-Intermediate scores appear consistent with normal cpA activation, as some are known to act like CO_2 as mosquito lures. **B**, Various physicochemical attributes (~300), after eliminating high correlations and low variance, are used to cluster the volatiles detected on human skin based on the k-means algorithm. The approach leads to a 2D plot illustrating the distance between chemicals from their physicochemical attributes. Four broad clusters are supported by the data. The top scoring prolonged cpA activators are shown differently colored (black dots) to indicate their location among the volatiles. These molecules occupy a specific region of the chemical-structural space, consistent with highly specific targeting of the evolutionary conserved CO_2 detecting neuron (cpA).

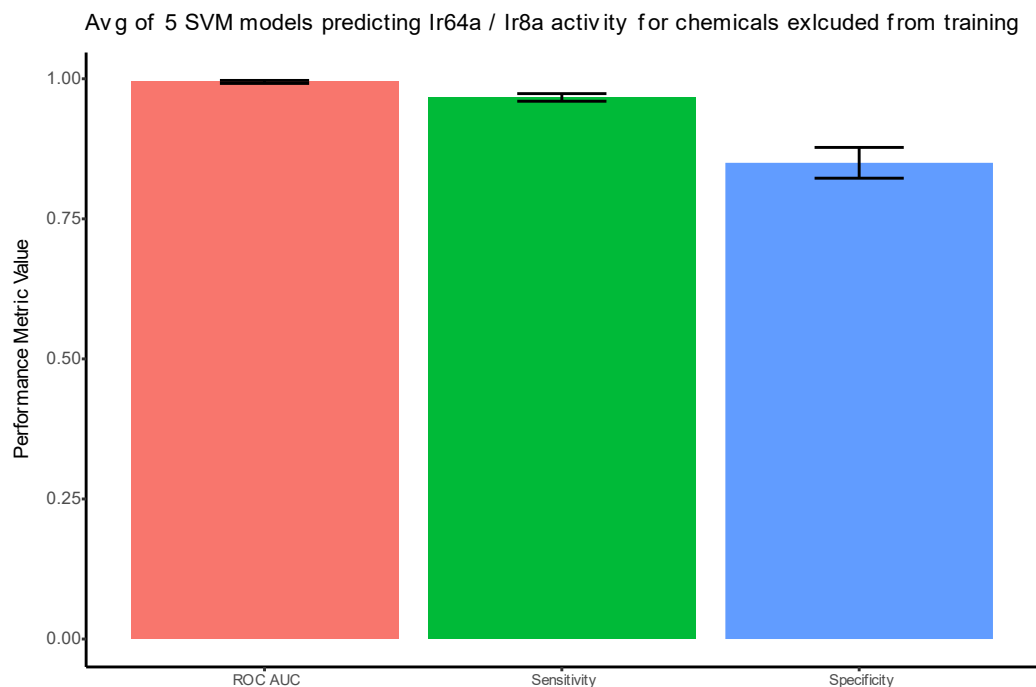
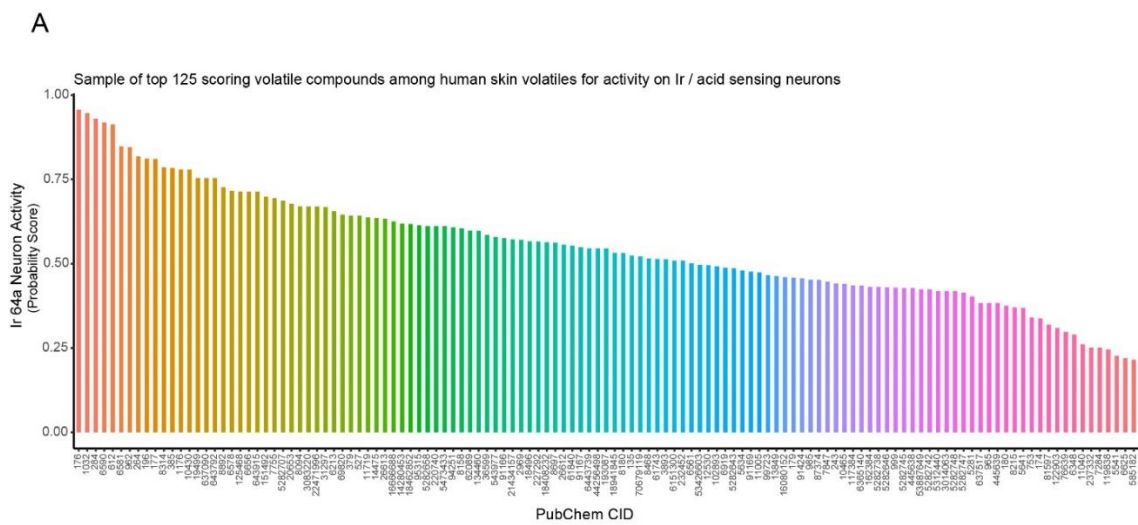


Figure 4.4. Validation of the machine learning models to predict Ir64a / 8a. Five support vector machine (SVM) models are trained on physicochemical features of chemicals with known activity on the *Drosophila* Ir64a / 8a pathway, which detects acids. *Drosophila* is used here as a model system to identify potential activity on acid-sensing pathways for chemical libraries such as volatile organic compounds and microbial metabolites on the surface of human skin. Cross validation procedure is applied to validate the models. This entails repeatedly splitting the data into training and testing portions. Bars represent the average performance for the labeled metric over 30 such training/testing splits. The metrics here summarize the results of the Receiver Operating Characteristic (ROC) analysis, which determines successful classification of the activity on this pathway (“Active” / “Inactive”). The prediction is compared to the experimentally observed result, giving rise to the Sensitivity (True Positive Rate), Specificity (False Positive Rate = 1-Specificity) and the area under the ROC curve (ROC AUC). These values have a maximum value of 1.0, Error bars show the standard error of the mean (SEM).



Clustering human skin volatiles based on their physicochemical attributes
Molecules with known or prospective relevance to mosquito behavior annotated

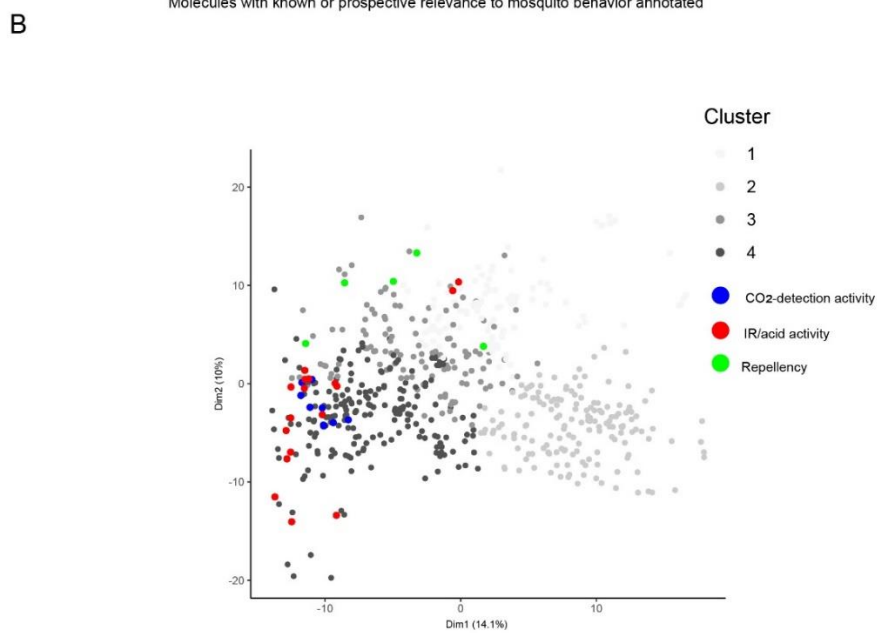


Figure 4.5

Figure 4.5. Adding Ir64a acid sensing activity in drosophila leads to comprehensive map of structure-insect behavior relationships for volatiles detected on human skin. **A**, Ir64a activity from *Drosophila* (fruit flies) is used to train machine learning models, followed by prediction of volatiles detected on human skin. The heat colors help distinguish differences in the scores. The top ranked volatiles are acids, with many having known activity on Ir64a. The scores are scaled 0-1 and are a probability-like value, with the highest score (1.0) indicating the chemicals that are very similar to “Actives” the machine learning model was trained on (e.g. activators of Ir64a). Here, the score is an aggregate of 5 different algorithms. Each algorithm takes a different perspective on the learning problem and the aggregation (averaging) then leads to better generalizability of the predictions. **B**, Clustering of human volatiles on human skin by structure, **shown previously in Figure 4.3**, is annotated further, illustrating how repellency behavior, activity on the CO₂-detecting cpA neuron and activation of Ir64a relate in terms of chemical structure. The four clusters are shown as in Figure 4.3B but in greyscale to better clarify the annotations. The annotations (colored dots) reflect both top known and predicted activities gathered from Figures 4.2-4.3 and the Ir64a model validated in Figure 4.4. The top scoring repellency annotation (green dots) is specially for volatiles linked to biosynthetic pathways in microbes.

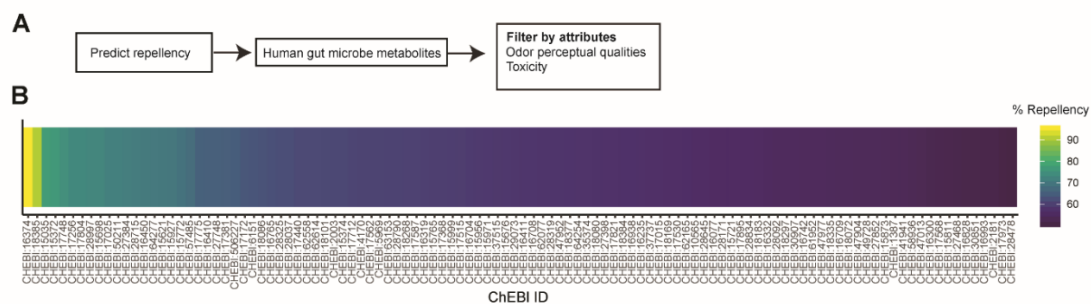


Figure 4.6. Human gut microbial metabolites offer some natural repellent candidates. **A**, The predicted % repellency (scaled relative to the repellent DEET (DEET = 100)) for human gut microbial metabolites. Metabolites are labelled according to the CHEBI identifier. ChEBI. Chemical Entities of Biological Interest is a curated database of biologically sourced chemicals or synthetics of relevance to biological systems.

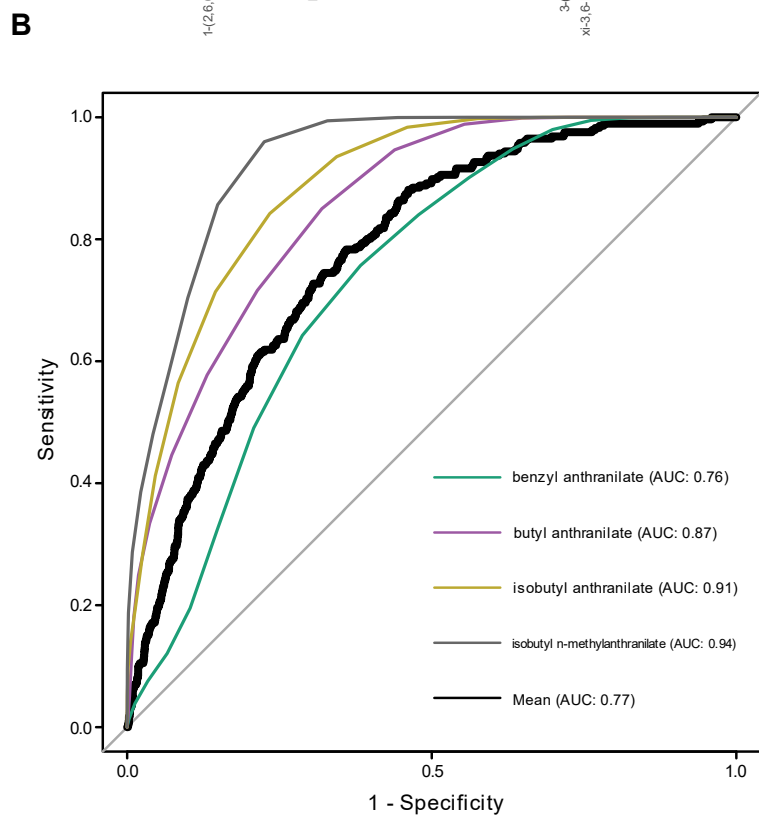
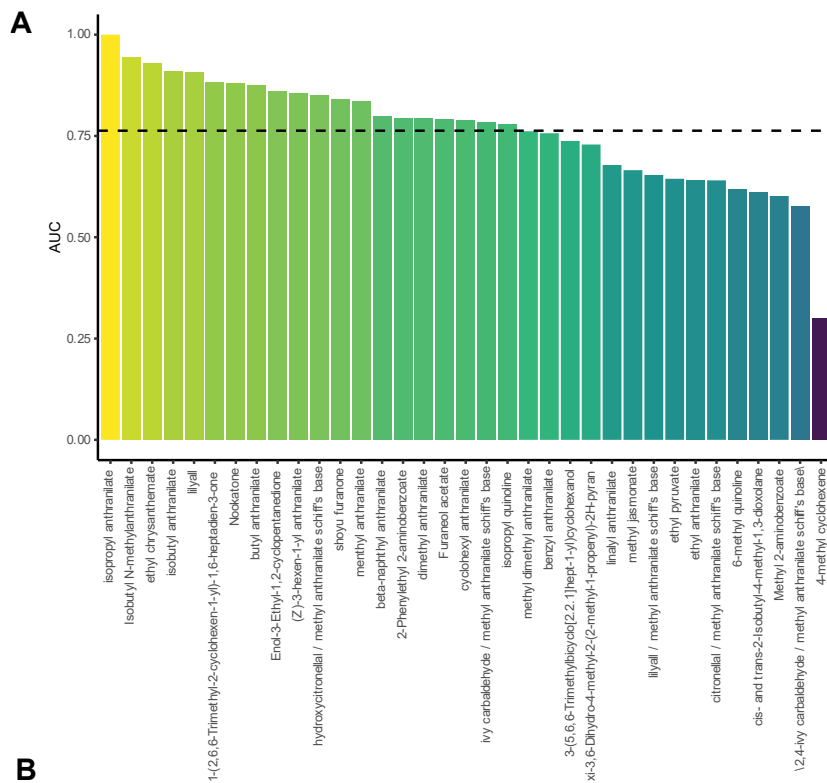


Figure 4.7

Figure 4.7. Models successfully predict odor perceptual qualities of repellents. A, Validation of models predicting the odor perceptual qualities (146 perceptual descriptors) of a set of compounds with newly experimentally verified repellency. The performance metric for the validation is the area under the Receiver Operating Characteristic (ROC) curve (AUC), which assesses the rate of correctly predicted perceptual descriptors (True Positive Rate or Sensitivity) relative the rate of incorrectly predicted perceptual descriptors (False Positive Rate or 1-Specificity). Dashed horizontal line is the mean AUC. **B,** Sample ROC curves for select compounds (colors) as well as the ROC curve across all compounds (black); the area under the curve (AUC) for each example is provided in the plot area. Details on the perceptual data and ROC analysis in Methods.









Image	ZINC_ID	Best Matching Repellent	Distance to Repellent	Score	LD ₅₀ mg/kg.
	ZINC000001394030	Ethyl Anthranilate	2.493159	110.9316	1293.0062
	ZINC000034195873	Ethyl Anthranilate	3.098310	110.5746	711.4268
	ZINC000005327991	Menthy N,N-dimethyl anthranilate	2.621302	110.4276	883.9981
	ZINC000094992683	2719-08-6	2.665383	110.4188	1067.0473
	ZINC000001735256	Menthy N,N-dimethyl anthranilate	2.619820	110.4169	883.8132
	ZINC000008609700	18189-02-1	1.987123	110.2547	1287.3842
	ZINC000253417706	Menthy N,N-dimethyl anthranilate	3.032478	109.7302	933.5598
	ZINC000034195871	Ethyl Anthranilate	3.038142	109.6462	684.8090
	ZINC000000506855	7149-26-0	2.630272	109.5113	1299.5812
	ZINC000191321911	7756-96-9	2.766004	109.5082	989.3195

Figure 4.8. Screen of 10+ million purchasable chemicals accelerates new discoveries in insect repellency. A, Tabulated predicted % repellency (relative to DEET) from a library of 10+ million chemicals, filtered to the top values. The best matching known repellent is displayed alongside the Euclidean distance, the predicted LD₅₀.

4.5. Tables

Table 4.1

Feature	Description
GATS2e	Geary autocorrelation of lag 2 weighted by Sanderson electronegativity
GATS4v	Geary autocorrelation of lag 4 weighted by van der Waals volume
GATS5v	Geary autocorrelation of lag 5 weighted by van der Waals volume
GGI3	topological charge index of order 3
H-047	H attached to C1sp3, C0sp2
MATS4i	Moran autocorrelation of lag 4 weighted by ionization potential
P_VSA_LogP_4	P_VSA-like on LogP, bin 4
P_VSA_m_2	P_VSA-like on mass, bin 2
P_VSA_s_3	P_VSA-like on I-state, bin 3
PDI	packing density index
PW4	path/walk 4 - Randic shape index
SHED_DA	SHED Donor-Acceptor
SIC1	Structural Information Content index (neighborhood symmetry of 1-order)
SpMAD_B(s)	spectral mean absolute deviation from Burden matrix weighted by I-State
SpMAD_EA(dm)	spectral mean absolute deviation from edge adjacency mat. weighted by dipole moment
SpMax_B(s)	leading eigenvalue from Burden matrix weighted by I-State
SpMaxA_AEA(ed)	normalized leading eigenvalue from augmented edge adjacency mat. weighted by edge degree
SpPosA_B(p)	normalized spectral positive sum from Burden matrix weighted by polarizability
VE1_B(p)	coefficient sum of the last eigenvector (absolute values) from Burden matrix weighted by polarizability
X2A	average connectivity index of order 2

Table 4.1. Top chemical features for predicting activity on Ir64a / 8a pathway. Features supplied are from the alvaDesc software package and were selected from > 5000 using the recursive feature elimination (RFE) algorithm and a support vector machine (SVM). This entailed fitting 100 SVM models on different portions of the data, predicting a “held out” or test set of chemicals; finally, rank ordering the features based on their area under the receiver operating characteristic curve (ROC AUC) over the 100 iterations. The “Description” is the detail for the shorthand label assigned to the features.

4.6. Methods

4.6.1. Repellency

The behavior data for the predictions of repellency are derived from the following:

Percentage repellency = $100 \times [1 - (\text{mean cumulative number of mosquitoes on the window of treatment for 5 seconds at time points 2,3,4,5 min} / \text{mean cumulative number of mosquitoes that remained on window of solvent treatment for 5 seconds at time points 2,3,4,5 min})]$. For computational modeling, the percent repellency was then scaled relative to DEET, setting, that is, 100% repellency as the avg. score obtained for the tests using DEET. This led to a more meaningful interpretation for the predictive scores.

4.6.2. Chemicalinformatics and Machine learning

Physicochemical features were selected using alvaDesc software, which offers > 5000 chemical features. Optimal chemical features were then selected using cross-validated recursive feature elimination (CV-RFE). The RFE algorithm takes a partition of the data and a subset of chemical features, and fits a model using a preferred machine learning algorithm; in this study, support vector machine. The portion of data that was not used to fit the model serves as a test set. Over 100s of such iterations the algorithm identifies an optimal number of features to maximize predictive success. Additionally, for every round the contribution of each chemical feature is assessed, giving rise to an aggregate rank; namely the frequency at which a feature has been identified as important to the predictive success.

The support vector machine (SVM) algorithm does not natively evaluate the contribution of a feature to successful predictions; some algorithms such as Random Forest (RF) permute (or shuffle) a feature's values, providing a simple comparison between the actual and permuted cases. This is then a proxy for importance. For the SVM algorithm, this must be done by the analyst, and a custom importance metric must be defined. Here, if the predicted outcome was a number (e.g. % Repellency), this was the pseudo R2, "pseudo" as it is the non-linear regression approximation of the R2 in linear regression. If it was classification (e.g. classifying metabolites as "Prolonged Activators" / "Not Prolonged Activators" of the cpA neuron), then it was the AUC for each feature on its own. Additional details are available in the documentation for caret package in R (Kuhn, 2008).

4.6.2.1. Machine learning predictions of odor perceptual qualities

Perceptual training data is from a study published in a reference book: ATLAS of odor character profiles (Dravnieks, 1985). In the study, panellists supplied ratings for 150 odorants and mixtures across 146 possible descriptors; the ratings are reported as the % usage. The % usage refers to the portion of panellists that rated an odorant (1-5) using a particular descriptor (1-146), indicating its relevance to the odorant. It is on the scale 0-100, with the maximum of 100% suggesting every participant found the descriptor to be relevant. Models were trained to classify chemicals in the top 10% of the % usage distribution; therefore, it is classification of chemicals that are highly "Fruity" versus those that are not, according to the human panellists. These trained models were

subsequently evaluated by comparing the labels that humans assigned to a set of chemical repellents using Receiver Operating Characteristic (ROC) analysis.

4.6.2.2. ROC analysis

The Area under the ROC Curve (AUC) assesses the true positive rate (TPR or sensitivity) as a function of the false positive rate (FPR or 1-specificity) while varying the probability threshold (T) for a label (Active/Inactive). If the computed probability score (x) is greater than the threshold (T), the observation is assigned to the active class. Integrating the curve provides an estimate of classifier performance, with the top left corner giving an AUC of 1.0 denoting maximum sensitivity to detect all targets or actives in the data without any false positives. The theoretical random classifier is reported at AUC = 0.5.

$$TPR(T) = \int_T^{\infty} f_1(x) dx$$

$$FPR(T) = \int_T^{\infty} f_0(x) dx$$

Where T is a variable threshold and x is a probability score

4.6.3. Clustering

Chemical features were computed in alvaDesc (> 5,000). The resulting chemical feature matrix is sparse and the features can be highly correlated. To ensure that the features were most relevant to the metabolite set (linked to human skin microbiota), low and zero variance as well high correlation (>.85) filters were applied, reducing the matrix down to ~300 information-rich chemical features. Clustering was then performed on the Z-

normalized matrix using the k -means algorithm. The “ k ” or number of clusters that best represents the data is estimated from the data. This entails setting k to different values, evaluating the quality of the clustering at each k according to the total within group cluster variability. The goal is to select the k that minimizes the within group variability (e.g. sum of squares), which implies tighter clusters. The variability here is the Euclidean distances among the metabolites based on the values of the ~ 300 chemical features. Here, the optimal k was 4.

Chapter 5

Predicting human odor perception of odorants including repellents

5.1. Introduction

Although an understanding of repellent receptor targets and the associated neural pathways is important, additional considerations such as the perceptual qualities of a candidate repellent are of potentially greater concern, particularly among biologically sourced chemicals. Many of the microbial metabolites introduced in Chapters 3 and 4, for instance, are characterized by strong, unpleasant scents, which limit their value as topically applied repellents. The predictability of odor-perceptual qualities from physicochemical properties was successfully applied to a set of newly discovered chemical repellents in Chapter 4. But there is a longstanding and comprehensive study of the physicochemical basis of smell predating that success. Not until the advent of modern machine learning has it become increasingly possible to evaluate the extent that chemical structure significantly determines odor perception. Despite correlations between functional groups and some odor perceptual qualities, one hypothesis is that human perception is unique. Specifically, that while prediction from chemical structure may be successful, these cases could represent outliers. In this chapter, I outline my work to study this claim in detail, further building on the value that this technology brings to repellent discovery.

6.1.1. A system-wide analysis of the human olfactory perceptual space

Human perceptual descriptions for olfactory stimuli are less stereotypic than for vision or auditory stimuli where perception can be predicted by clearly defined properties such as wave frequency. In fact olfactory perception may vary without an apparent relationship to the physicochemical properties of an odorant nor the molecular and cellular organization of the olfactory system (Buck L and Axel, 1991; P Mombaerts, 1999; P Mombaerts et al., 1996; Peter Mombaerts, 2001; Vassar, Ngai, & Axel, 1993). Yet the olfactory capabilities of humans appear to be close to that of species that rely heavily on olfaction for survival and mating (McGann, 2017). Genetic variation in olfactory receptors also explains a significant amount of variability in basic perceptual qualities like intensity as well as more complex perceptual qualities (Mainland et al., 2014; McRae et al., 2012; Trimmer et al., 2019). While culture and language also affect olfactory perception (Majid & Kruspe, 2018), individuals often show significant similarities in perceptual descriptions for the same chemical (Dravnieks, 1985; Keller & Vosshall, 2016), implying an underlying physicochemical basis for human olfactory perception. In fact, predicting percepts from physicochemical features is becoming increasingly plausible (Haddad, Medhanie, Roth, Harel, & Sobel, 2010; Keller et al., 2017; Kepple & Koulakov, 2017; Khan et al., 2007; Licon et al., 2019; Nozaki & Nakamoto, 2016; Snitz et al., 2013). However, the breadth and complexity of the human olfactory perceptual space and its physicochemical correlates remain poorly understood except for a select few (<10) perceptual descriptors (Keller et al., 2017). Moreover, because of the comparatively limited repertoire of olfactory receptors that have been functionally deorphanized (de

March et al., 2020; Hu et al., 2020; Keller, Zhuang, Chi, Vosshall, & Matsunami, 2007; Mainland et al., 2014; McClintock, Khan, et al., 2020; McClintock, Wang, Sengoku, Titlow, & Breheny, 2020; Pfister et al., 2020; Saito, Kubota, Roberts, Chi, & Matsunami, 2004; Shirasu et al., 2014), experiments with receptors and ligands are presently not an efficient method to comprehensively map physicochemical features to different perceptual qualities. There is subsequently an important role for computational modeling.

Previous attempts to predict ratings of odor perception from the physiochemical features of molecules have been successful to some degree, although these examples represent a small fraction of the perceptual descriptor space (Keller et al., 2017; Khan et al., 2007; Nozaki and Nakamoto, 2016). In these previous efforts, several perceptual descriptors tested were hard to predict and these descriptors may have been difficult to evaluate by study volunteers or lack a strong physicochemical basis. Nevertheless, a natural language processing (NLP) approach could successfully predict perceptual descriptors across studies, suggesting that descriptions of olfactory perceptual content are likely structured and not totally subjective (Gutiérrez, Dhurandhar, Keller, Meyer, & Cecchi, 2018). These earlier studies create an underlying framework that points to the intriguing possibility that the perceptual descriptions humans select to characterize odorants are associated with key physicochemical features, even those that are seemingly abstract and currently not well defined. While prior structure-activity studies predate modern machine learning, indicating features enriched among chemicals with shared perceptual qualities (Rossiter, 1996), some exceptions arise.

A few recent studies have modeled odor perception using large, non-experimental databases, training deep neural networks (Sanchez-Lengeling et al., 2019; Tran, Kepple, Shuvaev, & Koulakov, 2019) to predict perceptual descriptors from chemical features. These efforts have suggested that many complex perceptual descriptors are predictable, but the success can be challenging interpret. The chemical representation or input undergoes further abstraction, in which the predictive features are the network weights. We therefore established a pipeline to clarify the physicochemical properties that best predict diverse perceptual descriptors and to rigorously test using different metrics and controls that ensure the machine learning models are consistent with biological expectations. We find that chemical feature models can address many complex, biologically relevant tasks. As this suggests the important or predictive features that we identify are a resource for further research, we finally annotate a large commercially available chemical database with predicted odor qualities. These predictions reveal enriched structural motifs that help interpret the machine learning models.

5.2. Results

To better clarify the physicochemical basis of diverse perceptual descriptors, we designed a pipeline that begins with the identification of chemical features that contribute most to perceptual descriptors, followed by training machine learning (ML) models to predict percepts from these features and evaluating their predictions (Figure 5.1A, Methods). We used perceptual data from two human studies, Dravnieks (1985) and Keller (2016), conducted at different times and with different participant demographics (Dravnieks,

1985; Keller & Vosshall, 2016). In the Dravnieks (1985) study, fragrance industry professionals rated 137 individual volatile chemicals for 146 different odor qualities (perceptual descriptors). We identified ~30 predictive physicochemical features (from DRAGON software) for each of these perceptual descriptors (see Methods for details). Machine learning models that were trained with the physicochemical features successfully predicted most of the perceptual descriptors as seen by the computational validation (Figure 5.1B) (avg. Area Under Curve (AUC) = 0.81, avg. shuffle AUC = 0.62; $t = 24.17$, $p < 10^{-55}$; top 50 models avg. AUC = 0.90, avg. shuffle AUC = 0.62; $t = 55.54$, $p < 10^{-75}$). We also observed that altering the general classification cutoff from the top 10% usage to the top 15% or 25% changes the AUC value determined for different percepts (Table 5.1). Specifically, of note, is the increase in performance as the cutoff is lowered, suggesting these descriptors in the study dataset have fewer high scoring (% usage) examples for training and the high scoring chemicals may not be as physicochemically distinct as lower scoring (% usage) chemicals. In order to remove bias because of differences in the score distribution, we next evaluated other metrics for the validation such as root mean square error (RMSE), mean absolute error (MAE) and correlation between predicted and observed % usage (R); see details in Methods). Each chemical has a complex perceptual profile, we analyzed the correlation between predicted and observed % usage over the validation for the full (146 descriptor set), which suggested good results (Figure 5.5A). Next, for a set of hidden test chemicals, the predicted olfactory profile over all 146 perceptual descriptors also correlated well with

the known human ratings (avg. $r = 0.72$; best predicted chemical: $r = 0.86$; worst predicted chemical: $r = 0.67$) (Figure 5.1C).

The Dravnieks (1985) study used experienced human raters, and to generalize the utility of our approach we next applied it to the more recent Keller (2016) study of general public volunteers (Keller & Vosshall, 2016). As with the Dravnieks (1985) study, perceptual descriptors for a set of 69 hidden test chemicals (Keller et al., 2017) were also well predicted from physicochemical features (Figure 5.5B) or with multiple train/test sets from all 476 chemicals (Figure 5.5C).

The two studies, though differing significantly in methodology, evaluated a small number of identical perceptual descriptors. It was therefore possible to test whether models from the 1985 study could predict equivalent perceptual descriptors in the 2016 study (Cross-Study). Prior work has performed this analysis on a small number of overlapping chemicals using an approach involving semantic similarity and chemical features (Gutiérrez et al., 2018). We focused on 413 non-overlapping chemicals and more traditional modeling methods to evaluate across studies. Models for “Sweet“, “Warm“, “Sweaty” and “Chemical” trained on the Dravnieks (1985) study were successful at classifying the 413 chemicals unique to the Keller (2016) study (Dravnieks, 1985; Keller & Vosshall, 2016) (Figure 5.1D) (Avg. Cross-Study AUC = 0.73 ± 0.07 , maximum AUC = 0.82 ± 0.03 for “Sweet”). As a control, we compared the cross-study predictions with the Dravnieks (1985) model for a distinct percept, “Varnish,” which achieved good accuracy in Figure 5.1B and is similar to “Chemical” but expected to differ from the rest. Consistent with expectation, the overall average AUC using the Dravnieks (1985)

“Varnish” model cross-study was $0.41 \pm .12$. When we trained the Dravnieks (1985) models on randomly shuffled labels before the cross-study predictions, the overall average AUC was $0.52 \pm .07$ (Table 5.2). These results suggest that identical perceptual descriptors across studies are predictable from a set of physicochemical features, despite differences in study sample demographics and odor diversity.

We next analyzed if the descriptors within each study could be predicted equally well by a different descriptor model with good classification accuracy. For the Keller 2016 study, “Bakery,” which is similar to the many food-related descriptors in the study but differs from the rest, did not classify the 69 test chemicals as well as the percept-specific models (Figure 5.6A). Of the 146 Dravnieks (1985) study descriptors, ~96% were better predicted by the percept-specific model vs “Varnish” (avg. Varnish AUC = 0.51; $t = 21.65$, $p < 10^{-59}$) (Figure 5.6B). However, the “Varnish” model was indistinguishable from “Chemical,” “Paint,” and “Etherish,” implying chemical features are redundant in some cases. As this also suggested some descriptors in an arbitrarily large descriptor space might be predicted equally well by semi or even unrelated chemical feature models, we studied this exhaustively (Figure 5.7A). Overall, predictions with the actual descriptor model was often statistically better, even for some seemingly similar descriptors. However, this is not always the case, suggesting some descriptors may simply lack quality exemplar chemicals.

Apart from these two semiquantitative psychophysical studies, (Dravnieks (1985) and Keller), a large amount of perceptual data is available as text at various databases, some using identical or similar perceptual descriptors. While these databases are not

quantitative or methodical, we tested each of our 146 Dravnieks (1985) perceptual descriptor models on a unique set of 2, 525 chemicals from one such database maintained by the GoodScents company. The predicted perceptual scores of each chemical were evaluated against the known textual data using ROC analysis (Methods). Although this task differed dramatically from previous test datasets, on average, the predictions compared favorably to the observed percepts (AUC = 0.72, $t = 48.53$, $p < 10^{-15}$) (Figure 5.1E). Collectively, these examples of predictive success within and across datasets establish that many perceptual descriptors, even those that are seemingly abstract, have a physicochemical basis that can be identified.

In order to get an overview of the physicochemical basis of odor perception, we created network representations of the relationship between the percepts and the most predictive chemical features (Bullmore & Sporns, 2009; Koulakov, 2011; Meunier, Lambiotte, & Bullmore, 2010; Zhou, Smith, & Sharpee, 2018). For example, we expected that similar descriptors (“Fruity, Citrus”, “Lemon”, “Grapefruit”) were best predicted by similar chemical features and they would cluster together in the network (Figure 5.2A). Initially, we performed simple hierarchical clustering to compare the distances between the perceptual descriptors based on the % usage (Figure 5.8A), and then based on chemical feature sets in the machine learning models for comparison. While some chemical features were selected for multiple descriptor models, resulting in unconventional pairings in the hierarchical tree relative to perceptual ratings, we observed many similarities (Figure 5.8B).

We next turned our attention to the network-based visualizations, reducing the chemical features down to the top 3 for 93 of the most distinct perceptual descriptors in the Dravnieks (1985) study (117 features in total). Despite the limited information, distinct clusters were detectable. In general, networks using more chemical features (top 5 or 10) were better connected (Figure 5.2B, Figure 5.9A). Interestingly, these networks relate well to those assembled only from human participant ratings rather than physicochemical information (Castro, Ramanathan, & Chennubhotla, 2013). Taken together, these analyses suggested that perceptual descriptors with highly correlated % usage (e.g. descriptors that are fruit-like) may be subtly different in terms of the most important or predictive chemical features.

The human olfactory system discriminates similar smelling chemicals and does so presumably by detecting minor differences in key physicochemical features using an array of odorant receptors. To understand how a machine learning algorithm might achieve such discrimination, we selected two groups of closely correlated perceptual descriptors, fruit-like and soot-like and performed a network analysis as before. As expected, many top physicochemical features were shared among these similar descriptors, and yet separate sub-clusters were present (Figure 5.2C, top and bottom). Representative compounds with descriptors such as “Grape Juice” and “Peach, Fruit” are subtly different from each other, as are ones for “Sooty” and “Tar” (Figure 5.2D). When examining these differences in physicochemical features, it is evident how slight variations in structurally related chemicals could result in distinct perceptual responses. We also observed this in an additional analysis (Figure 5.9B). This suggests that

physicochemical information in machine learning models can address a complex challenge, similar to the biologically relevant discriminatory task.

An analysis of the chemical features selected for all the perceptual models suggested that the 3D structure of a chemical contributed significantly to predictions of odor perception, particularly the 3D-MoRSE (Schoor, Selzer, & Gasteiger, 1996) and GETAWAY (Consonni, Todeschini, & Pavan, 2002) chemical features (DRAGON), which are 3D representations weighted by physicochemical properties that are possibly without precise structural interpretations (Figure 5.9C). Simpler 2D features and functional group counts were less important (Figure 5.9C).

Only a miniscule portion of the odor-chemical space has been evaluated for perceptual information and this in part reflects the low throughput and high cost of human studies. One approach to overcome this is to extend small experimental datasets to large, unexplored chemical spaces. Subsequently, we predicted the 146 Dravnieks (1985) study perceptual descriptors for a ~440,000 chemical library (Boyle, Guda, et al., 2016; Boyle, McNally, Tharadra, & Ray, 2016) (Figure 5.3a, top and bottom). We evaluated ~68 million descriptor-chemical combinations and predicted numerous (hundred to thousands) new chemicals that smell like each descriptor. These chemicals represent a massive expansion (>3000 times) of the previously known chemical space with perceptual descriptors, which is likely to cover a substantial fraction of putative volatile chemicals with odorant properties. Ultimately, the predictions allowed us to create, for the first time, a comprehensive chemical space of all 146 Dravnieks (1985) perceptual descriptors.

Visualizing this massive chemical space in a 2D image is difficult, so we represented only a fraction of the top predictions in the form of a network (Figure 5.3B). We next clustered similar perceptual descriptors, highlighting the frequently occurring chemical features among the top predictions. Though the machine learning models incorporate potentially abstract chemical features, this type of analysis can help visualize structural features that may contribute to a certain percept (Figure 5.4).

5.3. Discussion

In this study, we provided a comprehensive analysis of odor perception prediction from physicochemical features of volatile chemicals, and have supplied important groundwork to understand optimal methods, metrics and approaches in modeling diverse perceptual descriptors. We do so with an additional focus on transparency and interpretability.

Of note, is the finding that most perceptual descriptors are best predicted by chemical features that describe 3D geometries. The value of 3D information was anticipated however when considering structurally similar odorants share many 2D features. To successfully discriminate odorant percepts, machine learning models utilize additional physicochemical properties, particularly 3D shape. In datasets with an arbitrarily large number of perceptual descriptors, the important chemical features could be redundant and cross-descriptor predictions overlap. However, we found that, although important chemical features overlap, the set of descriptors for a percept and the models themselves were indeed largely distinct. This would be consistent with evidence that

perceptual descriptors appear highly structured and are not arbitrary (Gutiérrez et al., 2018).

While caution is required in interpreting results from the Dravnieks (1985) or Keller (2016) datasets, which are small samples by typical machine learning standards, our validations and control analyses establish that they are nevertheless rich sources of information for uncovering structure-odor percept relationships. The generalizability of physicochemical feature-based models across the differing sample demographics and the mostly distinct odor panels is further evidence. To that end, we have ultimately outlined a simple pipeline that can be applied to facilitate data-driven theories about the human olfactory perceptual space and its physicochemical origins on a considerably larger scale.

This study differs in notable ways to others including placing the results in the context of diverse odor perception prediction efforts. We expanded cross-study analyses, where training and testing are performed on different psychophysics data sources. These results suggested that models trained on the Dravnieks data could be successfully adapted to predict the Keller study and a very different, non-experimental dataset in GoodScents. Although the size of the training set directly impacts success, and models trained on more data are always expected to perform better, the results are quite good relative to the size of the Dravnieks training data. Importantly, in previous modeling efforts focus has been placed on open source chemical feature representations. These include e-Dragon, a free web interface to an early version of Dragon software and moldred/RDKit. Here, we used proprietary geometry optimization tools such as OMEGA alongside the full version of Dragon. When we compared different feature representations, it is evident that there are

performance gains and losses depending on the perceptual descriptor. Subsequently, there is no consensus approach, but the tools used in this study appear to improve predictions, particularly within the Keller study; that is, when training is followed by prediction of 69 test chemicals from the same study.

The chemical features we report for the different Dravnieks perceptual descriptors are therefore a valuable resource and will likely support odor coding research and assist researchers in identifying new chemicals that smell a specific way. Predicted compounds from the large computational screen are a rich source of information about our potential olfactory chemical space. While this includes comparing predictive modeling efforts to define successes, failures and future directions for the field, differences in the methodology and chemical sets have thus far limited comparisons. We set the foundation here for such comparisons, analyzing several different perceptual datasets and evaluating various modeling efforts with multiple metrics. These comparisons are nevertheless broad in nature. But they help provide interpretation about perceptual descriptors and their predictability across considerably different modeling studies.

By applying machine learning alongside traditional chemoinformatic tools, we suggest it is now possible to extrapolate from the quality perceptual study data to large chemical spaces. These spaces can play an important role in translating the complex chemical features in machine learning models into visible, more interpretable patterns. We therefore anticipate that this study will provide a powerful approach and resource for the discovery of new flavors and fragrances.

5.4. Figures

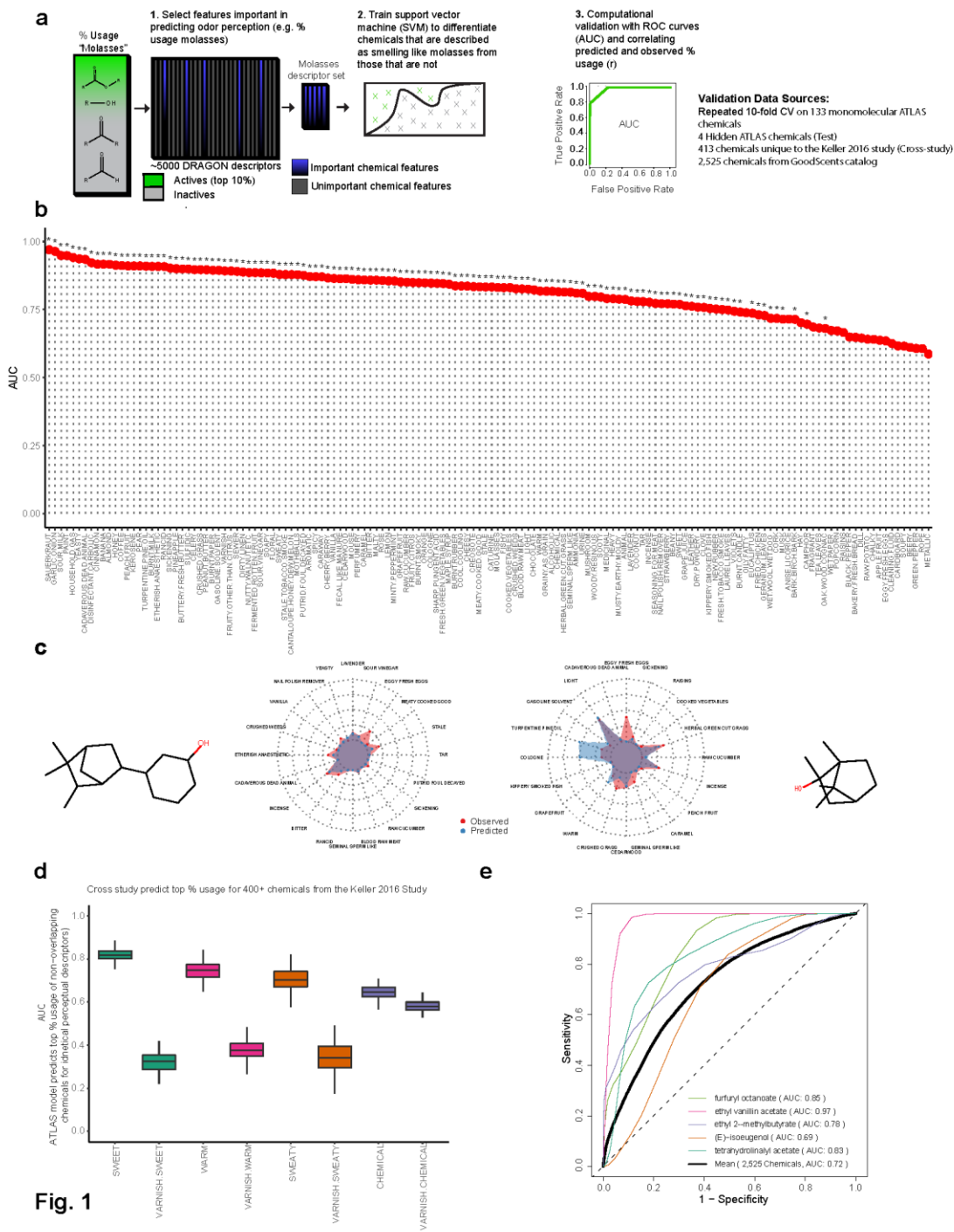


Fig. 1

Figure 5.1. Predicting perceptual descriptors from physicochemical features with machine learning. **A)** Pipeline for predicting Dravnieks (1985) ratings (% Usage) for perceptual descriptors, an example is provided for the descriptor, “molasses.” Important chemical features are detected that predict “Molasses.” A support vector machine (SVM) is fit; predictions are assessed by different methods such as the area-under-the curve (AUC) from Receiver Operating Characteristic (ROC) plots. **B)** Chemicals within the top 10% of ratings (% Usage) are labeled as “Active.” The AUC quantifies the relationship between sensitivity to actives (chemicals in the top 10% ratings) vs false positives. Plot bars represent the average AUC from three models trained using different chemical features. The AUC is computed on chemicals excluded from training (30 times, 10-fold CV repeated 3 times). Significance (*) from one-sided t-test, comparing the AUC to an identical model trained on shuffled “Active” labels. The number of “Active” labels remains unchanged. Significance is $p \leq .05$ after adjusting for false discovery rate. **C)** Predicted vs observed % usage for select test chemicals. For clarity, only a selection of perceptual descriptors is shown. **D)** Dravnieks (1985) trained models of “Sweet”, “Warm”, “Sweaty” and “Chemical” predict ratings for these same descriptors from a study of public volunteers for 413 test chemicals (Keller & Vosshall, 2016). Cutoffs to convert the public volunteer data into actives are from the Dravnieks (1985) study (top 10% usage). Significance is from one-sided t-test, comparing the perceptual descriptor models with a non-identical but top-performing Dravnieks (1985) model, “Varnish” over 100 bootstrap samples. Public volunteer data is averaged over dilution. **E)** Average prediction performance (AUC) when assigning 1-146 Dravnieks (1985) perceptual descriptor labels to 2,525 test chemicals with known labels in GoodScents database. CV: Cross-validation.

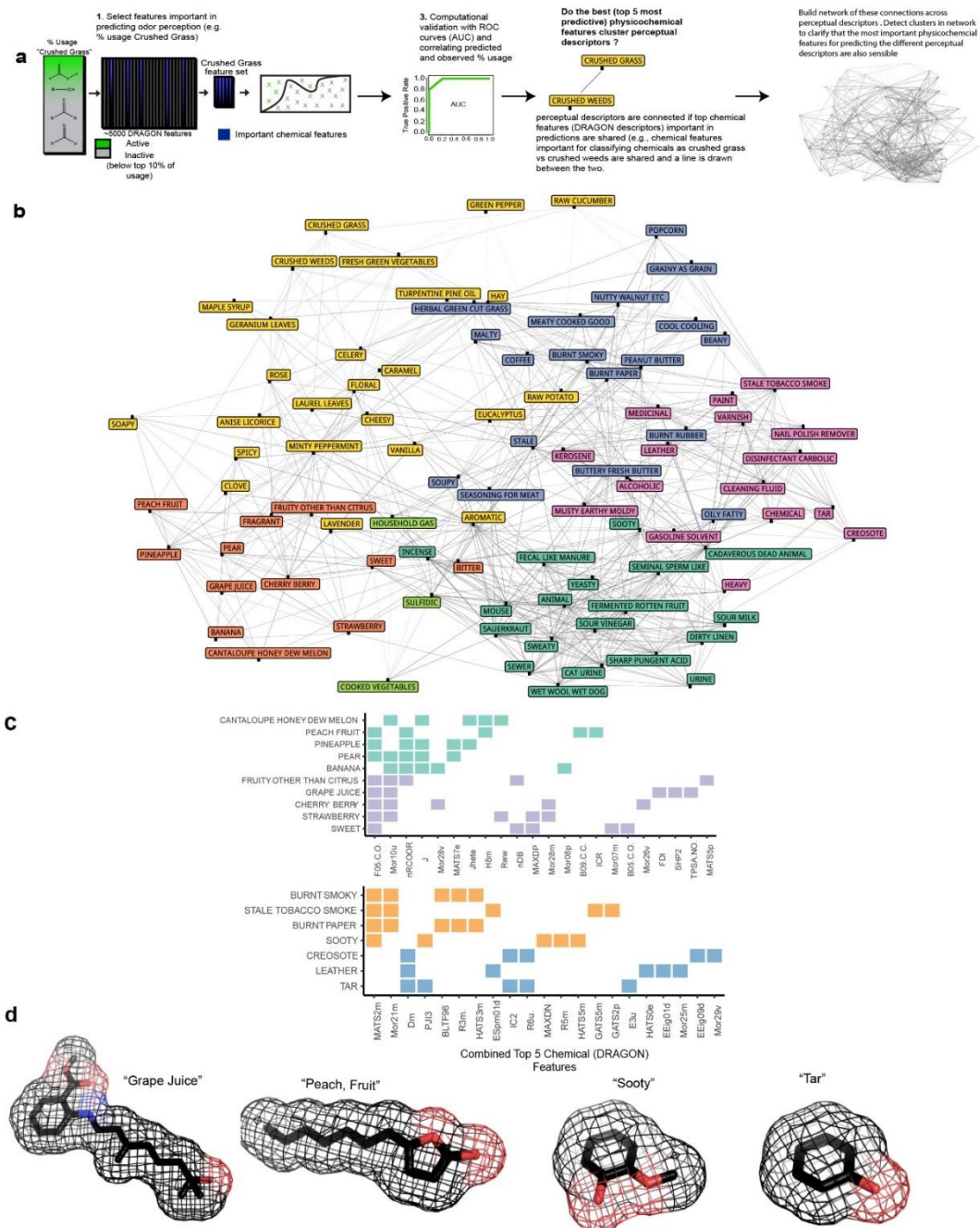


Figure 5.2. Building perceptual descriptor networks from few physicochemical features. **A)** Pipeline summarizing methods for selecting the most important chemical features for predictions of Dravnieks (1985) perceptual descriptors, followed by the construction of networks that help visualize relationships among these descriptors when considering physicochemical information alone. **B)** Assembled network from the top 5 chemical features per descriptor. Descriptors with shared top 5 chemical features are connected in the network. Similar perceptual descriptors are color-coded based on the Louvain algorithm. **C)** Two sets of correlated descriptors are analyzed based on the chemical features that are important (among the top 5) for predicting them. **Top**, matrix 1: “fruity” descriptors. **Bottom**, matrix 2: “sooty” descriptors. Louvain clustering (square color) shows the similar descriptors are separable into 2 sub-groups. Filled-in squares, regardless of color, represent the importance of the labeled chemical feature. **D)** Exemplar chemicals from the computationally inferred clusters.

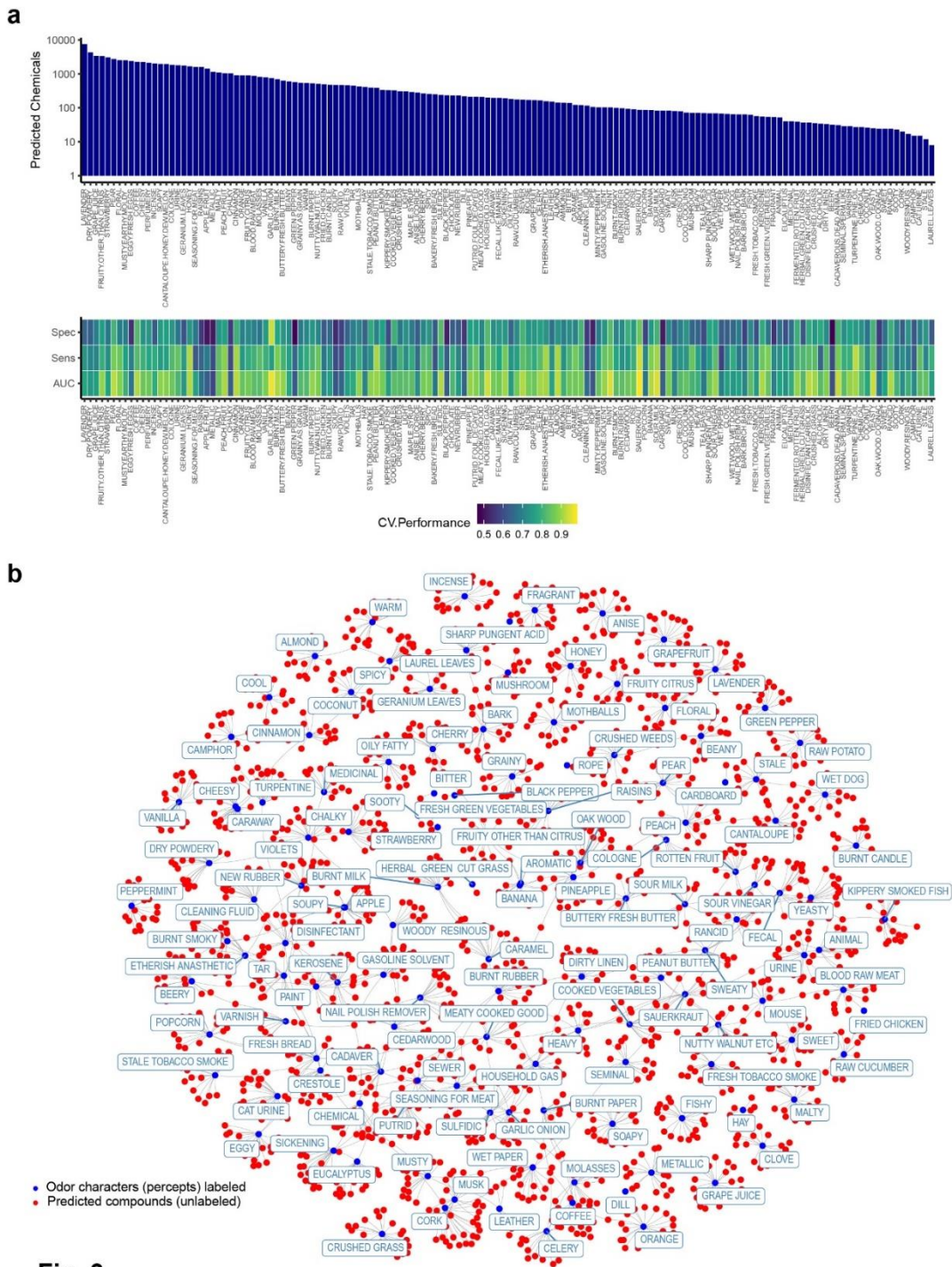


Fig. 3

Figure 5.3. Predicting and mining large commercially available chemical spaces. A) The machine learning models are used to predict perceptual descriptors from ~440,000 compounds. **Top**, predicted chemical counts are based on optimal thresholds from the ROC curves and structural similarity (atom pair similarity > .25) to training actives. An optimal threshold is the point on the curve that minimizes false positives and maximizes true positives. **Bottom**, detailed validation for the models ordered with respect to the number of predicted chemicals. **B)** A 2D representation of predictions for 15 hits for each perceptual descriptor (or all chemicals that exceed the % usage threshold for actives), with edges connecting compounds that are predicted for multiple descriptors. The newly predicted chemicals are indicated as unnamed red dots, and each descriptor as blue dots and labeled in rectangles. Predictions are from the support vector machine (SVM) algorithm with a radial basis function (RBF) kernel. See Methods for additional information.

a





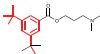










Image	Descriptor	ID Cluster	Image	Descriptor	ID Cluster
	CEDARWOOD	523689 1		PINEAPPLE	6167463 3
	EUCALYPTUS	885451 1		ANIMAL	530532 4
	VIOLETS	24356211		MOUSE	1970417 4
	CHEMICAL	532574 2		RANCID	4927318 4
	KEROSENE	484298 2		URINE	8296098 4
	PAINT	496365 2			
	VARNISH	61944272			
	AROMATIC	11218833			
	BANANA	505325 3			
	PEAR	973376 3			

Fig. 4

Figure 5.4. Enriched chemical features among predictions. A) Top predicted chemicals in eMolecules from the Figure 5.3 network are clustered and analyzed for common structural features (substructures or cores). These are highlighted (red) in images of representative chemicals from the predictions. The ID is the eMolecules identifier. Simple structural features are common among predicted chemicals, enabling basic comparisons between different perceptual descriptors based on chemical structure. Accordingly, this is an example of how a large network of predictions can offer additional insight. See Methods for details on the maximum common substructure algorithm for identifying the enriched features.

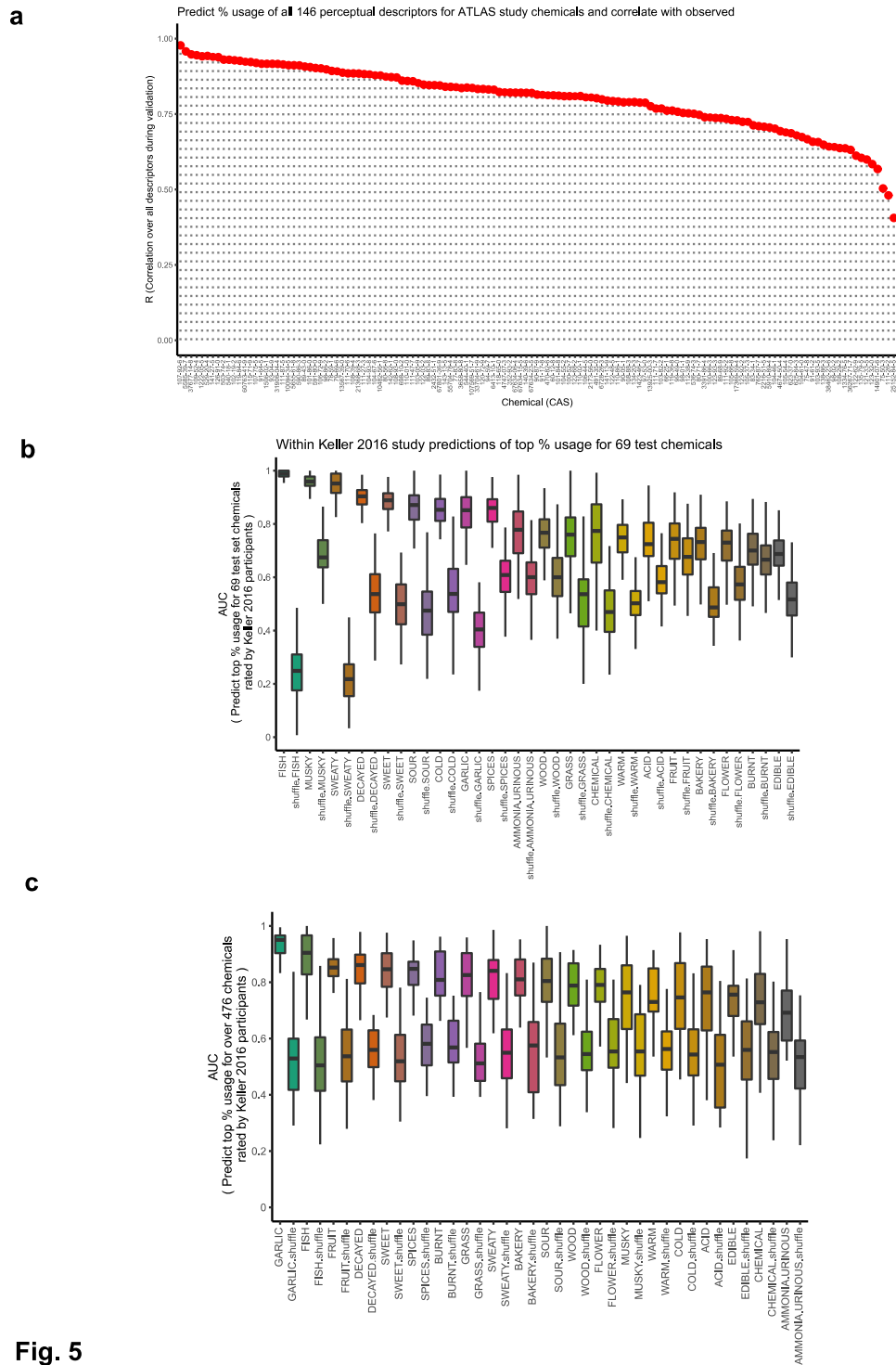
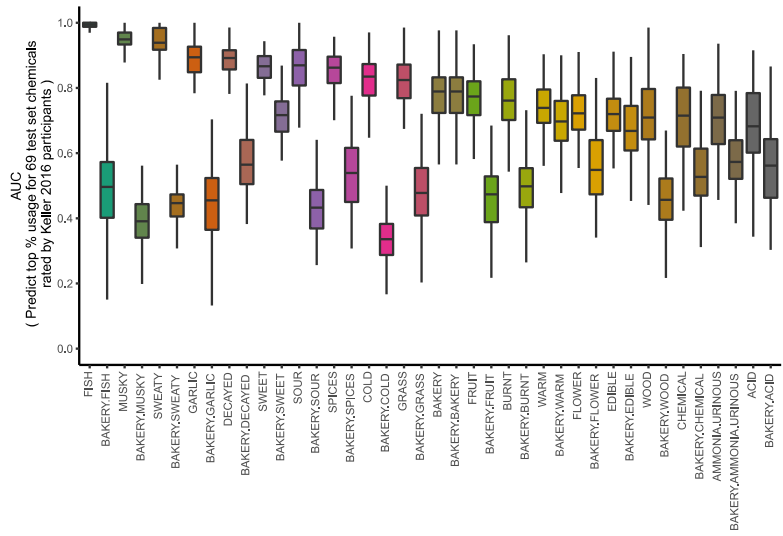


Fig. 5

Figure 5.5. Comprehensive evaluation of machine learning models with different metrics and data. **A)** Average correlation (R) between the predicted and observed % usage for the full set of perceptual descriptors over cross validation. Dravnieks (1985) study chemicals (x-axis) are abbreviated as the CAS identifier. **B)** Evaluation of chemical (DRAGON) feature models trained on the Keller 2016 study data. Models classify 69 test chemicals (used in the DREAM analysis) as smelling like a given descriptor (top 10% Usage). These chemicals were excluded from training and chemical feature selection. The area under the ROC curve (AUC) compares predictions to the data observed from the general public volunteers in that study. Chance performance is defined by training models identically but on mislabeled chemicals (shuffle). Error is the standard deviation over 100 bootstrap samples. **C)** A similar analysis is done using an alternative validation method where all 476 chemicals in the Keller 2016 study are repeatedly divided into training and testing chemical sets (10-fold cross-validation, repeated 3 times). This covers more diversity than the 69 test chemicals. Chemical features for these models were selected using a subset of the data to minimize biased validation. The predictions are aggregated from the support vector machine (SVM) and regularized random forest algorithms. Additional information on AUC calculation and its interpretation are in Methods. Chemical feature selection methods and biases that affect validation are also defined in Methods.

a



b

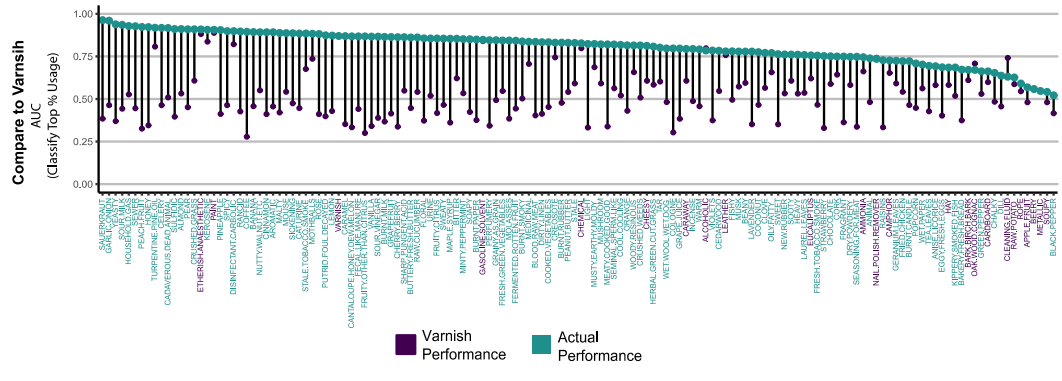


Fig. 6

Figure 5.6. An arbitrary model for predicting an perceptual descriptor fails to outperform actual model. **A)** Area under the ROC curve (AUC) for classifying the top 10% of usage on 69 test chemicals with chemical (DRAGON) features across perceptual descriptors from the 55 Keller 2016 study participants, averaging over dilution. The 69 test chemicals are as reported in the DREAM analysis (Keller et al., 2017). AUCs computed from aggregated scores of a RBF SVM and a regularized random forest. Performance of each perceptual descriptor model is plotted alongside performance if replacing the predictions with the “Bakery” model. Chemical features selected and models fit on 407 training chemicals. Error (standard deviation) is over 100 bootstrap samples of the 69 test chemicals. **B)** Classification (AUC) of top 10% of usage for the 146 Dravnieks (1985) perceptual descriptors descriptor models (teal dots) compared to predictions using a top performing “Varnish” (purple dots) model. Perceptual descriptors colored in purple failed to outperform “Varnish,” $p > .05$, adjusting for FDR (Benjamini-Hochberg). Plotted AUCs reflect the average of 3 RBF SVM models using different chemical features from a pool of ~ 70 over 30 cross validation folds (10- fold CV repeated 3 times) (RBF: Radial Basis Function; SVM: Support Vector Machine; FDR: False Discovery Rate).

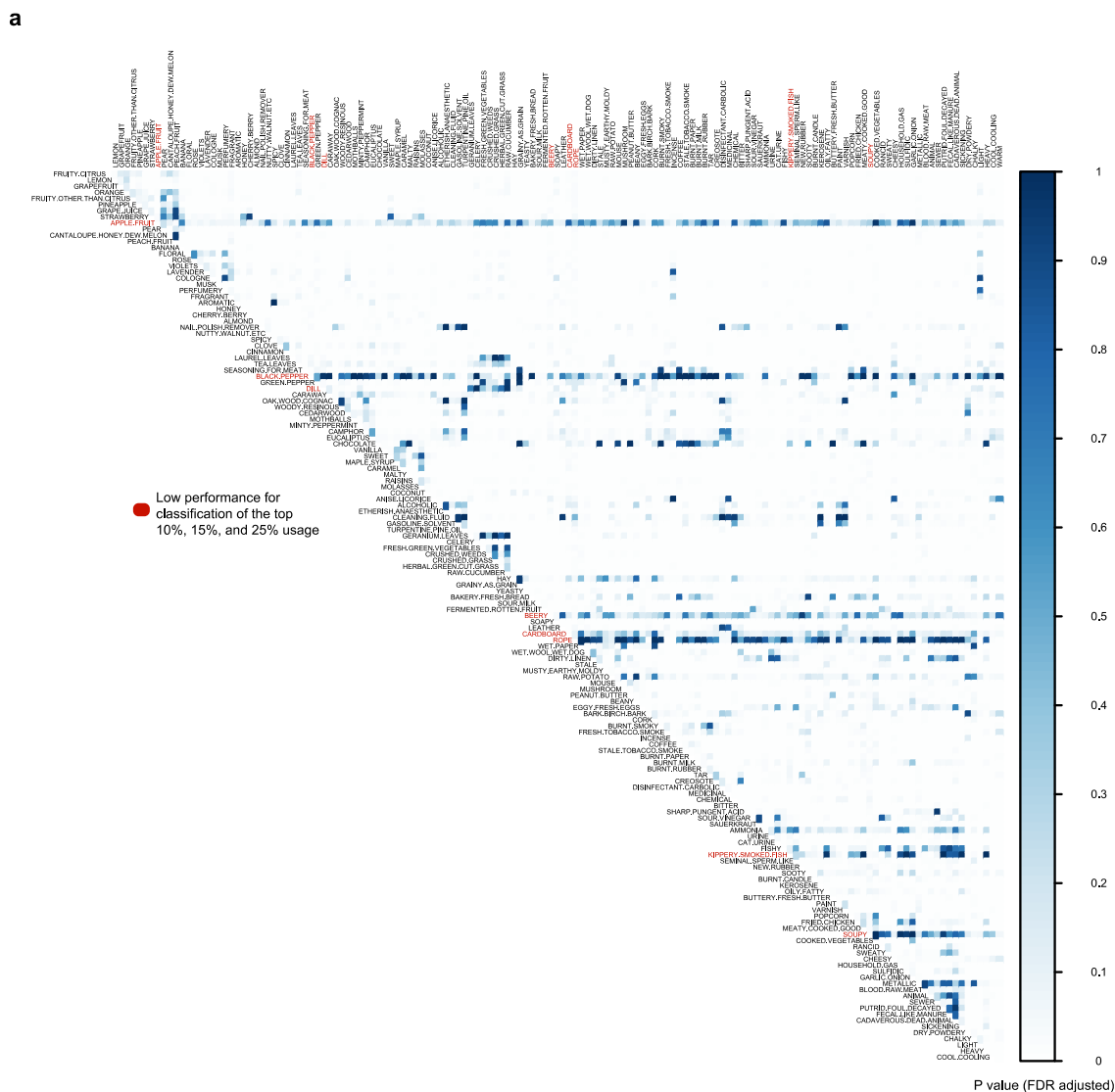
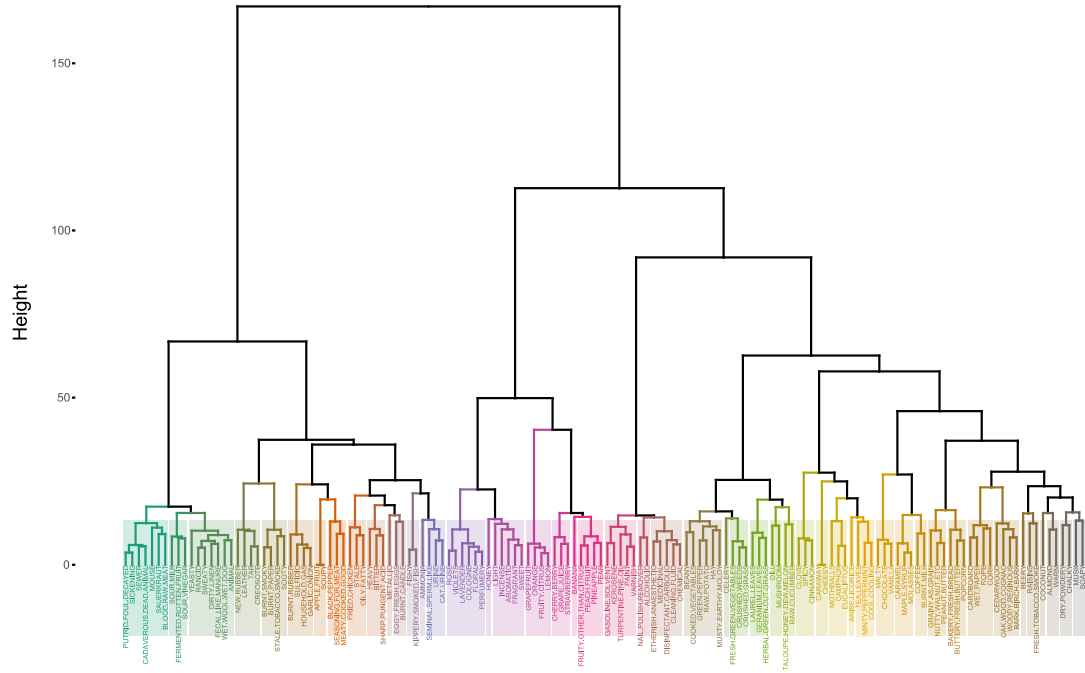


Figure 5.7. The actual predictive model for a perceptual descriptor is generally most accurate. A) Dravnieks (1985) study prediction performance over the cross validation where the percent usage of each perceptual descriptor is predicted by the models for the other descriptors. The color is the p value adjusted for FDR (T-test). Descriptor labels are colored (red) to distinguish the models that are of a lower quality rather than perceptual redundancy. These perceptual descriptors may fail for many reasons but notably most are not well represented among Dravnieks (1985) study chemicals (e.g. lack exemplars for classification training).

a

Cluster by Percent Usage in ATLAS Study



b Cluster by Top 30 Chemical Features for Perceptual Descriptors in ATLAS Study

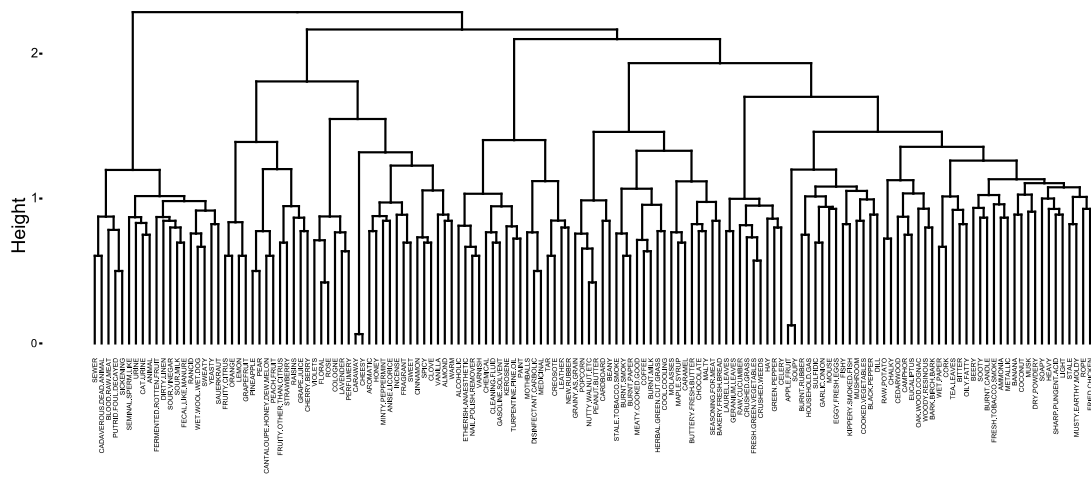


Figure 5.8. Chemical features cluster odor perceptual qualities similarly to the ground truth human rating. **A)** Hierarchical clustering of the Dravnieks (1985) study data by % usage. The cluster (colors) number is determined by the gap statistic over bootstrap samples. The distance is Euclidean. **B)** Hierarchical clustering is instead performed based on chemical feature sets appearing in the machine learning models. The distance is 1-Jaccard index, where the Jaccard index here indicates the similarity of binary strings (1,0) specifying if a chemical feature is or is not in the perceptual descriptor model.

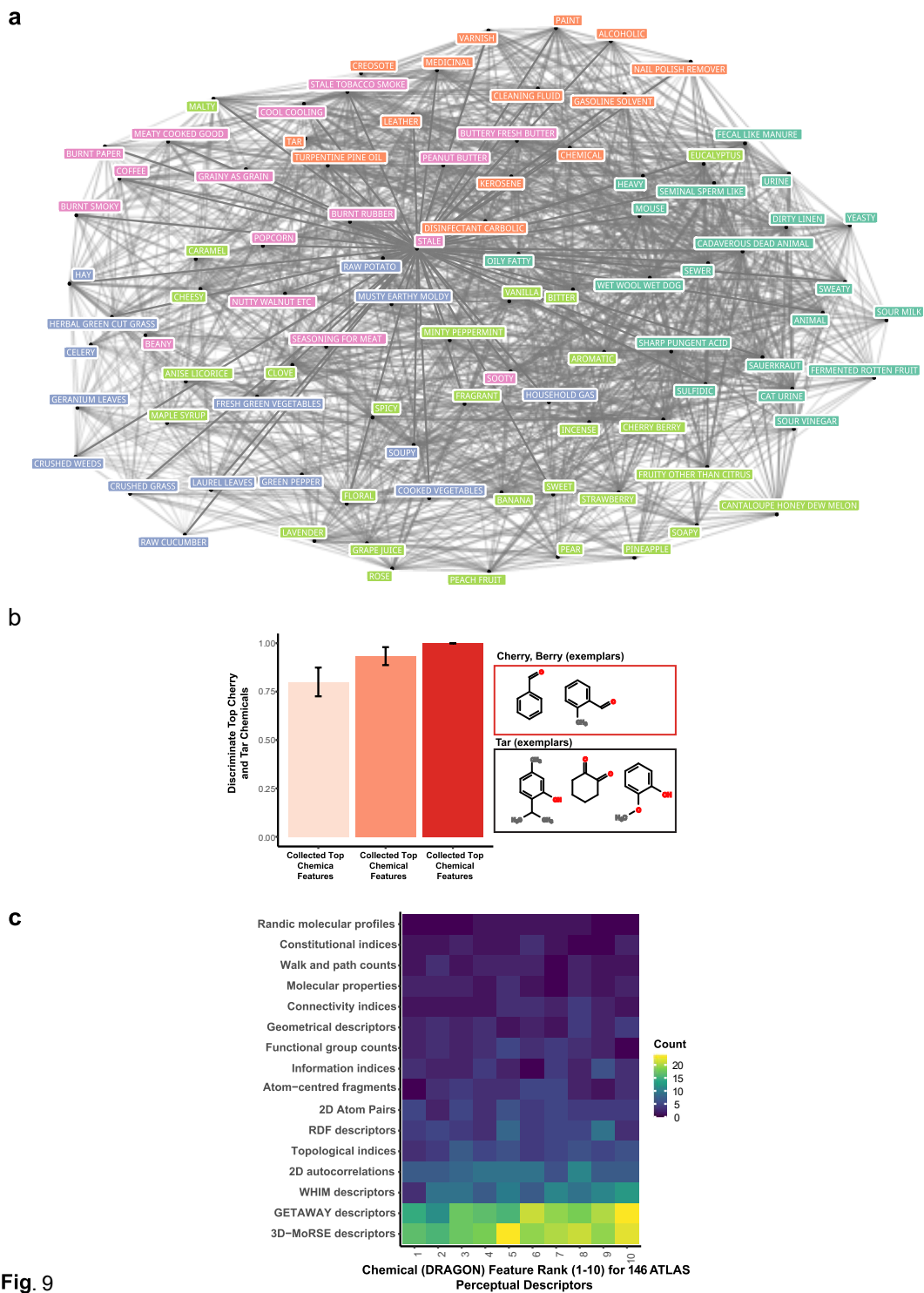


Fig. 9

Figure 5.9. Few predictive chemical features can in part differentiate diverse odor perceptual qualities. **A)** The 10 most important chemical (DRAGON) features for accurate predictions of perception (% usage) are used to build a network representation that shows relationships among the perceptual descriptors in terms of their prospective physicochemical similarity. Connectivity in the network signifies shared chemical features among 93 distinct perceptual descriptors and is used to infer clusters of similar perceptual descriptors according to the Louvain algorithm. The large number of features leads to a densely connected network but clusters detected. **B)** *Left*, discriminating top chemicals that smell like “cherry” versus “tar,” according to Dravnieks (1985) study respondents. The discrimination success is quantified by the average AUC across 30 cross validation folds (10-fold CV repeated 3 times) for models comprised of 1, 2, and 3 principal components (PC 1-3) that optimally retain information in the combined top 10 chemical (DRAGON) features (20 total). Error bars reflect the standard error. Note the 3 component model provides perfect classification. *Right*, exemplar chemicals for "cherry (berry)" and “tar" that are structurally similar but with subtly distinct chemical features. **c)** Counts of the chemical (DRAGON) features selected in bins from the top 1-10 (x-axis) for 146 perceptual descriptors with respect to the broad categories (y-axis) the features fall into.

5.5. Tables

Table 5.1

Perceptual Descriptor	Actives (Top 25%)	Actives (Top 15%)	Actives (Top 10%)	AUC (Top 10%)	AUC (Top 15%)	AUC (Top 25%)
FRUITY.CITRUS	34	20	14	0.850	0.852	0.825
LEMON	34	20	14	0.856	0.849	0.793
GRAPEFRUIT	36	20	14	0.851	0.847	0.858
ORANGE	35	20	14	0.849	0.848	0.849
FRUITY.OTHER.THAN.CITRUS	34	20	14	0.892	0.887	0.855
PINEAPPLE	35	20	14	0.901	0.888	0.823
GRAPE.JUICE	34	21	14	0.764	0.761	0.743
STRAWBERRY	34	20	14	0.771	0.802	0.819
APPLE.FRUIT	37	21	14	0.637	0.661	0.729
PEAR	34	20	14	0.910	0.828	0.808
CANTALOUPE.HONEY.DEW.MELON	34	21	14	0.878	0.898	0.850
PEACH.FRUIT	34	20	14	0.911	0.933	0.849
BANANA	34	20	14	0.917	0.812	0.747
FLORAL	34	20	14	0.884	0.830	0.834
ROSE	34	20	14	0.857	0.823	0.848
VIOLETS	34	20	14	0.744	0.793	0.787
LAVENDER	34	20	14	0.782	0.834	0.787
COLOGNE	34	20	14	0.848	0.869	0.850
MUSK	34	20	14	0.715	0.721	0.708
PERFUMERY	34	20	14	0.860	0.843	0.777
FRAGRANT	34	20	14	0.686	0.742	0.790
AROMATIC	34	20	14	0.872	0.834	0.810
HONEY	35	20	15	0.913	0.810	0.770
CHERRY.BERRY	34	20	14	0.866	0.838	0.770
ALMOND	34	20	14	0.916	0.876	0.866
NAIL.POLISH.REMOVER	35	20	14	0.772	0.788	0.699
NUTTY.WALNUT.ETC	34	20	14	0.886	0.858	0.825
SPICY	34	20	14	0.871	0.867	0.843
CLOVE	34	20	14	0.758	0.802	0.782
CINNAMON	34	20	14	0.917	0.855	0.837
LAUREL.LEAVES	38	20	16	0.748	0.717	0.783
TEA.LEAVES	34	20	15	0.682	0.776	0.707
SEASONING.FOR.MEAT	34	20	14	0.773	0.722	0.736
BLACK.PEPPER	34	20	14	0.649	0.630	0.732
GREEN.PEPPER	34	20	14	0.607	0.658	0.766
DILL	34	21	14	0.645	0.585	0.605
CARAWAY	36	20	14	0.871	0.710	0.713
OAK.WOOD.COGNAC	34	20	14	0.680	0.759	0.792

WOODY.RESINOUS	36	20	14	0.799	0.824	0.828
CEDARWOOD	34	20	14	0.863	0.807	0.809
MOTHBALLS	34	20	14	0.878	0.893	0.813
MINTY.PEPPERMINT	34	20	14	0.855	0.880	0.800
CAMPHOR	34	20	14	0.697	0.839	0.828
EUCALIPTUS	34	21	14	0.737	0.785	0.753
CHOCOLATE	34	20	14	0.822	0.823	0.721
VANILLA	35	20	14	0.864	0.929	0.840
SWEET	34	20	14	0.769	0.803	0.797
MAPLE.SYRUP	35	20	14	0.843	0.778	0.646
CARAMEL	34	20	14	0.859	0.825	0.819
MALTY	34	21	15	0.858	0.811	0.782
RAISINS	37	21	14	0.666	0.820	0.806
MOLASSES	34	22	14	0.831	0.784	0.708
COCONUT	34	20	14	0.780	0.794	0.777
ANISE.LICORICE	34	21	14	0.715	0.724	0.703
ALCOHOLIC	34	20	14	0.817	0.821	0.704
ETHERISH.ANAESTHETIC	34	20	14	0.908	0.801	0.783
CLEANING.FLUID	35	20	14	0.626	0.800	0.807
GASOLINE.SOLVENT	34	20	14	0.894	0.865	0.776
TURPENTINE.PINE.OIL	34	20	14	0.909	0.864	0.827
GERANIUM.LEAVES	36	20	14	0.728	0.766	0.815
CELERY	35	21	14	0.897	0.788	0.742
FRESH.GREEN.VEGETABLES	34	20	14	0.846	0.867	0.821
CRUSHED.WEEDS	35	20	14	0.827	0.738	0.767
CRUSHED.GRASS	34	20	14	0.897	0.866	0.788
HERBAL.GREEN.CUT.GRASS	34	20	14	0.814	0.832	0.822
RAW.CUCUMBER	34	20	14	0.851	0.819	0.748
HAY	34	20	14	0.702	0.760	0.735
GRAINY.AS.GRAIN	34	20	14	0.818	0.741	0.729
YEASTY	34	20	14	0.937	0.877	0.827
BAKERY.FRESH.BREAD	35	21	14	0.648	0.563	0.622
SOUR.MILK	34	20	14	0.949	0.848	0.741
FERMENTED.ROTTEN.FRUIT	34	20	14	0.886	0.889	0.791
BEERY	34	20	14	0.611	0.675	0.707
SOAPY	34	20	14	0.885	0.871	0.836
LEATHER	34	20	14	0.762	0.788	0.783
CARDBOARD	34	20	14	0.617	0.697	0.726
ROPE	34	20	14	0.607	0.634	0.582
WET.PAPER	34	20	14	0.674	0.729	0.728
WET.WOOL.WET.DOG	34	20	14	0.718	0.759	0.762
DIRTY.LINEN	34	20	14	0.888	0.885	0.864
STALE	34	20	14	0.833	0.868	0.810

MUSTY.EARTHY.MOLDY	34	21	14	0.788	0.814	0.754
RAW.POTATO	35	20	14	0.641	0.786	0.718
MOUSE	34	20	14	0.861	0.887	0.813
MUSHROOM	34	20	14	0.799	0.765	0.746
PEANUT.BUTTER	34	20	14	0.896	0.828	0.825
BEANY	34	21	14	0.770	0.711	0.690
EGGY.FRESH.EGGS	34	20	14	0.635	0.578	0.568
BARK.BIRCH.BARK	34	22	14	0.714	0.722	0.666
CORK	34	20	14	0.718	0.733	0.773
BURNT.SMOKY	34	20	14	0.849	0.898	0.913
FRESH.TOBACCO.SMOKE	34	21	15	0.750	0.738	0.675
INCENSE	34	20	14	0.778	0.817	0.826
COFFEE	34	20	15	0.912	0.882	0.735
STALE.TOBACCO.SMOKE	34	20	14	0.879	0.869	0.872
BURNT.PAPER	34	21	14	0.895	0.858	0.864
BURNT.MILK	35	21	14	0.908	0.745	0.675
BURNT.RUBBER	34	20	14	0.837	0.845	0.891
TAR	34	20	14	0.779	0.795	0.783
CREOSOTE	34	21	14	0.834	0.865	0.773
DISINFECTANT.CARBOLIC	34	20	14	0.922	0.737	0.859
MEDICINAL	34	21	14	0.791	0.677	0.771
CHEMICAL	34	20	14	0.815	0.841	0.814
BITTER	34	20	15	0.858	0.889	0.780
SHARP.PUNGENT.ACID	34	20	14	0.847	0.891	0.815
SOUR.VINEGAR	34	20	14	0.885	0.866	0.802
SAUERKRAUT	34	20	14	0.970	0.904	0.788
AMMONIA	34	21	14	0.811	0.848	0.792
URINE	36	20	14	0.810	0.810	0.797
CAT.URINE	35	20	14	0.833	0.862	0.748
FISHY	34	21	14	0.831	0.791	0.753
KIPPERY.SMOKED.FISH	34	20	14	0.754	0.701	0.660
SEMINAL.SPERM.LIKE	35	20	14	0.814	0.850	0.804
NEW.RUBBER	34	20	14	0.752	0.799	0.827
SOOTY	34	20	15	0.797	0.756	0.769
BURNT.CANDLE	34	20	14	0.741	0.800	0.771
KEROSENE	34	20	14	0.910	0.857	0.802
OILY.FATTY	34	20	14	0.739	0.821	0.883
BUTTERY.FRESH.BUTTER	34	21	14	0.899	0.872	0.786
PAINT	34	20	14	0.949	0.864	0.774
VARNISH	34	20	14	0.892	0.840	0.775
POPCORN	34	20	15	0.672	0.666	0.659
FRIED.CHICKEN	0	21	17	0.731	0.688	N/A
MEATY.COOKED.GOOD	37	20	14	0.833	0.830	0.801

SOUPY	37	21	14	0.616	0.677	0.752
COOKED.VEGETABLES	34	20	15	0.831	0.905	0.760
RANCID	34	20	14	0.908	0.888	0.816
SWEATY	34	20	14	0.879	0.848	0.764
CHEESY	36	20	14	0.836	0.677	0.716
HOUSEHOLD.GAS	34	20	14	0.941	0.930	0.890
SULFIDIC	34	20	14	0.899	0.923	0.821
GARLIC.ONION	34	20	14	0.964	0.810	0.823
METALLIC	35	21	14	0.587	0.675	0.713
BLOOD.RAW.MEAT	34	20	14	0.826	0.759	0.813
ANIMAL	34	20	14	0.787	0.775	0.731
SEWER	35	20	14	0.891	0.897	0.825
PUTRID.FOUL.DECAYED	35	20	14	0.876	0.903	0.831
FECAL.LIKE.MANURE	34	20	14	0.863	0.868	0.772
CADAVEROUS.DEAD.ANIMAL	37	20	14	0.934	0.893	0.817
SICKENING	34	20	14	0.902	0.917	0.847
DRY.POWDERY	34	20	14	0.759	0.840	0.812
CHALKY	35	20	14	0.641	0.739	0.772
LIGHT	34	20	14	0.825	0.770	0.753
HEAVY	34	20	14	0.790	0.842	0.806
COOL.COOLING	34	20	14	0.837	0.804	0.664
WARM	34	20	14	0.818	0.769	0.771

Table 5.1. Figure 1 performance using different classification cutoffs. The average AUC is shown for varying classification cutoffs. The % usage is transformed into active and inactive labels according to the top end of the % usage distribution (Top 10, 15, and 25), which changes the number of active and inactive chemicals.

Table 5.2

Perceptual Descriptor	AUC
SWEET	0.8162655
VARNISH.SWEET	0.3243550
SHUFFLE.SWEET	0.5812922
SWEATY	0.7071707
VARNISH.SWEATY	0.3413408
SHUFFLE.SWEATY	0.5426659
WARM	0.7472995
VARNISH.WARM	0.3774729
SHUFFLE.WARM	0.5421157
CHEMICAL	0.6452747
VARNISH.CHEMICAL	0.5801918
SHUFFLE.CHEMICAL	0.4234795

Table 5.2. Cross-Study classification performance. Dravnieks (1985) models predict the same perceptual descriptor in the Keller 2016 study for 413 chemicals unique to the study. The area under the curve (AUC) is averaged over 100 bootstrap samples. The perceptual descriptor is the model used for predictions. Each descriptor is appended with “Shuffle” or “Varnish,” showing the performance when the Dravnieks (1985) study model is trained on shuffled labels for exemplar chemicals or, alternatively, the Dravnieks (1985) “Varnish” model.

5.6. Methods

5.6.1. Psychophysical data

5.6.1.1 Keller (2016) Study

We used data from 55 general public volunteers (Keller & Vosshall, 2016) for external validation. Due to limited diversity in the selection of odor descriptors supplied by naïve volunteers and evidence indicating experience with odor language improves the quality of perceptual data (Dubois & Rouby, 2002; Lawless, 1984; Olofsson & Gottfried, 2015), we primarily considered a sample of industry professionals as reported in the atlas of odor character profiles (Dravnieks (1985)) (Dravnieks, 1985). Notably, the semantic descriptors (odor characters or perceptual descriptors) were sparsely used in some cases among the general public volunteers, suggesting that averaged ratings for a given descriptor (odor character) might represent a very small proportion of the respondents. This becomes particularly important for generating predictive models since missing data points (e.g. chemicals or odorants that are not rated by some participants) must be dealt with such as by averaging ratings for the nearest neighboring (k) odorants or filling-in with the median/mean rating across all odorants. While these approaches are valid in predictive modeling, they are a significant modification of the respondent data; the failure to provide a rating is a potentially important source of information. We maintained, as a result, the 0-100 scale for the general public volunteer data but converted ratings to a % usage metric instead. Dilution was not considered, averaging % usage over the different dilutions. In preliminary analyses there was however some evidence that models might benefit from training on data from a single dilution. Similarly, a small number of

replicates that were performed in this study were not included in the final training and testing datasets.

Although with the % usage each odorant is assigned numeric values more naturally, this modification was also in line with the Dravnieks (1985) study data. The % usage therefore provided a means to compare two sources that to a first approximation appear very different. Dravnieks (1985) also reports a percent applicability metric. The percent applicability is the sum of the ratings for a chemical or odorant over all participants divided by the maximum possible sum. This was not used for our cross-study comparisons as ratings from an experienced participant panel might scale differently and the sample size between the two studies is very different. Because cross-study comparisons are not well defined, we opted for the simplest possible metric, the % usage.

5.6.1.2. Atlas of odor character profiles, Dravnieks (1985)

Dravnieks (1985) summarizes odor profiles for 180 odorants, replicates and mixtures, with the latter not being used for predictions, from 507 industry professionals in total across 12 organizations. Each chemical was rated by between 120 and 140 participants. The participants scored a set of replicates, which were used to provide an index of discriminability for the data as the inverse of the squared correlation coefficient between replicates (RV). For this study, $RV = .11$. The scoring metric was on the range of 1-5 with 1 being slightly and 5 being extremely relevant. Raw scores were subsequently processed into two numeric values summarizing the participants' responses. We only focused on the % usage; the fraction of participants providing any response, 1-5 because

it is the simplest metric to interpret and relate to other studies. The perceptual descriptor (or character) set available for the Dravnieks (1985) study was extensive but empirically driven. Recommendations from the ASTM (American Society for Testing and Materials) sensory evaluation committee winnowed an initial set of 800 possible odor characters (perceptual descriptors) for sensory analyses down to 160. Prompted by additional research, this figure was later revised to 146 relevant perceptual descriptors, a final set that addressed concerns in which clear perceptual differences could result in identical descriptor usage from study participants. This final set of 146 perceptual descriptors and the percent usage was subsequently prepared for machine learning analyses.

5.6.1.3. GoodScents Test Data

GoodScents is a database of 2000+ chemicals, containing basic physicochemical information as well as perceptual descriptor labels, if available, from published reference materials. Since it is not possible to predict a descriptor for which there is no Dravnieks (1985) equivalent, we had to define exclusionary criteria to properly evaluate the predictions. This included in addition to removing non-unique chemicals those without descriptor labels matching or similar to Dravnieks (1985), leaving 2,525 chemicals for test set validation. Examples of similar descriptors in GoodScents include “weedy” and “nutty,” which correspond with “crushed weeds,” and “walnut” and “peanut butter in Dravnieks (1985), respectively.” The 146 Dravnieks (1985) descriptor models assigned a probability score. ROC curves were subsequently computed using the observed descriptor labels for each of the 2,525 chemicals. A chemical described simply as

“nutty,” for example, is expected to have high probabilities for “peanut butter” and “walnut” but not for “orange” and “chemical.” Cases where descriptors were correlated in Dravnieks (1985) ($>.85$) were also defined as a set to avoid overly penalizing the assignment of redundant descriptors to new chemicals. We identified earlier that models with this level of correlation are often interchangeable, with only a non-significant reduction in prediction performance. Namely, machine learning assignment of “Chemical” to an odorant described as “Varnish” was not incorrect given the data. The ROC curve assesses that high probabilities are correctly assigned to the observed descriptors. When the machine learning models predict descriptors that are unlike those observed, the area under the ROC curve decreases. An independent t-test comparison was made between actual AUCs and those using random probability scores.

5.6.2. Selecting optimally predictive chemical features

5.6.2.1. Optimizing chemical structures

Chemical features were computed with DRAGON 6 for Dravnieks (1985). Chemical structures were optimized and 3D coordinates computed with OMEGA. Molecular or chemical features were pre-computed and made publicly available for the DREAM study and these data files were used as is for analysis of the 55 public volunteers reported in the Keller 2016 study.

5.6.3. Chemical feature ranking and importance

5.6.3.1. Cross-validated recursive feature elimination (CV-RFE)

Recursive feature elimination iteratively selects subsets of features to identify optimal sets. The algorithm is a “wrapper” and therefore relies on an additional algorithm to supply predictions and quantify importance. Often this is a decision tree such as random forest, which was used here, since the algorithm computes feature importance internally. This distinction between internal and external simply means that while any arbitrary algorithm can supply the prediction error—here, the error in predicting the % usage value—many lack a well-defined method for quantifying feature importance. Feature importance and ranking must, in these instances, be supplied externally such as by non-linear regression models for each predictor and outcome compared to a constant.

Including cross-validation with the recursive feature elimination (RFE) partitions the training data into multiple folds. This step avoids biasing performance estimates but results in lists of top predictors over the cross-validation folds such that importance of a predictor is based on a selection rate.

5.6.3.2. Random Forest

Random forest is an extension of basic decision trees that overcomes the often-poor generalizability of these models by aggregating the predictions from multiple trees trained on bootstrap samples and different predictor sets, effectively limiting redundancy between trees. Rows that are excluded as part of bootstrapping process are used to estimate prediction performance on new data. This also provides a method for assigning importance to features through randomization; the % increase in prediction error after

randomizing a feature is accordingly the ranking metric that was used for tabulating chemical feature importance.

5.6.3.3. Selection Bias

Selecting features or predictors on the same dataset used for cross validation results in models that have already “seen” possible partitions of the data and therefore performance metrics will be biased. Selection bias (Ambroise & McLachlan, 2002) was addressed by bootstrapping and cross validation, which ensure some separation between predictor/feature selection and model-fitting/validation. In addition to these methods, we used hidden test sets and also showed that the models could be used to predict perceptual responses from a completely different experiment, removing methodological biases arising from odorant preparation and presentation or any unforeseen regularities that machine learning algorithms could exploit but that are fundamentally task irrelevant for the analyst or researcher interested understanding rather than predicting.

5.6.3. Selecting optimal machine learning algorithms

The support vector machine (SVM) with the radial basis function kernel (RBF) outperformed random forest, regularized linear models (ridge and lasso), and linear SVM, tuning over L1 versus L2 regularization. However, gradient boosted decisions trees and tree ensembles such as random forest nevertheless approximated performance of RBF SVMs on the public volunteer data (Keller 2016), which was used in part for the DREAM analysis, and in certain cases outperformed it. This emphasizes that the optimal

algorithm is context dependent. To ensure consistency in our analysis of different psychophysical data sources, we did not report the results in this manner, that is, fitting the best performing algorithm each time. We instead aggregated multiple SVM models to improve generalizability. Algorithm selection and training was done using the R package, caret (classification and regression training)(Kuhn, 2008; R Development Core Team, 2016).

5.6.4. Cross-Study Predictions

For cross-study predictions, models were fit as shown in the Figure 1a pipeline with Dravnieks (1985) data. Multiple SVM models were fit with slightly different chemical features and their predictions were aggregated. This ensemble approach limits the tendency to overfit during the training phase.

Notably, chemicals do overlap between the two studies. Removing these chemicals (58) from Dravnieks (1985) significantly reduces the available training data. We instead removed the overlap from the Keller 2016 dataset, leaving 413 chemicals as a test set. Although theoretically all 146 perceptual descriptors could be assessed, the choice of “warm,” “sweaty,” “sweet,” and “chemical” depended on key differences in the perceptual descriptors available for the two studies, Keller (2016) and Dravnieks (1985). For instance, while Dravnieks (1985) used word strings in many cases such as “putrid, foul, decayed” to provide greater context, Keller 2016 opted for “decayed.” It is unclear what impact this difference might have and if it is non-trivial. The interpretation of the cross-study prediction becomes ambiguous as a result. Identically presented descriptors,

like “chemical,” “warm,” “sweaty,” and “sweet” are well defined cases for testing models across studies.

5.6.5. Network analyses and visualizations

5.6.5.1. Matrices for network

Chemical and perceptual descriptor relationships were modeled as bipartite graphs from an incidence matrix with perceptual descriptors as rows and columns the combined, unique optimal chemical feature sets. The optimal feature sets are from iteratively fitting a random forest model on 100 different partitions of the Dravnieks (1985) training data. We ranked the features based on the random forest importance over the partitions. Several different perceptual descriptor-chemical feature matrices were assembled by varying the number of ranked features per descriptor (e.g. Top 3,5,10). Incidence matrices from the top 3, 5, or 10 chemical features are therefore identical except for the number of columns (unique chemical features). Factor analysis was performed to reduce the number of perceptual descriptors for clarifying network plots as in Figure 5.2B. This was run using the factanal function in addition to functions in the nFactors (Raiche, 2010) R package for factor extraction.

Specifically, values in the incidence matrices are 1 or 0; the optimal chemical features for each perceptual descriptor are 1, otherwise 0. This amounts to a sparse matrix with the non-zero values indicating relationships among the optimal physicochemical features and the perceptual descriptors. Collectively, these binary strings are likened to a set of

combinatorial chemical feature codes for the Dravnieks (1985) perceptual descriptors. We subsequently separated the bipartite graph for clarity into its constituent, adjacency matrices, which are symmetrical, $m \times m$ and $n \times n$, matrices, with m denoting rows (perceptual descriptors) and n the columns (chemical features) in the original incidence matrix. An adjacency matrix can be obtained by multiplying an incidence matrix by its transpose.

5.6.5.2. Clustering networks

Several methods are available for identifying modules, communities or clusters in networks assembled from adjacency matrices. We tested several, selecting the Louvain algorithm based on its higher modularity score for Dravnieks (1985) data. Actual or observed network properties were in turn compared to 10,000 random network simulations (Erdos-Renyi) of approximately identical size and density. The actual network properties differed from those generated through the random simulation.

5.6.5.3. Tools for network analysis and visualization

Graph analyses were done using the igraph package (Csardi & Nepusz, 2006) in R, plots with ggplot2 (Wickham & Chang, 2016) and functions from the ggnetwork package for visualizing the networks.

5.6.6. eMolecules Predictions and Network Representation

The eMolecules predictions are from Dravnieks descriptor models trained on the % Usage (0-100 ratings), with detailed performance in Figure 5.5A and Figure 5.6B. The regression-based models predict or estimate these ratings for the eMolecules chemicals. Because the Dravnieks training set is not structurally exhaustive, we applied two filters to further sort the predictions. These include (1) an atom pair fingerprint based on commonly occurring feature sets in biologically active compounds (Cao, Charisi, Cheng, Jiang, & Girke, 2008) and (2) the % usage values of the chemicals at the top end of the distribution (% Usage). Initially, the % usage values for the top chemicals (exemplars) per descriptor were applied to filter the predictions. For each descriptor, the reduced set was then compared to the physicochemical features of the exemplar chemicals using atom pair fingerprints. Since atom pairs are a coarse representation of complex 3D molecules, we applied a Tanimoto similarity coefficient threshold of .25. This ensured that predictions per descriptor displayed basic 2D features that overlapped with the Dravnieks exemplar chemicals, while exploring new structural patterns or motifs that are potentially missed in 2D comparisons. Notably, projecting from a small chemical training set to a larger chemical set potentially amplifies noise in the training data, which should be considered in the interpretation.

5.6.7. Enriched Substructures/Cores

Enriched cores were analyzed using RDKit through Python (Landrum, 2006; Python Core Team, 2015). The algorithm performs an exhaustive search for maximum a

common substructure among a set of chemicals. In practice, larger sets often yield less substantive cores. To remedy this, the algorithm includes a threshold parameter that relaxes the proportion of chemicals containing the core. We used a threshold of .5, requiring that half of the chemicals from the top 10 contained the core.

5.6.7. Support Vector Machine

Training the support vector machine (SVM) involves identifying a set of parameters that optimize a cost function, where cost 1 and cost 0 correspond to training chemicals labeled as “Active” and “Inactive,” respectively. θ^T is the scoring function or output of the support vector machine. If the output is ≥ 0 , the prediction is “Active.” The function (f) is a kernel function.

$$SVM\ Cost = \min_{\theta} C \sum_{i=1}^m y^{(i)} cost_1(\theta^T f^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T f^{(i)}) + \frac{1}{2} \sum_{j=1}^n \theta_j^2$$

The kernel determines the shape of the decision boundary between the active and inactive chemicals from the training set. The radial basis function (RBF) or Gaussian kernel enables the learning of more complex, non-linear boundaries. It is therefore well suited for problems in which the biologically active chemicals cannot be properly classified as a linear function of physicochemical properties. This kernel computes the similarity for each chemical (x) and a set of landmarks (l), where σ^2 is a tunable parameter determined by the problem and data. The similarity with respect to these landmarks is used to predict new chemicals (“Active” vs. “Inactive”).

$$\text{Gaussian Kernel} = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$$

5.6.7.1 Model Performance Metrics

The Area under the ROC Curve (**AUC**) assesses the true positive rate (TPR or sensitivity) as a function of the false positive rate (FPR or 1-specificity) while varying the probability threshold (T) for a label (Active/Inactive). If the computed probability score (x) is greater than the threshold (T), the observation is assigned to the active class. Integrating the curve provides an estimate of classifier performance, with the top left corner giving an AUC of 1.0 denoting maximum sensitivity to detect all targets or actives in the data without any false positives. The theoretical random classifier is reported at AUC = 0.5.

$$\mathbf{TPR}(T) = \int_T^\infty \mathbf{f}_1(x) \, dx$$

$$\mathbf{FPR}(T) = \int_T^\infty \mathbf{f}_0(x) \, dx$$

Where \mathbf{T} is a variable threshold and \mathbf{x} is a probability score

However, we generated classifiers that are more authentic than theoretical random classification, shuffling the chemical feature values in the models and statistically comparing the mean AUCs across multiple partitions of the data. This controls against optimally tuned algorithms predicting well simply because of specific predictor attributes (e.g. range, mean, median, and variance) or models that are of a specific size (number of predictors) performing well even with shuffled values. Additionally, biological data sets

are often small, with stimuli or chemicals that—rather than random selection—reflect research biases, possibly leading to optimistic validation estimates without the proper controls. We used the AUC with classification-based training, such as to predict binary labels (Active/Inactive). For classification-based training we initially converted the % usage into a binary label (Active/Inactive) using the top 10% of the distribution as the cutoff. To provide additional context, we showed performance estimates varying the cutoff as well. The basis for a classification-based performance metric was the often top-heavy distribution of the % usage. It is for instance possibly not as relevant for models to accurately predict chemicals with minimal % usage. Rather, it is preferable for models to accurately predict whether a chemical will smell “Sweet” or not.

To provide further clarity we also reported multiple performance metrics including the correlation between the predicted and observed % usage, the root mean squared error (RMSE), and mean absolute error (MAE): **RMSE:** Root mean squared error is the square root of the mean difference between predicted values and those observed (% usage). It is the average prediction error on the same scale as the target or outcome being predicted. We supplied this metric because the correlation coefficient (R) is not always an accurate representation of model performance and classification of exemplar chemicals required an arbitrary cutoff (e.g. 90th percentile). We reported the correlation coefficient, R , between the predicted and observed % usage due to its previous use with human perceptual data. **MAE:** Mean absolute error is the mean of the absolute difference between predicted and observed (% usage). It thus assigns equal weight to all prediction errors, whether large or small.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y - \hat{y})^2}{N}}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y - \hat{y}|; \text{ where, } \hat{y} = \text{predicted and } y = \text{observed}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}; \text{ where, } TP = \text{True Positive and } FN = \text{False Negative}$$

$$\text{Specificity} = \frac{TN}{TN+FP}; \text{ where, } TN = \text{True Negative and } FP = \text{False Positive}$$

Chapter 6

Predicting human olfactory perception using odorant receptor activities

The prediction of human odor perception from 3D structure (Chapter 5) implies a combinatorial coding scheme, as different perceptual qualities of an odorant were best predicted by unique sets of structural and physicochemical features. Organizing diverse perceptual qualities according to such features is nevertheless abstract. Fundamentally, it is the odorant receptor that is activated or inactivated by these features and therefore computational models that deal directly with the receptor activation data are potentially more accurate and interpretable. Though the structure-based approach of Chapter 5 proved successful, I next analyzed the use of human odorant receptor data to address some of the limitations of the Chapter 5 approach.

Importantly, *in vivo* odorant receptor recordings are only possible under special circumstances such as in surgical patients that are not representative study samples. Efforts to understand the human odorant receptor code have therefore relied on *in vitro* heterologous systems. This in combination with whole genome sequencing studies have offered some progress. But the *in vitro* modeling step still represents a significant bottleneck. Accordingly, here, I ask three questions that build off of the previous chapter: (1) Can computation expand the putative ligand space and help accelerate discoveries?; (2) is *in vitro* odorant receptor activity a suitable alternative to the structure-based approach to predicting human perception?; and (3) what is gained with the approach in terms of insight into odor coding?

6.1. Introduction

In humans a single odorant molecule might be described by different perceptual descriptions, influenced by culture, language, and experience (Majid and Kruspe, 2018). Such complexities suggest that while olfactory circuitry is structurally similar across species, language or experience, which is dynamic and constantly evolving, could be a strong determinant of perceptual experience for humans. But even though the implication is that odor perception should be highly subjective, studies have shown that genetic variability in odorant receptors contributes to odor perception. Equally, machine learning has accurately predicted perceptual descriptors of odorants from chemical features, suggesting that physicochemical properties influence perception (Debnath et al., 2019; Gutiérrez et al., 2018; Keller et al., 2017; Khan et al., 2007; Licon et al., 2019; Nozaki and Nakamoto, 2016; Sanchez-Lengeling et al., 2019). Moreover, modeling human odor perception using a large semantic similarity space has shown that accurate predictions of perceptual ratings are possible even when training and prediction are done on completely different study samples. That is, in aggregate human perceptual descriptors do not appear to be arbitrarily used and are generalizable (Gutiérrez et al., 2018).

The connection between odorant receptor activity and perception is not as well defined. It is unclear that the activity of specific ORs confers odor identity. For instance, while the human odorant receptor OR5AN1 is highly selective to musk-smelling chemicals, less selective ORs also respond to these chemicals (Ahmed et al., 2018). In simpler systems like insects, there is some evidence that activation or inhibition of certain odorant receptors is sufficient to drive behaviors from attraction and aversion to

courtship, supporting the possibility of an underlying olfactory receptor code for perception (Chihara et al., 2014; Dweck et al., 2013; Kurtovic et al., 2007; MacWilliam et al., 2018; Stensmyr et al., 2012; Suh et al., 2004). Since these genetic studies are not feasible in humans, it is not yet clear how an olfactory receptor code can be generalizable, or whether it exists. It is however becoming increasingly plausible that there is indeed a perceptual code in humans. A few key odorant receptors have been reported for perceptual attributes other than musk (Shirasu et al., 2014) such as onion (Noe et al., 2017), general food-related volatiles (Geithe et al., 2017) and steroids (Keller et al., 2007). Sequence variation in the OR7D4 receptor has been shown to alter the perception of androstenone from a “sweaty,” unpleasant smell to one that is mildly “sweet” and pleasant (Keller et al., 2007). More recently, the specific amino acid residues of OR5AN1 that are responsible for its high selectivity to musk-smelling chemicals have also been confirmed (Ahmed et al., 2018). These studies were possible due to three types of information: (1) perceptual responses of humans (2) the odorant receptors that detect the chemicals from heterologous expression systems, and (3) genetic studies (Trimmer et al., 2019). Obtaining this information is not trivial for reasons that include the difficulty of receptor deorphanization and that behavioral responses are known for only a fraction of the purported volatile space, due to low throughput data collection with human volunteers.

Although some of these limitations are not easily overcome, we reasoned that it would be of interest to leverage machine learning/artificial intelligence to better understand the ligands of odorant receptors and clarify the role of odorant receptor

activity on human perceptual coding. Most prior machine learning efforts have focused on modeling odor perception according to the chemical features of odorants. While these studies have shown promise and provide evidence for the physicochemical basis of odor perception, chemical features alone do not offer clear insight into biological coding, as this would require additional information about the olfactory receptors that odorants activate. Moreover, it is an extremely challenging task to isolate the olfactory receptors that are relevant to a percept.

Here, we tested if human odorant receptor responses from heterologous assays could be used in lieu of chemical features for modeling human odor perception, and also developed models incorporating both approaches. We first created machine learning models to predict ligands for 34 human ORs. We could then use these models to evaluate how OR activity predicted perceptual descriptors. To start, we focused on hundreds of chemicals that human volunteers previously evaluated (Keller and Vosshall, 2016), and selected ORs that best predicted perceptual descriptors on a portion of training chemicals. Surprisingly, the prediction accuracy for models of only a few top scoring ORs compared favorably to large physicochemical feature models on 69 test chemicals (Keller et al., 2017), emphasizing that a small percentage of the OR pool is particularly useful for a given percept. This also suggested that specific subsets of ORs may be highly tuned to certain perceptual qualities, as implied in a prior network analysis of odorant receptors and perceptual descriptors (Bak et al., 2019).

6.2. Results

6.2.1. Modeling OR responses using chemical features

Each odorant receptor is activated by a unique set of chemicals, and together the large olfactory receptor family can detect a vast chemical space. We compiled a database of 84 deorphanized human ORs and 54 allelic variants which have been tested with multiple odorants, altogether adding up to ~170 (Adipietro et al., 2012; Braun et al., 2007; Charlier et al., 2012; Cook et al., 2009; Fujita et al., 2007; Gonzalez-Kristeller et al., 2015; Jacquier et al., 2006; Jaeger et al., 2013; Keller et al., 2007; Mainland et al., 2014; Mashukova et al., 2006; Maßberg et al., 2015; Matarazzo et al., 2005; McRae et al., 2012; Menashe et al., 2007; Neuhaus et al., 2006; Saito et al., 2009; Sanz et al., 2005; Schmiedeberg et al., 2007; Shirasu et al., 2014; Spehr et al., 2003; Topin et al., 2014). In order to generate more comprehensive odor response profiles of these ORs, we used machine learning to model structure-activity relationships. Among the 138 ORs, only 34 have a sufficient number of known ligands for machine learning models. For each of the 34 ORs, predictive chemical features were identified from the known ligands (Figure 6.1A). We validated the models by predicting ligands on a subset of odorants that were randomly left out of the training data set, repeating this several times. The prediction success was high for the 34 models (avg. AUC = 0.88; shuffled chemical features avg, AUC = .051, $p < 10^{-32}$) (Figure 6.1B; Figure 6.7A-B; Table 6.1).

The OR-ligand predictive models also gave us an opportunity to identify new ligands for the 34 ORs from a large chemical library (~450,000). In doing so, we developed a theoretical space that expands the existing data by a factor of 10 (Figure

6.1C). Enriched structural features were identifiable among the top predicted ligands for each OR, illustrating simple 2D features that are presumably important for activating each receptor (Figure 6.1D; Table 6.2).

6.2.2. Modeling odorant percepts from OR responses

A key question in olfaction is how activities of ORs contribute to different perceptual qualities. Specific receptors contribute to androstenone perception (Keller et al., 2007), however little is known about odorants commonly perceived as flavors and fragrances. One possibility is that their perception depends on a model similar to androstenone, and one or few receptors contribute to perception. Alternatively, a model involving a combinatorial code of a large number of ORs is also possible, particularly since unlike androstenone, most odorants activate multiple ORs. In order to test these possibilities, we performed a series of analyses on a large dataset of human odor perception (Keller and Vosshall, 2016). Not only were a large number of chemicals tested by volunteers in this study, but computational studies have successfully demonstrated structure-percept relationships (Gutiérrez et al., 2018; Keller et al., 2017; Kepple and Koulakov, 2017; Sanchez-Lengeling et al., 2019). However, several odorants used in the behavior study have not been tested for OR activities. We therefore used the OR-ligand models in the previous section to estimate activity for chemicals, designating similar training and testing chemicals as described before (Keller et al., 2017) (407 training; 69 testing chemicals) (Figure 6.8A). Models containing only a few optimal ORs successfully predicted the perceptual descriptors for test chemicals (average test AUC = 0.78) (Figure

6.8B), particularly when compared to a similar approach based on different physicochemical feature encodings rather than ORs (Figure 6.8C). Lastly, because the activity on the 34 ORs was known for some chemicals in the Keller 2016 study, and it was unclear if this might affect the results, we revisited the analysis with these chemicals removed (326 train; 54 test chemicals). Test performance was not significantly reduced, compared to the earlier analysis ($p = 0.234$).

We next turned to another psychophysical study (ATLAS) that evaluated 146 perceptual descriptors for ~150 odorants. As before, most perceptual descriptors were well predicted from a small subset of ORs, despite the larger, more diverse descriptor pool in this study (Figure 6.2A) (Top 50 best-performing: 10 ORs: avg. AUC = 0.84). When we compared the performance of the OR activity to the optimal chemical features, 47/146 perceptual descriptors were better predicted using the ORs. In light of this excellent performance, we further investigated ORs whose contributions to percept predictions are highest. Interestingly, only a few select ORs contributed strongly to the prediction of some perceptual descriptors (Figure 6.3A).

In order to expand the scope and utilize activity information from the 104 ORs with few known ligands, we computed 3D similarity between chemicals in the ATLAS study and the OR ligands (Mahé et al., 2006) and identified the most likely active compound for each of the 104 ORs (Methods). When incorporating these additional ORs into the pipeline, predictions improved slightly for some perceptual descriptors. Among the top 50 best predicted descriptors, smaller OR models were significantly better than all 138 ORs on the test data (10 ORs AUC = 0.84; 138 OR AUC = 0.80, $t = 2.76$, $p = 0.007$),

suggesting that the additional information was not often useful (Figure 6.9A). These 138 ORs still represent just a third of the human OR repertoire, and we anticipate our approach will help identify even better sets of ORs that are tuned to specific perceptual qualities as more human ORs get deorphanized.

6.2.3. Modeling odorant percepts from OR responses and chemical features

Because many previous efforts have focused on predicting odor perception with chemical features (Keller et al., 2017), we tested if adding ORs could improve the predictions. We selected OR6P1, an OR ranked highly for “Cinnamon,” as a test case and added it to 34 optimal chemical features. Interestingly, we found a notable increase in predictive success on test chemicals (mean AUC chemical features: 0.77, mean AUC chemical features + OR6P1 = 0.81) (Figure 6.10A).

To determine if ORs could improve predictive models in an unbiased manner across the 146 perceptual descriptors, we combined the odor response information of the 138 ORs and the chemical features, selecting a small subset of important ORs and chemical features to create Machine Learning models (Figure 6.4A). We found that removing the top ranked ORs and replacing them with those of lesser importance negatively impacted predictions for some descriptors (Figure 6.4B). If we permute the activities of the optimal or top-ranked ORs for a given descriptor, the overall test performance significantly dropped ($p < 10^{-7}$), with 82% of descriptors better predicted with non-permuted ORs (Table 6.3). Collectively, these results indicate that specific ORs

appear to contribute more than others and perceptual predictions are generally improved by including ORs (Table 6.4)

In order to visualize relationships among the perceptual descriptors based on predictive ORs and chemical features, we next performed a cluster analysis. When examining the clustering based only on perceptual ratings of chemicals (Figure 6.5A), we found the top 5 predictive ORs grouped the perceptual descriptors similarly (Figure 6.5B). Notably, randomly selecting 5 ORs failed to produce any meaningful groups or clusters of perceptual descriptors (Figure 6.5C). Combining the most predictive ORs and chemical features improved the clustering of perceptual descriptors (Figure 6.5D). Overall, the descriptors that were best clustered in Figure 5A (Silhouette Width > 0.3) matched completely or partially with Figure 6.5B and 6.5D, with the exception of “Fishy” and “Kippery.” This indicates that relationships among perceptual descriptors in the ATLAS training set are somewhat preserved in OR activity or chemical feature models, even when only a small amount of chemical or OR information is included in each model.

6.2.4. Modeling with in vivo OR response data from Drosophila

One of the interesting observations we have is that only a few ORs are picked and are sufficient to create predictive models of odor perception. However, the perceptual descriptor – to – OR mapping we analyzed here represents data from only ~20% of the human OR repertoire and one possibility is that when more ORs are available to pick from, a larger number will be selected computationally as optimal. In order to understand

the contribution of specific olfactory receptors to behavior in a system where a large fraction of odorant receptors have been deorphanized, we turned to the *Drosophila melanogaster* model system. In vivo odor-response spectra are known for several odorants for the majority of odorant receptors (Ors) and olfactory receptor neurons (ORNs) in the adults, as well as the behavioral valence (attraction vs aversion) to these odorants.

We adapted our approach to predict behavioral valence of flies (Figure 6.6A) and we could do so with significant success using a small number of important chemical features and electrophysiologically measured responses from sensory neurons. Similar to what we observed with human ORs, a subset of the in vivo *Drosophila* Or activities was favored for odor valence predictions, beyond collections of numerous chemical features (Figure 6.6B). Evaluating the best valence predictors for test chemicals from a combined set of Or/ORN activities and chemical features indicated that the Or/ORNs significantly contributed to odor valence predictions, consistent with the in vitro human data ($R^2 = 0.66$; Shuffle ORs + Chemical Features: 0.51 , $p = 0.007$) (Figure 6.6C). These results also suggested that a small number of *Drosophila* Or/ORN activities is highly predictive on the same set of test chemicals. Interestingly, additional Ors/ORNs failed to improve predictions (Or/ORN subset: $R^2 = 0.53$; All other ORs: $R^2 = 0.40$, $p = 0.015$) (Table 6.5). While this type of analysis remains to be done in humans, the results from flies suggest that even when a more comprehensive receptor or neuron array is added, only a small subset of the available receptors appears information rich as far as behavioral decisions are concerned (Figure 6.6D).

6.3. Discussion

While previous machine learning pipelines have found some success using chemical features, selecting the optimal feature sets for predictions of perception not well defined. We found that human odorant responses from heterologous assays could be used with comparable and sometimes better predictive success. In part, the result is anticipated by the fact that each OR is presumably selective to very specific physicochemical features themselves. Both the human perceptual descriptor and fly valence predictions suggest that a substantive portion of odor identity arises early in the processing stream, at the olfactory receptors, based on high predictive success rates (~76-91%). It is likely that the remaining portion depends on experience-dependent modulation, supporting a downstream model with reliance on distributed neuronal networks for human perceptual coding. Our findings support a “primacy model” which holds that a small number of distinct and overlapping olfactory receptor activity profiles encode odor identity (Wilson et al., 2017). Although increasing concentration activates more receptors, the highest sensitivity receptors start responding first as an animal approaches an odor source and presumably continue to convey the identity. Such a model is consistent with the findings reported here and others (Weiss et al., 2012) because it appears that only a few ORs contribute to a perceptual descriptor and it is therefore also tractable to predict how a chemical smells from specific physicochemical properties.

Nevertheless, it is unclear how information arising early in the olfactory pathway is preserved along the complex circuits and can in fact lead to generalizable perceptual features. The spatial organization of the olfactory receptor neurons and glomeruli are for

one not well preserved in the piriform cortex. Unlike the retinotopic and tonotopic patterning observed in the visual and auditory cortices, representing spatiotemporal properties of visual and auditory stimuli as they are processed at sensory neurons, piriform activity appears randomly distributed, without a clear mapping of physicochemical features (Stettler and Axel, 2009). A combination of computational models and calcium imaging has however shown piriform circuits, though they are qualitatively different, can support perceptual invariance amid changes in concentration and across different odorants (Roland et al., 2017; Schaffer et al., 2018). Similarly, neural tracing experiments in mice support that while olfactory circuitry differs from other sensory modalities, odor related-information is represented along equally structured neuroanatomical pathways, as in the piriform output projecting to the orbitofrontal cortex (Chen et al., 2014).

One possibility is that only 1 or few receptors of the many that detect an odorant actually convey percept. The evolutionary landscape should accordingly be coupled to biologically relevant or frequently encountered features of the chemical space, as has been implied by characterizations of receptors highly tuned for musk and onion-related compounds (Ahmed et al., 2018; Noe et al., 2017), in addition to the highly conserved trace amine-associated receptors (TAARs) and their importance in modulating behavioral output in mice (Dewan et al., 2018). In our analyses, the OR specialized for musk was not a top candidate for musk predictions but contributed strongly to predictions of “sweaty.” Since methods for selecting and ranking ORs depend on characteristics of the available data, interpretations should be cautious, acknowledging that the human OR data

are sparse and the participants and chemical sets from the ATLAS and Keller studies are not exhaustive. Yet from these same considerations the positive results achieved are unexpected, especially when compared to predictions of odor perception using chemical features.

Odorant receptors (ORs) are also expressed in non-olfactory tissues. Ligands for certain ORs have been shown to modify the function and proliferation of multiple cell types. Although the precise mechanisms are not well defined, ORs represent promising therapeutic targets. Ligands for ORs such as OR51E1, OR10G7 and OR1D2, which were included in this study, are candidate treatments for conditions ranging from prostate cancer and chronic obstructive pulmonary disease (COPD) to atopic dermatitis (Kalbe et al., 2016; Maßberg et al., 2016; Tham et al., 2019). We therefore anticipate that the predictions and the analysis of known and candidate OR ligands from this study will also have value in non-olfactory studies.

6.3.1. Limitations of the study

The computational approach presented in the study is restricted by training sets from previously deorphanized human odorant receptors (OR) determined by in vitro assays. Only a small fraction of the human ORs family has been deorphanized in vitro, therefore limiting the identification of the optimal predictive ORs in this study. Moreover, the number of chemicals with well-defined perceptual profiles determined behaviorally is small relative to the space of chemicals that are likely to have odorant properties. Since

the computational approach we outlined depends on the size and complexity of OR and perceptual datasets, our results should be interpreted alongside these limitations.

6.4. Figures

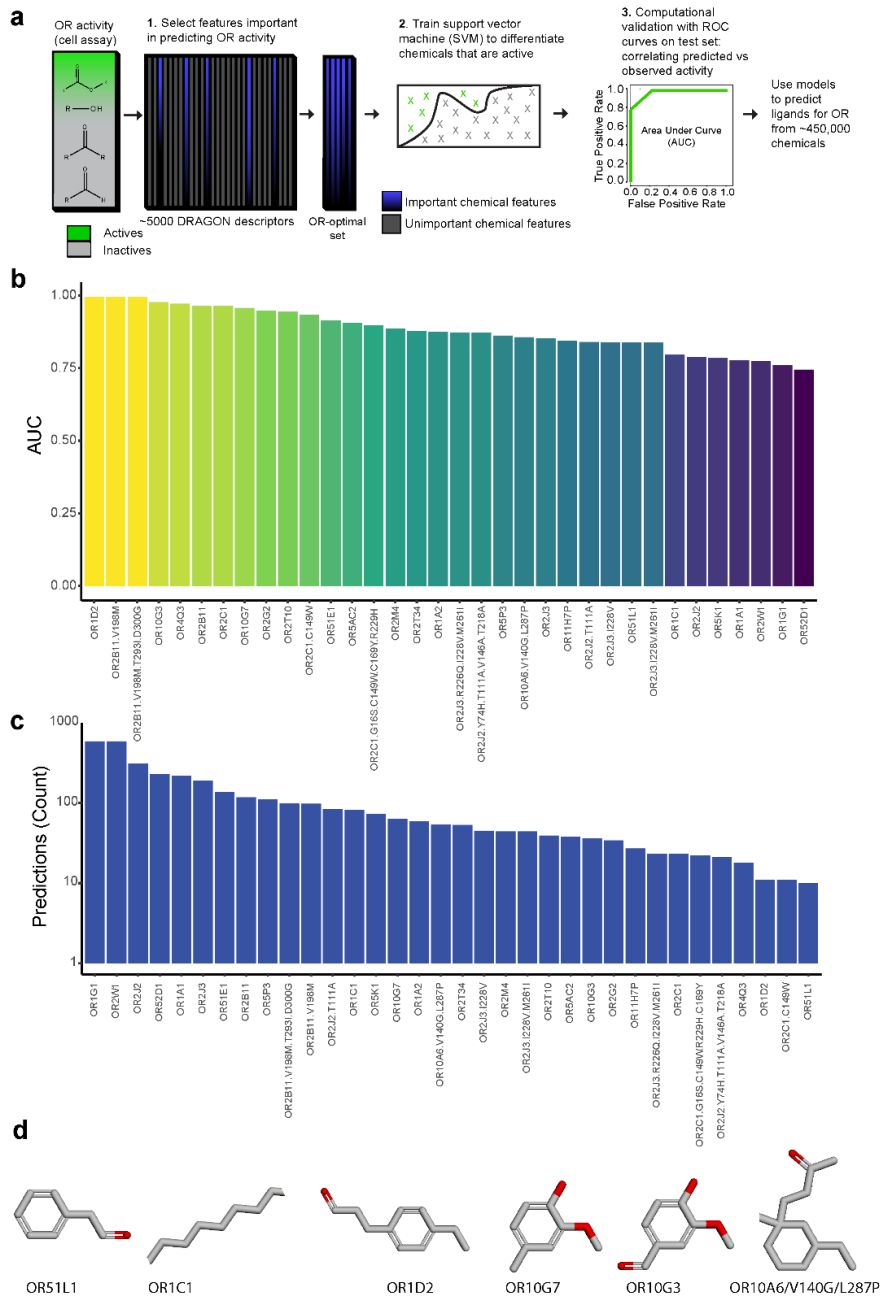


Fig. 1

Figure 6.1. Features of human Odorant Receptor ligands can be learned and new ligands predicted. a) Pipeline for generating probability scores for chemicals with perceptual data. Starting with lists of ligands from heterologous assay data SVM models learn predictive physicochemical features for a subset of human ORs and OR variants with >2 ligands (34 total). These trained models in turn predict new chemicals such as those with known perceptual profiles. b) Average performance of 34 OR models using repeated 10-fold cross validation. c) Number of ligands predicted for each of the 34 ORs in ~400,000 eMolecules library after filtering based on optimal probability score cutoffs and structural similarity to known ligands. d) Sample of enriched substructures among the top 10 predicted chemicals for indicated ORs. Only substructures that were non-trivial and present in at least half of the 10 highest scoring chemical ligands. A comprehensive table of substructures for other receptors is provided in Table 6.2.

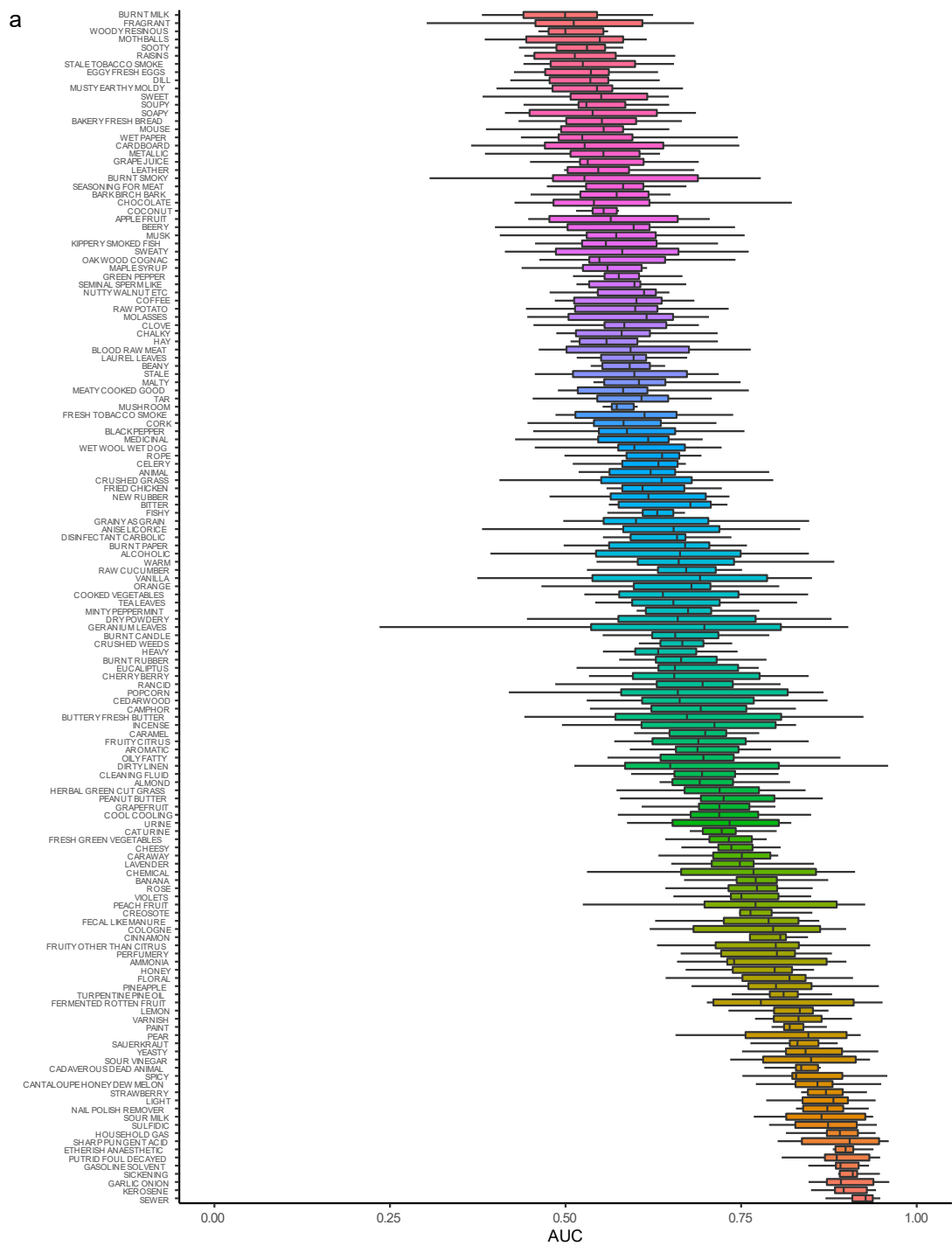


Fig. 2

Figure 6.2. OR activity can model diverse olfactory percepts in human studies.

Performance of RBF SVM models trained with 10 ORs for ATLAS study data. The top 10% usage chemicals are predicted for all 146 perceptual descriptors in the study. Successful classification of these chemicals is reported as the mean Area-Under-the-Curve (AUC) over repeated 10-fold cross validation (10-fold repeated 5 times; 50 folds total). To limit biased validation, the procedure was run twice, setting aside different test chemicals, determining important OR subsets to predict the descriptors with these chemicals excluded, then ensuring that the cross-validated AUC comprised 60% completely hidden chemicals. The variability in the plot is the standard deviation over these two distinct runs. High variability may arise as the top 10% usage is computed from the training data. SVM: Support Vector Machine; RBF: Radial Basis Function; additional algorithm details in Methods.



Fig. 3

Figure 6.3. Contribution of ORs to perceptual models. A) Importance of individual ORs for machine learning models of each of the 146 ATLAS perceptual descriptors. The heatmap is generated by fitting models for each OR separately and scaling relative to maximum AUC (100). Importance is shown with the most important ORs in blue. Labels for the perceptual descriptors (Y-axis) and ORs (X-axis) are arranged relative to similar importance values.

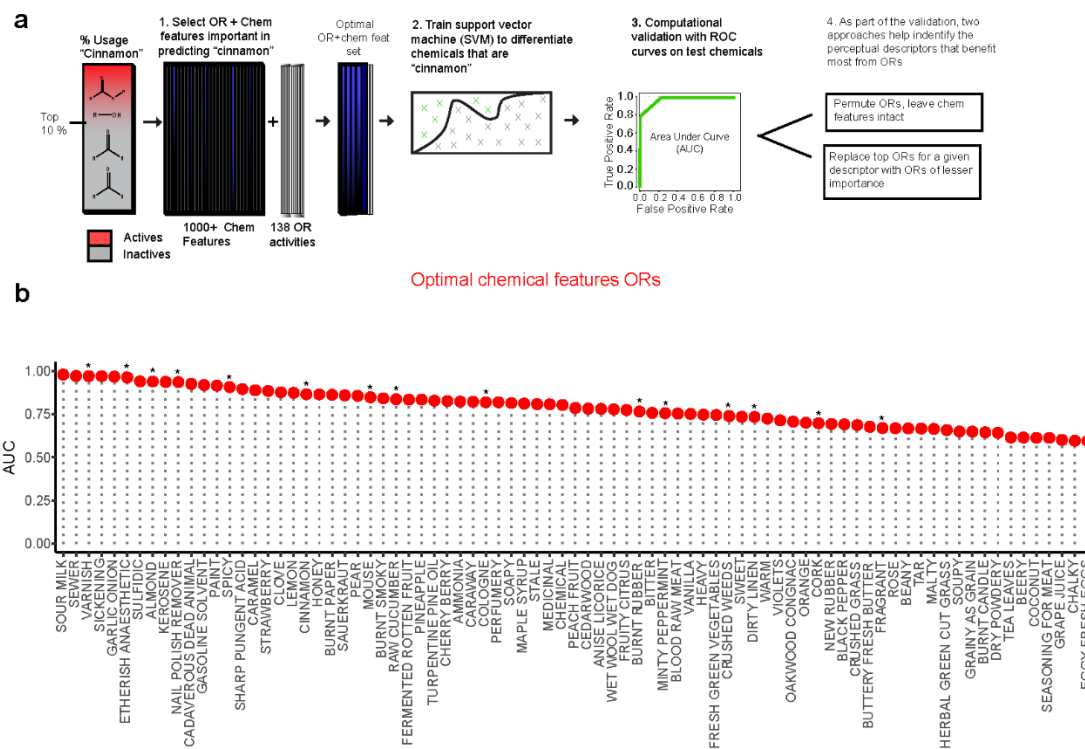


Figure 6.4. Few Odorant receptors are needed to predict perceptual descriptors. A) Schematic of the approach to selecting a small number of important chemical features and ORs, followed by model-fitting. Two methods, including replacing top-ranking ORs with those of lesser importance and permuting (shuffling) the OR activities, help identify perceptual descriptors where ORs contribute relative to chemical features. To standardize the analysis, the training and validation are as outlined in Figure 6.2. B) Combined chemical feature-OR models predict the top 10 % usage of ATLAS perceptual descriptors. The (*) symbol signifies a notable decrease in performance occurred if the ORs were replaced with ones of lesser importance (One-tailed Independent Samples T-test, $p \leq 0.05$). For the comparison with permuted or shuffled OR activities, other metrics, and benchmarking relative to chemicals features, see Tables 6.3-6.4.

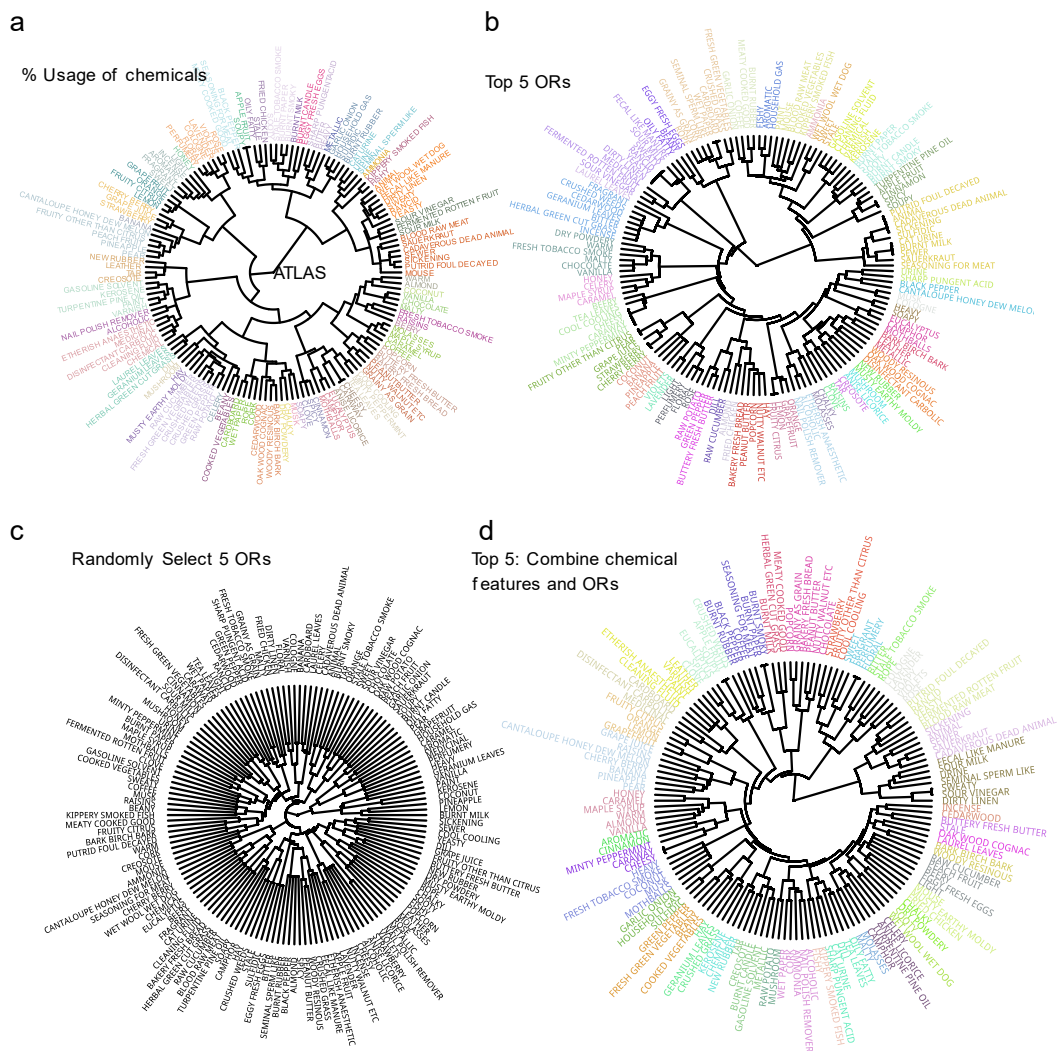


Figure 6.5. A few key ORs or chemical features sensibly cluster the perceptual descriptors. A) Dendrogram representation of the Euclidean distances among perceptual descriptors based on overlap of perceptual response data (% Usage) from chemicals in the ATLAS study. B) Dendrogram from the top 5 ORs picked per perceptual descriptor. C) Dendrogram created from 5 randomly chosen ORs per perceptual descriptor. D) Dendrogram from the 5 best overall predictors including OR and chemical features per perceptual descriptor. Clustering is hierarchical and based on Euclidean distance (A) or the Jaccard distance (B-D). Cluster number (colored branches) inferred from gap statistic across bootstrap samples.

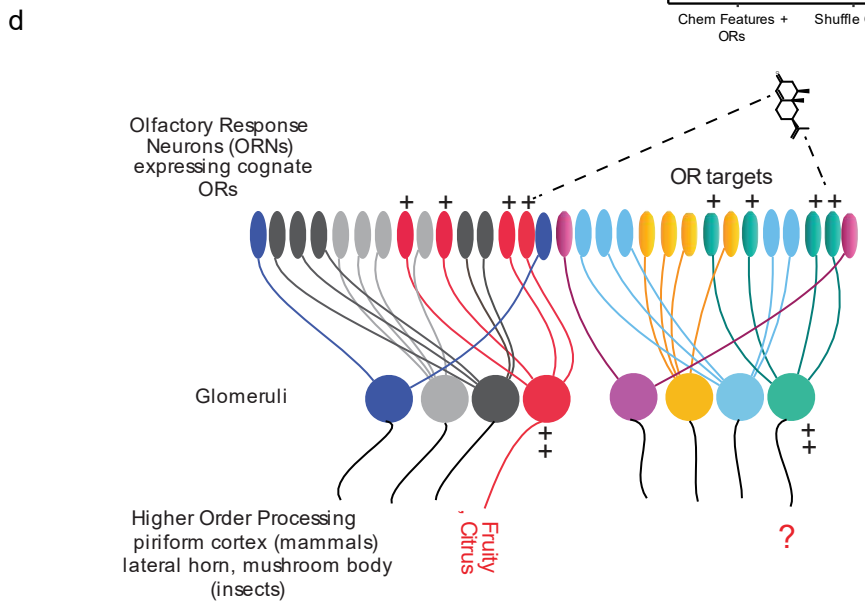
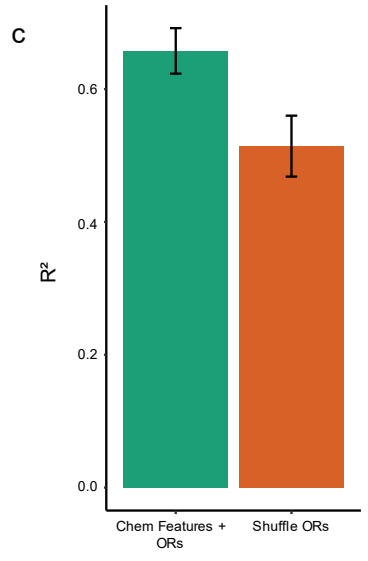
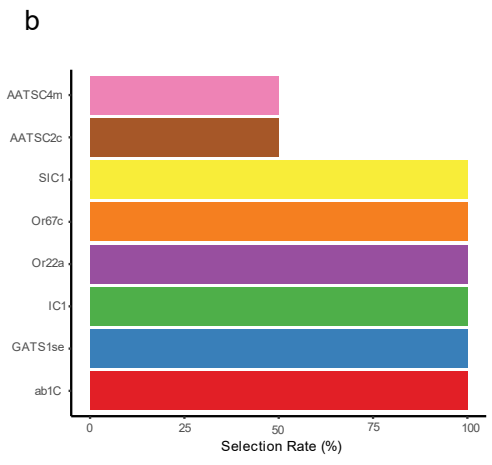
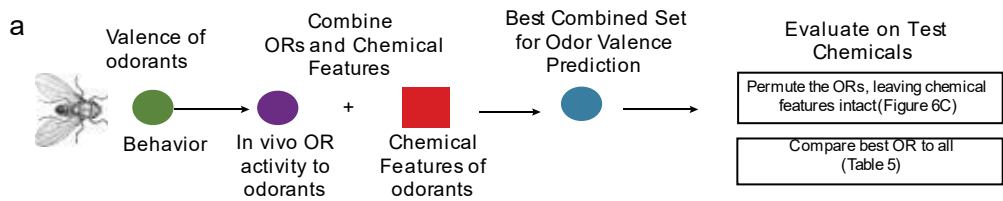


Fig. 6

Figure 6.6. Few odorant receptor activities in *Drosophila* are highly predictive of valence. A) Schematic for applying machine learning to identify optimal predictors of odor valence in *Drosophila* from in vivo neural activity and chemical features. The best combined model is evaluated on test chemicals. OR contributions to *Drosophila* odor valence are assessed by shuffling the OR activities in the combined model as well as comparing the best OR vs all (Table 6.5). B) Selecting chemical features and in vivo OR activities that optimally predict odor valence. Recursive feature elimination (RFE) is run twice to accomplish this. Selection in the top 10 over these runs is plotted as a percent. Additional details on selecting optimal models in methods. C) The best combined model is evaluated on test chemicals, with and without the OR activities shuffled. D) Generic model displaying a many-to-one mapping between ORNs and glomeruli. Although there are > 1 responding units (ORs), information that confers perceptual character is restricted to a smaller subset of the input.

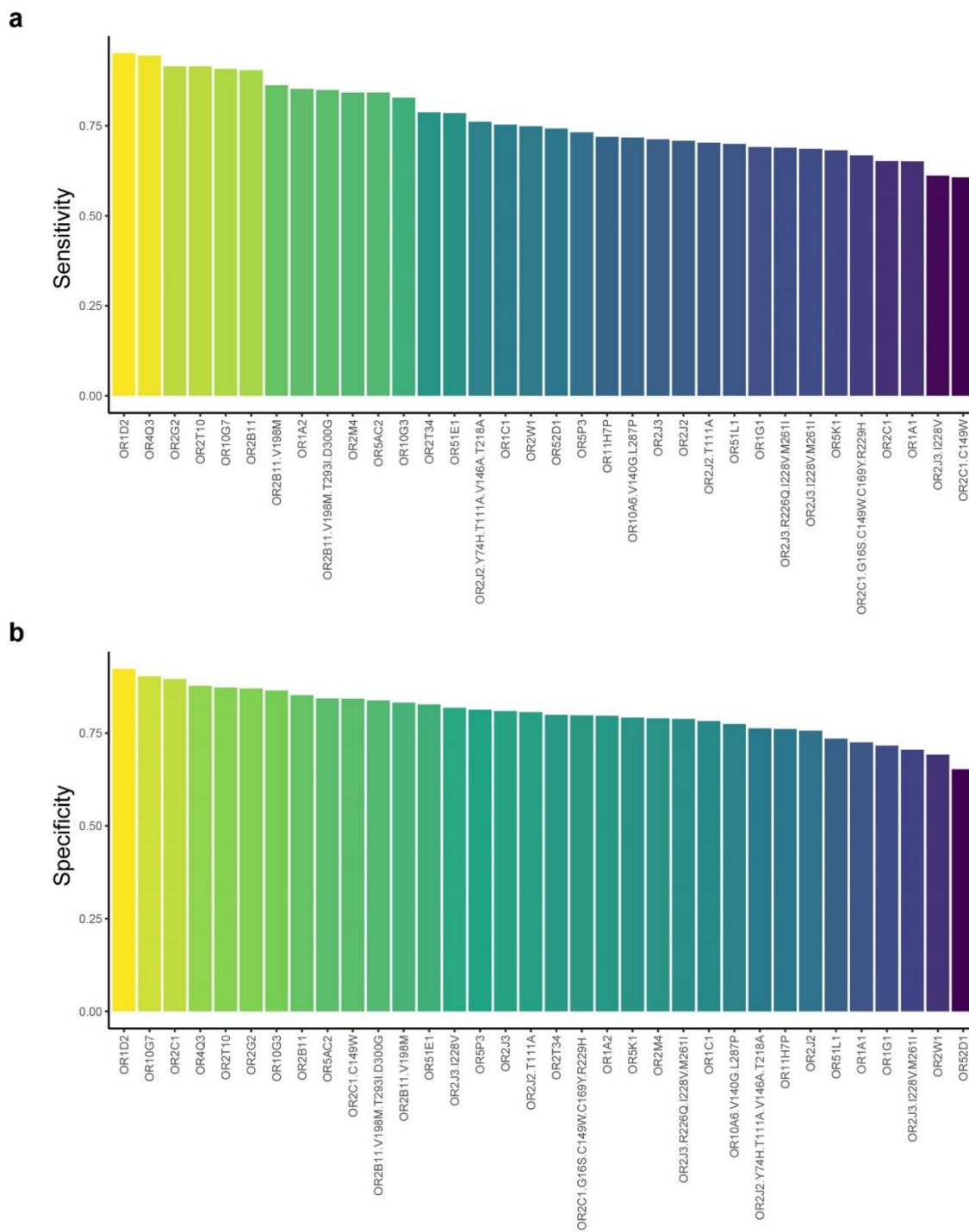


Figure 6.7. Detailed performance of models predicting activity on 34 human ORs. A) The average sensitivity of the 34 OR models and B) average specificity over repeated cross validation folds (10-fold CV repeated 10 times).

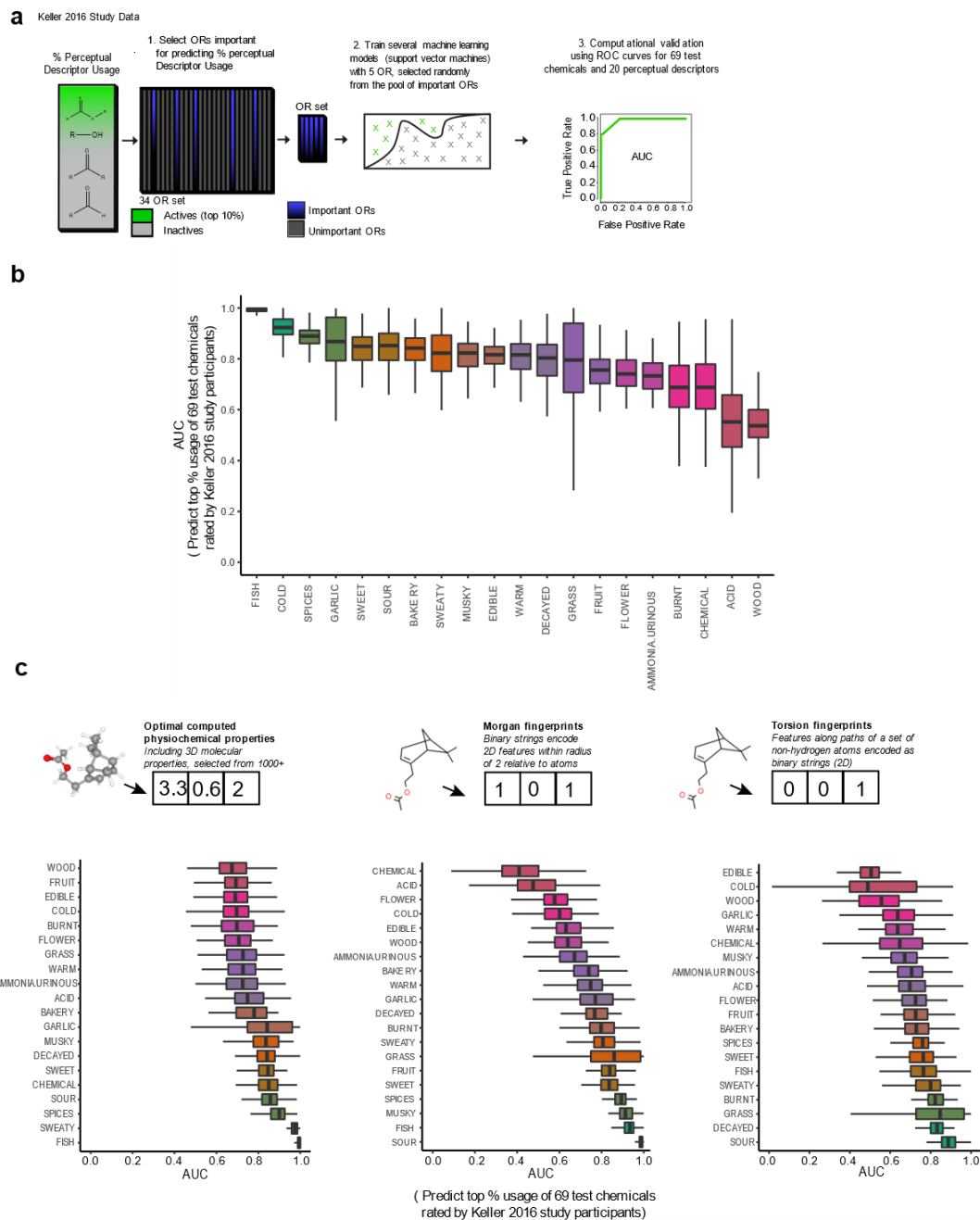


Figure 8

Figure 6.8. Human OR activity or chemical features predict perceptual data from Keller 2016 study participants. A) Pipeline for making predictive models for odor perception from ORs for the Keller 2016 perceptual data. Classification cutoffs for the 69 test chemicals are determined from 407 training chemicals. B) Classifying the top 10% of usage for 69 hidden test set chemicals; performance is reported as the area under the ROC curve (AUC). Prediction of the % usage is an aggregate of 5 SVM models, each sampling 5 ORs from the top 10. The OR ranking is determined by recursive feature elimination over cross validation (10-fold repeated 10 times) with 407 training chemicals. C) Prediction of the 69 test chemicals with models trained on various chemical feature representations. *Left*, physicochemical features are computed for optimized 3D structures and 5 SVM models sample 35 top ranked chemicals features. Plotted performance is the aggregated prediction. *Middle*, predictions from an SVM model trained on Morgan circular fingerprints. During training, low variance bit positions are dropped to improve the fit. *Right*, predictions from an SVM model trained on topological torsion fingerprints, dropping low variance bit positions during training. All plots display the standard deviation over 100 bootstrap samples of the 69 test chemicals.

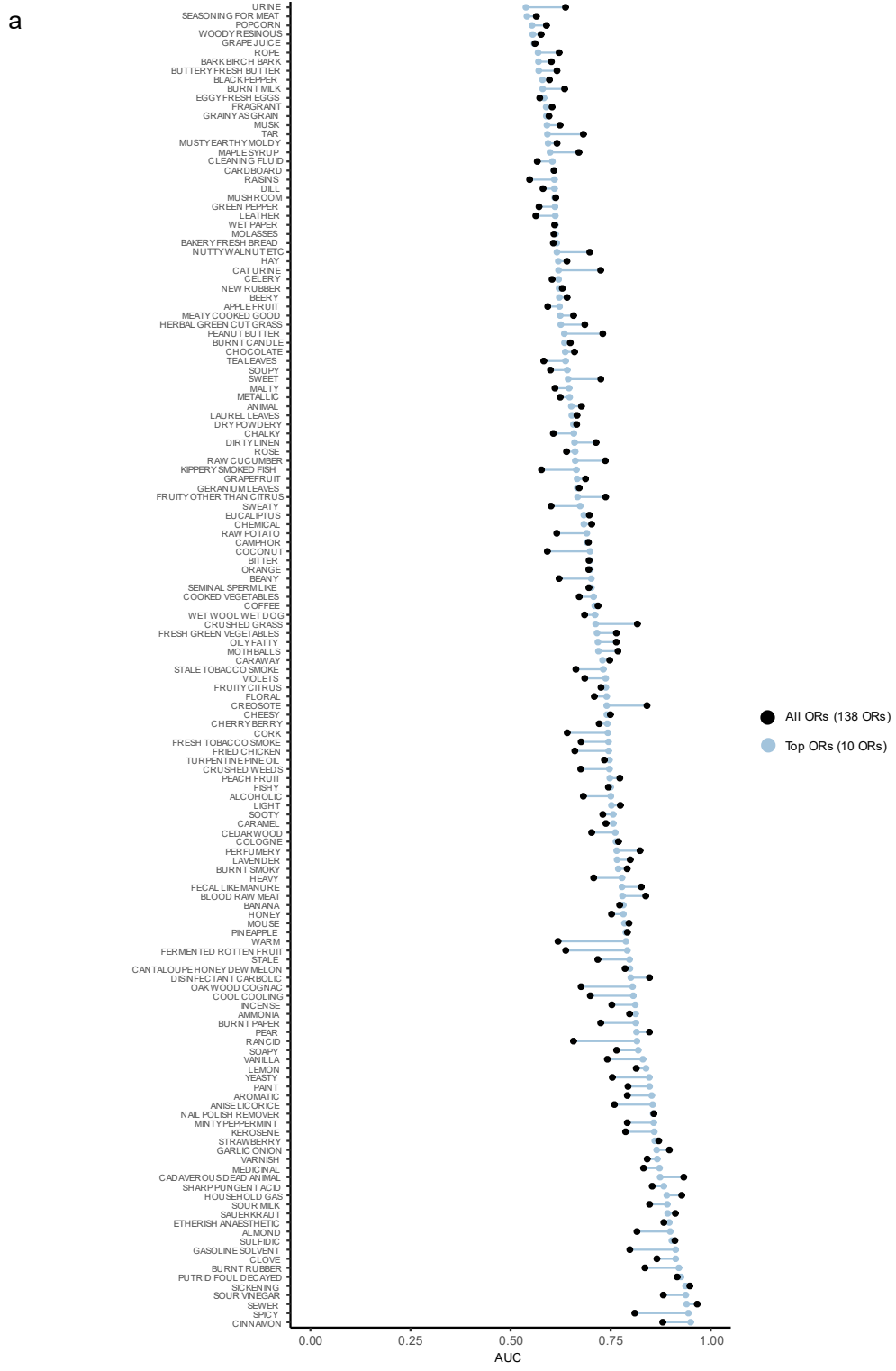


Fig. 9

Figure 6.9. Small subsets of ORs optimize predictions of most perceptual descriptors.
A) Comparison between models fit with 10 or 138 ORs on ATLAS study data. Black colored dots show the performance using all ORs while blue dots show the performance using 10 ORs.

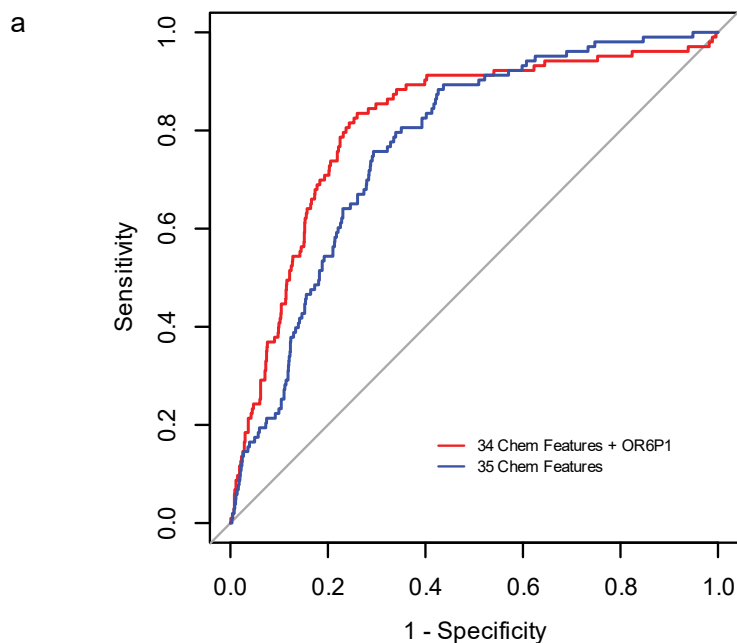


Figure 6.10. Adding an OR to a chemical feature model improves odor perception predictions. A) An OR (OR6P1) that was selected for predicting the % usage of the descriptor “Cinnamon” in the ATLAS study is validated with and without physicochemical features by ROC analysis. The examples are RBF SVM models that are trained and tested on equivalent chemicals sets. (RBF = Radial Basis Function; SVM = Support Vector Machine).

6.5. Tables

Table 6.1

OR	Metric	Value
OR10G3	AUC	0.9764246
OR10G7	AUC	0.9571523
OR11H7P	AUC	0.8438965
OR1A2	AUC	0.8741326
OR1C1	AUC	0.7959467
OR1D2	AUC	0.9958090
OR1G1	AUC	0.7593999
OR2B11	AUC	0.9649722
OR2G2	AUC	0.9473866
OR2J2	AUC	0.7880475
OR2J3	AUC	0.8525910
OR2M4	AUC	0.8861412
OR2T10	AUC	0.9438511
OR2T34	AUC	0.8784901
OR2W1	AUC	0.7741072
OR4Q3	AUC	0.9717862
OR51E1	AUC	0.9142729
OR51L1	AUC	0.8378983
OR52D1	AUC	0.7448325
OR5AC2	AUC	0.9059954
OR5K1	AUC	0.7841736
OR5P3	AUC	0.8616590
OR10A6.V140G.L287P	AUC	0.8566833
OR2B11.V198M	AUC	0.9949333
OR2B11.V198M.T293I.D300G	AUC	0.9946667
OR2C1.C149W	AUC	0.9331333
OR2C1.G16S.C149W.C169Y.R229H	AUC	0.8975333
OR2J2.T111A	AUC	0.8400286
OR2J2.Y74H.T111A.V146A.T218A	AUC	0.8713667
OR2J3.I228V.M261I	AUC	0.8373600
OR2J3.I228V	AUC	0.8386250
OR2J3.R226Q.I228V.M261I	AUC	0.8731667
OR1A1	AUC	0.7773069
OR2C1	AUC	0.9636074
OR10G3	Sens	0.8283333
OR10G7	Sens	0.9080000
OR11H7P	Sens	0.7196667
OR1A2	Sens	0.8520000
OR1C1	Sens	0.7533333

OR1D2	Sens	0.9516667
OR1G1	Sens	0.6909206
OR2B11	Sens	0.9042286
OR2G2	Sens	0.9150000
OR2J2	Sens	0.7081444
OR2J3	Sens	0.7126667
OR2M4	Sens	0.8422222
OR2T10	Sens	0.9150000
OR2T34	Sens	0.7876190
OR2W1	Sens	0.7489190
OR4Q3	Sens	0.9450000
OR51E1	Sens	0.7851667
OR51L1	Sens	0.7000000
OR52D1	Sens	0.7425333
OR5AC2	Sens	0.8422222
OR5K1	Sens	0.6816667
OR5P3	Sens	0.7321667
OR10A6.V140G.L287P	Sens	0.7173333
OR2B11.V198M	Sens	0.8626667
OR2B11.V198M.T293L.D300G	Sens	0.8493333
OR2C1.C149W	Sens	0.6066667
OR2C1.G16S.C149W.C169Y.R229H	Sens	0.6680000
OR2J2.T111A	Sens	0.7034286
OR2J2.Y74H.T111A.V146A.T218A	Sens	0.7613333
OR2J3.I228V.M261I	Sens	0.6864000
OR2J3.I228V	Sens	0.6120000
OR2J3.R226Q.I228V.M261I	Sens	0.6893333
OR1A1	Sens	0.6514000
OR2C1	Sens	0.6520000
OR10G3	Spec	0.8646471
OR10G7	Spec	0.9032486
OR11H7P	Spec	0.7610861
OR1A2	Spec	0.7966139
OR1C1	Spec	0.7823186
OR1D2	Spec	0.9231667
OR1G1	Spec	0.7163778
OR2B11	Spec	0.8521181
OR2G2	Spec	0.8702770
OR2J2	Spec	0.7565145
OR2J3	Spec	0.8094917
OR2M4	Spec	0.7900806
OR2T10	Spec	0.8730319
OR2T34	Spec	0.7994472

OR2W1	Spec	0.6918247
OR4Q3	Spec	0.8771373
OR51E1	Spec	0.8272121
OR51L1	Spec	0.7352230
OR52D1	Spec	0.6525470
OR5AC2	Spec	0.8433472
OR5K1	Spec	0.7919028
OR5P3	Spec	0.8127833
OR10A6.V140G.L287P	Spec	0.7745800
OR2B11.V198M	Spec	0.8317600
OR2B11.V198M.T293I.D300G	Spec	0.8378600
OR2C1.C149W	Spec	0.8425800
OR2C1.G16S.C149W.C169Y.R229H	Spec	0.7978600
OR2J2.T111A	Spec	0.8063800
OR2J2.Y74H.T111A.V146A.T218A	Spec	0.7627000
OR2J3.I228V.M261I	Spec	0.7053000
OR2J3.I228V	Spec	0.8181000
OR2J3.R226Q.I228V.M261I	Spec	0.7882000
OR1A1	Spec	0.7249889
OR2C1	Spec	0.8959733

Table 6.1. Summary of ROC analysis for models predicting activity on 34 human ORs. Averages for the prediction performance of Figure 1 models over validation, including the sensitivity (true positive rate), specificity (false positive rate = 1-specificity), and overall AUC.

Table 6.2










Image	IDs	OR
	608374	OR10G7
	631771	OR1G1
	486255	OR2B11
	479671	OR2G2
	540140	OR2M4
	481045	OR2T10
	516498	OR2W1
	478561	OR51E1
	478221	OR51L1


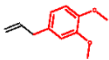

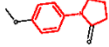





Image	IDs	OR
	477085	OR52D1
	494660	OR5K1
	479790	OR5P3
	631730	OR10A6.V140G.L287P
	535081	OR2B11.V198M
	483236	OR2B11.V198M.T293I.D300G
	509362	OR2C1.C149W
	481575	OR2C1.G16S.C149W.C169Y.R229H
	492074	OR2J2.T111A

Table 6.2. Enriched substructures among predicted ligands for 34 human ORs.

Additional enriched cores/substructures for the labeled ORs, highlighting the core on an exemplar chemical from the eMolecules predictions. ID is the eMolecules identifier for the representative chemical, which is among the top 10 predictions for the labeled OR. Bonds and atoms are colored black. The enriched substructure is in red.

Table 6.3

Perc.Descriptor	AUC.Actual.ORs	AUC.Perm.ORs	P (Actual v. Perm)	AUC.Random.ORs	P (Actual v. Random)
ALCOHOLIC	0.68	0.62	6.624121e-03	0.72	0.723178607
ALMOND	0.94	0.93	2.487726e-01	0.91	0.034877069
AMMONIA	0.82	0.76	7.437701e-02	0.81	0.411008447
ANIMAL	0.62	0.62	4.653889e-01	0.70	0.964195594
ANISE LICORICE	0.78	0.63	9.318801e-03	0.76	0.361514077
APPLE FRUIT	0.56	0.61	6.724894e-01	0.66	0.800616700
AROMATIC	0.71	0.61	7.262549e-02	0.73	0.689074811
BAKERY FRESH BREAD	0.57	0.62	9.059523e-01	0.58	0.719289158
BANANA	0.85	0.64	9.809746e-03	0.87	0.709005236
BARK BIRCH BARK	0.57	0.63	8.966433e-01	0.59	0.615803346
BEANY	0.67	0.67	5.570947e-01	0.61	0.156366133
BEERY	0.61	0.60	3.367454e-01	0.59	0.288927182
BITTER	0.76	0.61	4.906788e-03	0.72	0.274457885
BLACK PEPPER	0.69	0.63	3.041965e-02	0.61	0.060708625
BLOOD RAW MEAT	0.75	0.57	3.027819e-04	0.69	0.131544847
BURNT CANDLE	0.64	0.59	1.217739e-01	0.62	0.298663633
BURNT MILK	0.61	0.60	4.028164e-01	0.62	0.554476157
BURNT PAPER	0.86	0.67	3.743915e-03	0.84	0.304135032
BURNT RUBBER	0.76	0.64	1.367015e-03	0.68	0.006654245
BURNT SMOKY	0.84	0.69	1.459191e-02	0.81	0.130453646
BUTTERY FRESH BUTTER	0.68	0.76	9.155857e-01	0.66	0.366338507
CADAVEROUS DEAD ANIMAL	0.92	0.83	1.386069e-04	0.92	0.388750235
CAMPHOR	0.63	0.57	4.832609e-02	0.70	0.945089522
CANTALOUPE HONEY DEW MELON	0.79	0.73	3.018547e-02	0.82	0.787234025
CARAMEL	0.89	0.61	8.795053e-04	0.86	0.175338953
CARAWAY	0.82	0.57	9.910237e-07	0.82	0.467041687
CARDBOARD	0.63	0.60	1.545006e-01	0.65	0.643885930
CAT URINE	0.55	0.62	9.862892e-01	0.59	0.887941827
CEDARWOOD	0.78	0.66	7.725876e-02	0.76	0.392239690
CELERY	0.61	0.70	9.874588e-01	0.66	0.861058602
CHALKY	0.60	0.59	4.697072e-01	0.58	0.362309551
CHEESY	0.79	0.58	5.814025e-04	0.85	0.969988448
CHEMICAL	0.80	0.72	5.454888e-04	0.76	0.187394329
CHERRY BERRY	0.82	0.65	1.870925e-03	0.82	0.400904651
CHOCOLATE	0.65	0.79	9.861704e-01	0.74	0.953634802
CINNAMON	0.86	0.78	3.247772e-02	0.73	0.003079151
CLEANING FLUID	0.76	0.75	2.807639e-01	0.81	0.921645505
CLOVE	0.88	0.60	1.088595e-04	0.85	0.264212595
COCONUT	0.61	0.59	2.942493e-01	0.60	0.373755189
COFFEE	0.74	0.57	1.325299e-04	0.76	0.689137219
COLOGNE	0.82	0.74	4.557177e-02	0.75	0.034619751
COOKED VEGETABLES	0.79	0.60	5.312181e-04	0.79	0.539445780
COOL COOLING	0.77	0.64	5.920177e-03	0.80	0.795019055
CORK	0.70	0.61	5.878820e-02	0.61	0.022586784
CREOSOTE	0.80	0.61	6.339020e-01	0.83	0.893222941
CRUSHED GRASS	0.69	0.62	1.174218e-02	0.64	0.143682506
CRUSHED WEEDS	0.74	0.61	4.921776e-03	0.66	0.013566817
DILL	0.57	0.73	9.992897e-01	0.59	0.781535973
DIRTY LINEN	0.73	0.61	7.730862e-04	0.67	0.044158634
DISINFECTANT CARBOLIC	0.79	0.90	9.958961e-01	0.89	0.988115114
DRY POWDERY	0.64	0.63	4.372117e-01	0.62	0.342789868
EGGY FRESH EGGS	0.59	0.56	7.667215e-02	0.59	0.461463704
ETHERISH ANAESTHETIC	0.96	0.90	2.183860e-03	0.93	0.027905380
EUCALIPTUS	0.69	0.57	1.904247e-03	0.76	0.981148786
FECAL LIKE MANURE	0.70	0.77	9.785591e-01	0.71	0.548529155
FERMENTED ROTTEN FRUIT	0.83	0.64	1.552895e-04	0.73	0.068225780

Perc.Descriptor	AUC.Actual.ORs	AUC.Perm.ORs	P (Actual v. Perm)	AUC.Random.ORs	P (Actual v. Random)
FISHY	0.76	0.58	4.824185e-04	0.77	0.584032635
FLORAL	0.75	0.81	8.525523e-01	0.76	0.571314301
FRAGRANT	0.67	0.54	5.315367e-02	0.57	0.002180182
FRESH GREEN VEGETABLES	0.74	0.81	8.271564e-04	0.72	0.251533345
FRESH TOBACCO SMOKE	0.63	0.57	5.314839e-02	0.64	0.612188803
FRIED CHICKEN	0.64	0.62	2.208670e-01	0.64	0.503480983
FRUITY CITRUS	0.77	0.74	1.227108e-01	0.77	0.415394866
FRUITY OTHER THAN CITRUS	0.75	0.54	2.031718e-04	0.76	0.545691914
GARLIC ONION	0.97	0.63	2.531550e-05	0.96	0.231282678
GASOLINE SOLVENT	0.92	0.85	2.785957e-03	0.90	0.128023122
GERANIUM LEAVES	0.62	0.54	8.120492e-02	0.65	0.746064731
GRAINY AS GRAIN	0.65	0.88	7.357469e-01	0.61	0.241183239
GRAPE JUICE	0.60	0.63	8.148000e-01	0.58	0.249827462
GRAPEFRUIT	0.73	0.69	6.902408e-02	0.76	0.832562571
HAY	0.63	0.58	2.191222e-01	0.64	0.635096113
HEAVY	0.75	0.52	8.808084e-03	0.70	0.180959523
HERBAL GREEN CUT GRASS	0.66	0.57	2.178443e-02	0.61	0.135613486
HONEY	0.86	0.72	6.441975e-03	0.84	0.105893090
HOUSEHOLD GAS	0.91	0.81	1.735901e-05	0.93	0.730659466
INCENSE	0.74	0.75	5.484723e-01	0.77	0.716919634
KEROSENE	0.94	0.76	1.019985e-05	0.93	0.327626884
KIPPERY SMOKED FISH	0.57	0.59	7.332818e-01	0.58	0.577061232
LAUREL LEAVES	0.63	0.56	8.698926e-02	0.67	0.838973837
LAVENDER	0.73	0.71	2.760870e-01	0.74	0.619993047
LEATHER	0.68	0.61	1.345040e-02	0.79	0.992390918
LEMON	0.67	0.67	2.979404e-01	0.65	0.136759987
LIGHT	0.77	0.73	1.434929e-01	0.85	0.917798531
MALTY	0.66	0.70	7.065725e-01	0.65	0.398116520
MAPLE SYRUP	0.81	0.62	6.505141e-04	0.78	0.196680862
MEATY COOKED GOOD	0.68	0.55	1.619274e-03	0.69	0.532624907
MEDICINAL	0.81	0.70	2.199711e-02	0.72	0.094122938
METALLIC	0.62	0.59	3.428863e-01	0.63	0.625422059
MINTY PEPPERMINT	0.75	0.59	4.713820e-04	0.67	0.005963238
MOTHBALLS	0.76	0.75	4.015789e-01	0.80	0.939709264
MOUSE	0.85	0.69	6.052005e-05	0.74	0.001296670
MUSHROOM	0.67	0.61	5.507031e-02	0.68	0.612807182
MUSK	0.64	0.61	2.600284e-01	0.65	0.641149881
MUSTY EARTHY MOLDY	0.54	0.59	8.496130e-01	0.56	0.726368591
NAIL POLISH REMOVER	0.93	0.93	4.717329e-01	0.87	0.008637722
NEW RUBBER	0.69	0.60	1.422405e-02	0.69	0.486580423
NUTTY WALNUT ETC	0.79	0.80	6.001993e-01	0.79	0.541252687
OAK WOOD COGNAC	0.71	0.60	5.334112e-02	0.70	0.449304705
OILY FATTY	0.75	0.68	5.442055e-02	0.78	0.703181219
ORANGE	0.70	0.63	4.857429e-02	0.70	0.487048230
PAINT	0.91	0.79	8.734811e-05	0.91	0.461934408
PEACH FRUIT	0.79	0.58	5.484549e-05	0.76	0.266667783
PEANUT BUTTER	0.83	0.78	1.108906e-01	0.86	0.826427173
PEAR	0.86	0.80	1.820044e-02	0.84	0.381644634
PERFUMERY	0.82	0.78	1.393034e-02	0.81	0.429208350
PINEAPPLE	0.83	0.83	4.383805e-01	0.82	0.400655067
POPCORN	0.59	0.68	9.254312e-01	0.65	0.890432139
PUTRID FOUL DECAYED	0.87	0.80	2.422996e-03	0.82	0.971155085
RAISINS	0.57	0.59	7.519006e-01	0.60	0.789009023
RANCID	0.80	0.69	1.046710e-02	0.80	0.573830524
RAW CUCUMBER	0.84	0.61	5.236949e-04	0.76	0.027564181
RAW POTATO	0.63	0.56	8.758358e-03	0.66	0.714216712
ROPE	0.60	0.58	3.075818e-01	0.61	0.668058005

Perc.Descriptor	AUC.Actual.ORs	AUC.Perm.ORs	P (Actual v. Perm)	AUC.Random.ORs	P (Actual v. Random)
ROSE	0.67	0.59	1.072022e-01	0.65	0.330191386
SAUERKRAUT	0.66	0.68	8.345338e-01	0.63	0.187159140
SEASONING FOR MEAT	0.61	0.58	1.750854e-01	0.59	0.258163703
SEMINAL SPERM LIKE	0.68	0.59	2.888288e-02	0.75	0.911515570
SEWER	0.97	0.65	1.684756e-03	0.97	0.404678453
SHARP PUNGENT ACID	0.69	0.64	7.879184e-03	0.67	0.285085817
SICKENING	0.97	0.66	9.330160e-04	0.96	0.173703843
SOAPY	0.61	0.78	2.807849e-01	0.77	0.176364339
SOOTY	0.64	0.57	6.044304e-02	0.68	0.830745272
SOUPY	0.65	0.61	2.153393e-01	0.57	0.201392365
SOUR MILK	0.98	0.68	1.311867e-03	0.97	0.395627632
SOUR VINEGAR	0.66	0.78	2.440581e-02	0.88	0.781798703
SPICY	0.91	0.73	1.473271e-04	0.85	0.044485570
STALE	0.61	0.71	3.902480e-03	0.80	0.384288452
STALE TOBACCO SMOKE	0.68	0.65	1.798344e-01	0.71	0.741998826
STRAWBERRY	0.68	0.56	1.729606e-04	0.65	0.273525575
SULFIDIC	0.94	0.75	8.166529e-04	0.93	0.298005800
SWEATY	0.68	0.61	6.619308e-02	0.70	0.661394584
SWEET	0.73	0.65	3.036319e-02	0.72	0.397528666
TAR	0.67	0.66	3.736882e-01	0.66	0.451093414
TEA LEAVES	0.61	0.62	5.105895e-01	0.57	0.113352387
TURPENTINE PINE OIL	0.63	0.72	3.150109e-04	0.80	0.237355229
URINE	0.74	0.62	1.306574e-03	0.76	0.694702395
VANILLA	0.75	0.72	9.793389e-02	0.69	0.180342803
VARNISH	0.97	0.63	6.922921e-04	0.94	0.017280947
VIOLETS	0.71	0.74	7.678331e-01	0.71	0.448119936
WARM	0.72	0.57	1.437458e-03	0.70	0.319931645
WET PAPER	0.60	0.69	9.795118e-01	0.62	0.673009454
WET WOOL WET DOG	0.78	0.59	1.104795e-03	0.74	0.098478406
WOODY RESINOUS	0.65	0.60	1.157541e-01	0.68	0.734801220
YEASTY	0.66	0.64	2.672057e-01	0.69	0.690830166

Table 6.3. Detailed analysis of odor perception predictions using chemical features and ORs. Combined OR and chemical feature model performance using the ATLAS study data. In one condition, ORs are replaced with those of lesser importance (“Random”). In the second condition, OR activities for the best combined set are permuted (shuffled). The chemical features are intact in both conditions. Training and testing chemicals are equivalent.

Table 6.4

Best Predicted	Metric	Chem Features	ORs + Chem Features
Top 5	AUC	0.9502134	0.9708750
Top 10	AUC	0.9210808	0.9565362
Top 20	AUC	0.8969860	0.9253587
Top 25	AUC	0.8822129	0.9117761
Top 50	AUC	0.8314386	0.8558157
Top 5	R	0.6559822	0.7265567
Top 10	R	0.5977552	0.6949485
Top 20	R	0.5818210	0.6447922
Top 25	R	0.5734141	0.6267827
Top 50	R	0.5271159	0.5474018

Table 6.4. Comparing predictions of odor perception with chemical features or chemical features and ORs. Summary table containing the average test performance for the best predicted perceptual descriptors in the ATLAS study across two metrics (R and AUC) and different predictor set combinations (e.g. ORs and chemical features). R is the correlation between the predicted and observed % usage of the perceptual descriptors. The AUC is the classification success for chemicals in the top 10% of usage.

Table 6.5

RMSE	Rsquared	MAE	Model	Method
0.307164	0.403876	0.252166	All ORs	RBF SVM
0.261997	0.531187	0.218281	Optimal ORs	RBF SVM

Table 6.5. A subset of ORN activities best predict the drosophila preference index compared to the full set of ORNs. Shows the root mean square error (RMSE) and R squared metrics (Methods), metrics which quantify the relationship between the predicted and observed T-maze Preference Index (PI). The RMSE quantifies the average error in the prediction, where a smaller value is better. Mean absolute error (MAE) is a metric that is related to RMSE; however, by excluding the square root, this metric does not overemphasize one or few predictions that may be very far from the observed. The R squared, in contrast, quantifies the relationship between the variability in the predicted vs observed T-maze Preference Index (PI). The square root of the value is simply the correlation coefficient, often designated as ‘r’ or ‘R.’ This value should be as large as possible, with the maximum being 1.0. The “Model” column provides details about the type of fit, a subset of few ORNs versus all, whereas “Method” clarifies the algorithm that was used. ORN: Olfactory Response Neuron; OR: Olfactory Receptor; RBF SVM: Support Vector Machine with a Radial Basis Function kernel (details in Methods). Additional details on metrics provided in Methods under the Metrics heading.

6.6. Methods

6.6.1. Modeling OR ligands from chemical features

We trained SVM models to learn physicochemical features of the confirmed ligands for a subset of ORs whose response profiles are currently better characterized (34 total).

Different chemical features were encoded as binary fingerprints (1,0) (Klekota-Roth (Klekota and Roth, 2008), Morgan/Circular (Morgan, 1965), MACCs, Shortest Path, and Hybridization (Steinbeck et al., 2003). Chemical fingerprints can encode up to ~1000 bits and many are possibly uninformative. Kullback–Leibler (KL) divergence (Nisius and Bajorath, 2010) was used to select only those bits that maximized the distance between active and inactive compounds in the heterologous assay data. Predictions from these models provided probability scores for each OR-chemical pair for the ATLAS chemicals. This work relied on the chemistry development kit (CDK) (Steinbeck et al., 2003) as well as its R interface (Guha, 2007).

6.6.2. Enriched Substructures/Cores

Enriched cores were analyzed using RDKit through Python (Landrum, 2006). The algorithm is an exhaustive search for the maximum common substructure among chemicals. In practice, larger chemical sets often yield less substantive cores. To remedy this, the algorithm includes a threshold parameter that relaxes the proportion of chemicals containing the core. We used a threshold of .5, requiring that half of the top predicted chemicals contained the core.

6.6.3. ORs as predictors of perception

Despite several available data sources, most in vitro assays typically report a handful of ORs with multiple ligands and many others with few ligands (1 or 2 compounds that pass statistical thresholds). To incorporate the more narrowly tuned receptors, we computed an approximation of the 3D pharmacophore kernel (Mahé et al., 2006). Pharmacophore kernels are a versatile method for computing pairwise similarities among chemicals according to a set of standard features that are related to biological activity. Namely, similarity between ATLAS chemicals and known OR ligands was defined by the three-point Tanimoto coefficient, which is scaled to 0-1, with 1 being maximally similar. In cases where there were > 1 ligands for an OR the maximally similar ligand was used.

To incorporate the ORs with more ligands, we trained SVM models on physicochemical features of odorants with known activity. There were 34 ORs with sufficient training data for this approach. These models assigned probability scores for the 34 ORs to the perceptual study chemicals (ATLAS and Keller 2016). The Keller 2016 perceptual ratings were converted to the % usage, or the % of participants using a perceptual descriptor; that is, supplying a rating (0-100) for a given descriptor. The ATLAS study provides this metric.

The receiver operating characteristic (ROC) analysis or, in particular, the area under the curve (AUC) is based on transforming the rating that had been assigned to a perceptual descriptor by study participants into a classification label (active/inactive). The active chemicals are those within the top 10% of the ratings (% usage). However, as this cutoff is arbitrary, other metrics are supplied in supplementary materials for

comparison. These, in addition to the classification-based metric (ROC analysis), are explained in detail in the metrics section alongside their strengths and weaknesses for this specific problem. Unless noted in the figure legends the importance of an OR is not based on classification. Specific methods for evaluating importance are discussed below.

6.6.4. Computing chemical features to predict perceptual descriptors

We computed chemical features using the Python wrapper for the open source RDKit software (Landrum, 2006). This included chemicals features that were raw values, pertaining to features such as functional group counts and 3D geometries, which closely resemble the proprietary DRAGON software; the whole library is accessible through the mordred module (Moriwaki et al., 2018). We also computed Morgan/circular (radius =2) and topological torsion fingerprints. These use a hash function to encode different chemical features as fixed length binary strings (1024 bits).

6.6.5. Selecting important ORs in prediction of human perception

Important ORs were selected using a cross validated recursive feature elimination (10-fold, repeated 10 times), with the random forest (RF) algorithm or the support vector machine (SVM) algorithm. Random forest defines importance by permuting predictors and reporting the % increase in error. Random forest fits multiple decision trees on different bootstrap samples and supplies a consensus vote over the trees as the prediction. Bootstrap sampling leads to a portion of data being left out; the “out of bag” sample which is used to estimate the prediction performance. When a model is fit, the predictor

importance (% increase in error) is computed. The support vector machine, however, does not include an ‘out of bag’ sample and therefore the OR/chemical feature importance is computed externally by fitting non-linear regression models for each predictor.

By including the model-fitting inside a cross validation loop the importance is computed over multiple folds or portions of the training data rather than on the complete training set, which reduces bias in the predictors that are selected. The importance is in this context redefined as a selection rate (e.g. the rate the predictor was highly ranked).

6.6.6. Clustering

Clustering was performed with the `hcust` function in R using the Ward D2 method and the Euclidean distance for numerical matrices such as the perceptual ratings (Figure 5A) or 1-Jaccard distances for binary matrices (Figure 5B-D). 1000 bootstrap samples were used to select the optimal number of clusters, according to the gap statistic (1-standard error (SE) rule).

Quantification and statistical analysis

6.6.7. Support Vector Machine

Training the support vector machine (SVM) involves identifying a set of parameters that optimize a cost function, where cost 1 and cost 0 correspond to training chemicals labeled as “Active” and “Inactive,” respectively.

$$SVM\ Cost = \min_{\theta} C \sum_{i=1}^m y^{(i)} cost_1(\theta^T f^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T f^{(i)}) + \frac{1}{2} \sum_{j=1}^n \theta_j^2$$

Additionally, a kernel determines the shape of the decision boundary between the active and inactive chemicals from the training set. The radial basis function (RBF) or Gaussian kernel enables the learning of more complex, non-linear boundaries. It is therefore well suited for problems in which the physicochemical properties vary among the biologically active chemicals. This kernel computes the similarity for each chemical (x) and a set of landmarks (l), where σ^2 is a tunable parameter determined by the problem and data. The similarity with respect to these landmarks is used to predict new chemicals (“Active” vs. “Inactive”).

$$Gaussian\ Kernel = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$$

6.6.8. Metrics

The area under the roc curve (AUC) assesses the true positive rate (TPR or sensitivity) as a function of the false positive rate (FPR or 1-specificity) while varying the probability threshold (T) for a label (Active/Inactive). If the computed probability score (x) is greater than the threshold (T), the observation is assigned to the active class. Integrating the curve provides an estimate of classifier performance, with the top left corner giving an AUC of 1.0 denoting maximum sensitivity to detect all targets or actives in the data without any false positives. The theoretical random classifier is reported at AUC = 0.5.

$$TPR(T) = \int_T^{\infty} f_1(x) dx$$

$$FPR(T) = \int_T^{\infty} f_0(x) dx$$

Where T is a variable threshold and x is a probability score

However, we generated classifiers that are more authentic than theoretical random classification, shuffling the chemical feature (or OR) values in the models and statistically comparing the mean AUCs across multiple partitions of the data. This controls against optimally tuned algorithms predicting well simply because of specific predictor attributes (e.g. range, mean, median, and variance) or models that are of a specific size (number of predictors) performing well even with shuffled values. Additionally, biological data sets are often small, with stimuli or chemicals that—rather than random selection—reflect research biases, possibly leading to optimistic validation estimates without the proper controls. We used the AUC with classification-based training, such as to predict binary labels (Active/Inactive). For classification-based training we initially converted the % usage into a binary label (Active/Inactive) using the top 10% of the distribution as the cutoff. The basis for a classification-based performance metric was the often top-heavy distribution of the % usage. It is for instance possibly not as relevant for models to accurately predict chemicals with minimal % usage. Rather, it is preferable for models to accurately predict whether a chemical will smell “Sweet” or not.

To provide further clarity we also reported multiple performance metrics including the correlation between the predicted and observed % usage, the root mean squared error (RMSE), and mean absolute error (MAE): RMSE: Root mean squared error is the square root of the mean difference between predicted values and those observed (% usage). It is the average prediction error on the same scale as the target or outcome being

predicted. We supplied this metric because the correlation coefficient (R) is not always an accurate representation of model performance and classification of exemplar chemicals required an arbitrary cutoff (e.g. 90th percentile). We reported the correlation coefficient, R, between the predicted and observed % usage due to its previous use with human perceptual data. MAE: Mean absolute error is the mean of the absolute difference between predicted and observed (% usage). It thus assigns equal weight to all prediction errors, whether large or small.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y - \hat{y})^2}{N}}$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y - \hat{y}|; \text{ where, } \hat{y} = \text{predicted and } y = \text{observed}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}; \text{ where, } TP = \text{True Positive and } FN = \text{False Negative}$$

$$\text{Specificity} = \frac{TN}{TN+FP}; \text{ where, } TN = \text{True Negative and } FP = \text{False Positive}$$

Chapter 7

Adapting the computational pipeline to discovery of odorant-based Covid-19 drugs

7.1. Introduction

Earlier emphasis was on implementing machine learning pipelines to understand chemosensory processing in insects, humans, and then subsequently incorporating these findings into applications, such as safe chemical insect repellents. However, the framework outlined throughout is not tied to a specific problem. Indeed, toxicity, physical and cosmetic property modeling remain important for numerous problems. With the emergence of the COVID-19 the science has had to move at a faster rate, creating large amounts of data. To process and analyze this scientific data, computational tools, specifically machine learning, have become essential. Since this is consistent with the major theme of this work, it seemed critical to adapt the machine learning pipeline to help discover putative treatments for COVID-19. This chapter describes that effort.

7.1.1. Drug repurposing and discovery for COVID-19

SARS-CoV-2 is a novel coronavirus that is responsible for the COVID-19 disease which is a rapidly evolving global pandemic. Coronaviruses primarily target the upper respiratory tract and the lungs, with varying degrees of severity. Related corona viruses such as the SARS-CoV emerging in China in 2002 and the MERS-CoV in the Middle East in 2012 result in severe respiratory conditions. The SARS-CoV-2 also produces similarly severe respiratory conditions, albeit at a lower rate but with a higher contagion

factor (Sanche et al., 2020). Alarmingly, infected individuals may be asymptomatic carriers, presumably harboring the viral infection in the upper airway tract, increasing the likelihood of infecting populations that are most susceptible to severe complications (Bai et al., 2020; Z. Chen et al., 2020).

Although the mechanisms underlying SARS-CoV-2 infection are not completely understood, select human proteins are targets for the virus including ACE2 (Wan, Shang, Graham, Baric, & Li, 2020). The SARS-CoV-2 receptor binding domain (RBD) interacts strongly with the human ACE2 receptor and TMPRSS2 to enter a human cell (Yan et al., 2020). In addition to ACE2, a recent systems-level analyses of protein-protein interaction with peptides encoded in the SARS-CoV-2 genome identified ~300 additional human proteins, of which, 66 were considered suitable candidates for identification of therapeutics (Gordon et al., 2020). Gordon et. al. performed an in vitro assay with human cells expressing 26 SARS-CoV-2 proteins, which was followed by an analysis for high-confidence interactions. Of the 100s of reported interactions 66 were prioritized, and the authors subsequently mined and tested FDA approved drugs that were known or suspected to target these human proteins. Most of the human target proteins are overexpressed in the respiratory tract. Of particular note is the entry receptor ACE2 which is expressed at high levels in a few cell types of the nasal epithelium, as well as elsewhere (Gordon et al., 2020; Sungnak et al., 2020). This could be an unusual opportunity for volatile inhaled therapeutics and prophylactics that will have direct access to the cells that are infected by the virus.

The Gordon et al study also identified FDA-approved drugs that have known activity against these human protein targets or are structurally related to chemicals with known activity on the targets. While these drugs have yet to be tested directly on the virus, another study performed high-throughput testing of ~12,000 FDA-approved or clinical stage drugs on viral replication in cell lines (Riva et al., 2020). This study identified at least 6 potential leads that include a kinase inhibitor, a CCR1 inhibitor and 4 cysteine protease inhibitors that are candidates for testing in clinical trials.

Since the regulatory process for the approval of new drugs can take several years, the repurposing of FDA approved drugs for COVID-19 offers a potential fast-track to approval. One of the more promising candidates being tested is the antiviral Remdesivir, which has been effective in vitro (Wang et al., 2020) as well as in non-human primates (Williamson et al., 2020), with human trials currently ongoing. The other drug being tested is the antimalarial, hydroxychloroquine, which showed some promise alongside the antibiotic, azithromycin, in small clinical trials (Z. Chen et al., 2020; Gautret et al., 2020). However, hydroxychloroquine has shown less promise in larger trials for treating COVID-19 (Mahevas et al., 2020).

While drug repurposing is expedient, it is possible that drugs designed for other diseases will not be as well suited to respiratory organs, where a large percentage of putative human proteins targeted by the virus are enriched (Gordon et al., 2020), or to the nervous system, implicated by neurological symptoms as well as prior evidence that coronaviruses can cross the blood brain barrier (Li, Bai, & Hashikawa, 2020; Mao et al., 2020). Drug-development strategies are also often guided by minimizing off-target

interactions. Repurposed drugs might have to be used in combination, and the side effects and interactions that this entails are presently not well defined. While there are recent efforts exploring novel, directed therapies from small molecule libraries (Sheahan et al., 2020), it is desirable to identify 100-1000s of putative chemicals as the majority may be difficult to synthesize in mass, prove toxic at therapeutic concentrations, or yield inconsistent benefits across patients due to genetic variability. These shortcomings have significantly increased the demand for additional drugs or small molecules that might interfere with viral entry and replication. Additionally, if prophylactics or non-toxic, easy to use therapeutics were available even for mild cases that do not require hospitalization and experimental drug treatments, it may nevertheless impact long-term health and community transmission (Bagheri et al., 2020).

There are subsequently unmet needs in COVID-19 research, including identification of compounds that target the relevant SARS-CoV-2 human proteins from (1) approved drugs, (2) FDA registered chemicals or (3) a large repository of ~14 million purchasable chemicals from the ZINC 15 database (Sterling & Irwin, 2015), which we computed additional properties for such as mammalian toxicity, vapor pressure, and logP. For 65 human protein targets that SARS-CoV-2 interacts with that had publicly available bioassay and chemical data (Gordon et al., 2020), we first generated a database of predictions based on structural similarity to chemicals that interact with the targets and then machine learning models (34). Many chemicals we have identified have little or no known biological activities and are predicted to have low toxicity in addition to a wide range of vapor pressures. These data are a resource to rapidly identify and test novel, safe

treatment strategies for COVID-19 and other diseases where the target proteins are relevant.

7.2. Results

7.2.1. Identification of important structural features from known inhibitors of human target proteins.

In order to test whether there is a structural basis for inhibitors of the target proteins identified previously (Gordon et al., 2020; Yan et al., 2020), we used two complementary approaches to evaluate each target's training set of compounds with known activity, compiled from the literature. First, we performed an exhaustive search for maximum common substructures among active chemicals. In some cases, enriched substructures were apparent among known ligands, with slight variation in the substructure based on the sensitivity to the targets, suggesting physicochemical features may be relevant in predicting activity against these targets. Next, we used a machine learning pipeline for predicting chemicals that interfere with SARS-CoV-2 targets. It involves selection of important physicochemical features for each target, followed by fitting support vector machines (SVM) with these features and then evaluating the predictions using various computational validation methods (Figure 7.1A). The chemical features that best predicted activity for the different targets included simple 2D information, describing the type and number of bonds, but also more abstract 3D geometries (Tables 7.1-7.2). Identification of each target-specific feature set provides a foundation to better

understand the physicochemical basis of the activity (details about the feature ranking algorithms in Methods).

7.2.2. Machine learning models can successfully predict activity from chemical structure

We identified 24 targets with training sets large enough to model the log IC₅₀, K_i, or AC₅₀ (Figure 7.2A). Rigorous computational validation was performed and the results on training (Figure 7.2B, left) and test data that had been set aside (Figure 2C, left) indicated good overall performance according to the average mean absolute error (MAE) and the correlation between predicted and observed assay measures (MAE = 0.48; R = 0.62). Predictions of log K_i for the viral entry receptor, ACE2, were also accurate (test set R = 0.92; test set mean absolute error (MAE) = 0.53) (Figure 7.2C, left).

For some of the viral targets, we noticed that assay data included additional inhibitory measurements. Some of the available data such as % inhibition, for instance, are less quantitative. However, to include as much of the available data as possible, we created models to identify physicochemical features that might broadly contribute to inhibition. We therefore assigned binary, active and inactive, labels to the chemicals, then trained models as outlined before (Figure 7.2A; Methods). The models that were developed using this classification approach similarly proved successful, validating over partitions of the training data (avg. AUC = 0.87, avg. Shuffle AUC = 0.50, $p < 10^{-19}$) (Figure 7.2B, right), as well as over sets of external test chemicals (avg. AUC = 0.83, avg. Shuffle AUC = 0.51, $p < 10^{-8}$) (Figure 7.2C, right). Collectively, these results

suggested the models provided accurate predictions and could be used to screen approved drug libraries as well as databases of commercially available chemicals for novel therapeutics.

7.2.3. Predicting candidates for repurposing of FDA-approved drugs

Repurposing of existing FDA approved drugs offers a path towards rapid deployment of therapeutics against SARS-CoV-2. Approved drugs may have activity that extend beyond the original target protein. Accordingly, we used the machine learning models to predict activities of ~100,000 FDA registered chemicals (UNII database) as well as the DrugBank (Wishart et al., 2018) and Therapeutic Targets (Chen, 2002; Zhu et al., 2009) databases, which include information on drug interactions, pathways, and approval status. Interestingly, some of the approved drugs are predicted to have high activity against the SARS-CoV-2 targets (Figure 7.3A). In order to identify more efficacious candidates, we isolated the drugs scoring in the top 25 for multiple targets and found a few of high priority (Figure 7.3B).

7.2.4. Predicting inhaled drugs for SARS-CoV2 from FDA-approved and a large ~14M chemical space

Given that many of the human target proteins are overexpressed in the respiratory tract, including the entry receptor ACE2 in only a few cells types of the nasal epithelium, the upper airways and lungs (Gordon et al., 2020; Sungnak et al., 2020), we reasoned that volatile chemicals may offer a unique opportunity as inhaled therapeutics that will have

direct access to the cells and tissues that are infected by the virus. We used the machine learning models to search a large database of ~14 million commercially available chemicals (ZINC) for volatile candidates. We initially isolated the top 1% of the predicted scoring distribution (Figure 7.4A, left), which resulted in > 1 million chemicals in total (Figure 4A, right). To prioritize the hits for potential human use, we next developed machine learning models to predict volatility (vapor pressure) (Figure 7.7) and mammalian toxicity (LD₅₀) (Figure 7.8). The toxicity and vapor pressure estimates helped identify smaller priority sets (Figure 7.4B). Although the vapor pressures were not especially high, we rank ordered the top candidates according to the best values (Figure 7.4C).

Chemicals with suspected odorant properties, however, represent only a fraction of the chemical space, and these chemicals may not have the activity levels suited for COVID-19 cases. Volatile compounds, for instance, may be biased towards structurally simple chemicals that do not resemble drugs. We therefore also focused on additional chemicals with highest predicted activities for their targets and low estimated toxicities regardless of vapor pressure. We identified numerous candidates with potential activity against multiple viral targets (Figure 7.5A) and many other others with significant activity against a single target (Figure 7.6A).

7.2.5. Large-scale toxicity Prediction for chemicals of interest

Although mammalian toxicity estimates are critical in prioritizing candidates for screening, the effect a chemical may have on additional biological processes remains

relevant. For approved drugs, some testing has been carried out using in vivo models, but the breadth of such testing is not exhaustive, and therefore in silico modeling is in demand. Toxicity prediction itself predates modern machine learning, however interest has grown considerably of late, with successful modeling of toxicity using decision trees, support vector machine, and deep neural networks. Similarly, large-scale government initiatives in the United States and European Union suggest a commitment toward virtual or in silico toxicity modeling over laboratory animals. The Tox21 program, operating under the National Institute of Health (NIH), conducts and reports data from high-throughput in vitro assays. To date, the database includes 64 target assays for 10,000+ chemicals. Machine learning studies have successfully modeled 10 of these targets; however, the remaining targets have not been studied extensively with machine learning. It is therefore important to assess the suitability of machine learning for the entire set of targets and develop a comprehensive database outlining the structural motifs and physicochemical properties that are associated with different toxicities.

To start, approaches were evaluated to predict activity on the 10 protein targets that had been previously studied. This ensured that the machine learning could be benchmarked relative to prior efforts. Success here would then justify applying this approach to the more comprehensive set of targets and chemicals. The 10 targets were divided into two broad categories: stress response (SR) and nuclear receptor signaling (NR). The chemicals screened across these 10 in vitro assays were divided into training (8,000) and testing sets (~700) as before. Winning approaches from among 400 previous submissions achieved prediction accuracies, on average, of ~80-85%, which is consistent

with the success rates achieved for this study (Table 7.3; Figure 7.9A). The performance in this study for some targets was ranked among the very top of previous efforts, placing alongside leading performance (out of the 400) for the other protein targets as well (Table 7.3; Figure 7.9A). Accordingly, these results suggested that the combination of binary fingerprints as well as other physicochemical properties adequately captured the structure-activity relationship and this approach would generalize to the protein targets that have not been modeled with machine learning.

Of the proteins that have not been modeled a subset was selected: sonic hedgehog (SHH), a critical developmental regulator that is affected by known teratogenic chemicals; estrogen receptor beta (ERB), an important endocrine regulator; caspase3 (CASP3), a protease that is a key part of the apoptosis cascade; constitutive androstane receptor (CAR), which mediates the response to xenobiotic chemicals; androgen receptor (AR), which, similar to estrogen receptors, plays a key role in endocrine regulation; the thyroid stimulating hormone receptor (TSHR), responsible for regulating thyroid function; lastly, retinoic acid receptor alpha (RAR), a heterodimer that participates in epigenetic regulation, particularly through stimulating deacetylation; the biochemical reaction that enhances chromatin packing, thereby repressing gene expression (<https://ncats.nih.gov/tox21>). These protein targets are then further characterized according to the type of activity that the in vitro screen was designed to detect (agonist vs antagonist) and the cell line (e.g. Kidney Cell, HEK293; Chinese Hamster Ovary, CHO, Murine Embryo fibroblast) (Table 7.4). Although the chemicals screened differ by target, the composition of training and testing compounds that for the 10 previously modeled

protein targets was kept as consistent as possible. When applying the same approach as Figure 7.9A and as reported in Table 7.3, prediction of the test chemicals was successful, at a rate that compared to the 10 previously modeled protein targets (Figure 7.9A), with an average accuracy of 81% (AUC = .81) (Figure 7.9B). This suggests that in silico modeling based on physicochemical properties can be applied to comprehensively evaluate undesirable off target toxicities of known drugs and small molecules that have potential as novel therapeutics.

7.3. Discussion

SARS-CoV-2 is a significant world health crisis. The full scope of COVID-19 disease and any long-term health complications following infection remain unclear. Although vaccines are the best long-term solution, treatments will be necessary to mitigate disease severity in the short term. What is concerning is that, while several repurposed drugs have already been tested in some form of clinical trial, and only one drug Remdesivir has shown a clear benefit in randomized clinical trials. Additionally, there is no guarantee that an effective vaccine can be found for the SARS-CoV2 virus, and therefore drug candidate pipelines are extremely important to pursue for the long-term research effort against COVID-19. A vaccine against SARS-CoV-2 would likely need to stimulate local immunity, since the infection is limited to mucosal surfaces, and these could be short-lived immunities.

We have therefore taken a comprehensive approach to try and provide a pipeline for short and long-term use, and for a potentially local application route via inhalation.

Existing FDA approved drugs that target a single protein important for viral replication and host entry are currently the highest priority for repurposing as new COVID-19 drugs. However, we think that there are compelling reasons to create pipelines to explore many putative targets, and chemical spaces that are far larger and more diverse than the known approved drugs. We have therefore screened ~14 million potentially purchasable compounds from the ZINC database and also predicted toxicity values for the numerous candidates. In addition, we have identified chemicals that are predicted to affect more than one of the host proteins, suggesting these may have more efficacy. One unusual category we have emphasized is volatiles, as these compounds may be biologically sourced, and therefore microbes could be genetically engineered to produce them in mass (Hug, Krug, & Müller, 2020). This would subsequently reduce the strain on global supply chains for chemicals that are necessary in synthesizing certain pharmaceuticals. These chemicals are also intriguing options for drug cocktails. If present in metabolic pathways, they possibly already interact *in vivo*. Therefore, short-term therapeutic concentrations may be better tolerated in humans.

It is nevertheless important to note that machine learning depends on available data. Because the size and diversity of publicly available bioassay data are limited, caution is required in interpreting the predictions. It is common to find past bioassays focused on similar shaped chemicals, limiting the scope of the machine learning approach to find new chemistries. Importantly, apart from ACE2, the other human proteins that were identified to interact with SARS-CoV-2 are yet to be tested *in vivo* for drug-ability. And although some of the candidate chemicals we identified may be biologically

sourced, the concentrations are not well defined or unknown, nor is there any understanding of a therapeutic concentration in this scenario. These data are presented as a forward-looking resource and a pipeline to evaluate chemical data with additional research. While our motivation was the evolving COVID-19 pandemic, the 64 SARS-CoV-2 targets are relevant to a range of other diseases and conditions. We therefore anticipate that the AI-based predictions of purchasable compounds from 10+ million chemicals will accelerate drug discovery in general and facilitate research on these chemicals in the future for a number of diseases. In general, the use of AI-driven tools could provide additional valuable solutions for tackling Covid-19 (Santosh, 2020).

7.4. Figures

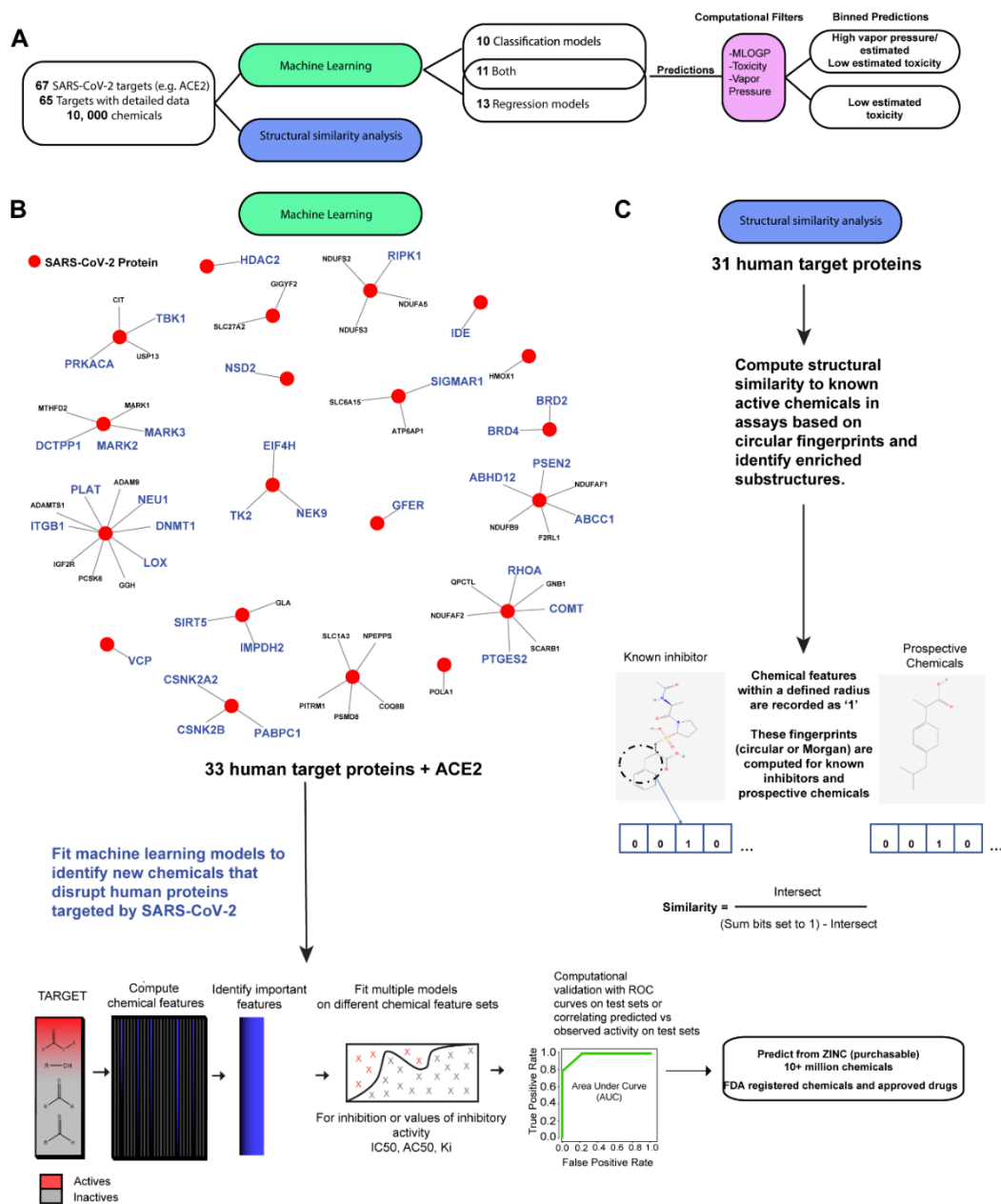


Figure 7.1. Machine learning pipeline to identify chemicals that interfere with SARS-CoV-2 targets. **A)** Overview of the pipeline to predict chemicals for 65 SARS-CoV-2 human targets selected from Gordon et al., 2020 and using bioassay data from publicly available databases. **B)** Graphically depicts the pipeline details. Available bioassay data on the viral targets were mined for information to use in machine learning or structural analysis. This resulted in 24 targets that could be modeled using values for the most abundant inhibitory assay measure (e.g. K_i or IC_{50}) and 21 targets modeled by classifying broad inhibition (34 unique targets in total). The remaining targets with limited data were funneled into a structural similarity analysis, which aids in developing more bioassay data and helps clarify the chemical features contributing to bioactivity. For targets modeled with supervised machine learning, optimal chemical features were identified on subsets of training data. The top features were sampled by support vector machines (SVM). These models were then aggregated. External chemicals were used to verify successful predictions. Models trained for the 34 targets predicted large chemical databases including FDA registered chemicals and approved drugs, as well as 10+ million purchasable chemicals from the ZINC database. Top scoring predicted chemicals were subsequently assigned theoretical toxicity, log vapor pressure, and MLOGP, which estimates membrane permeability.

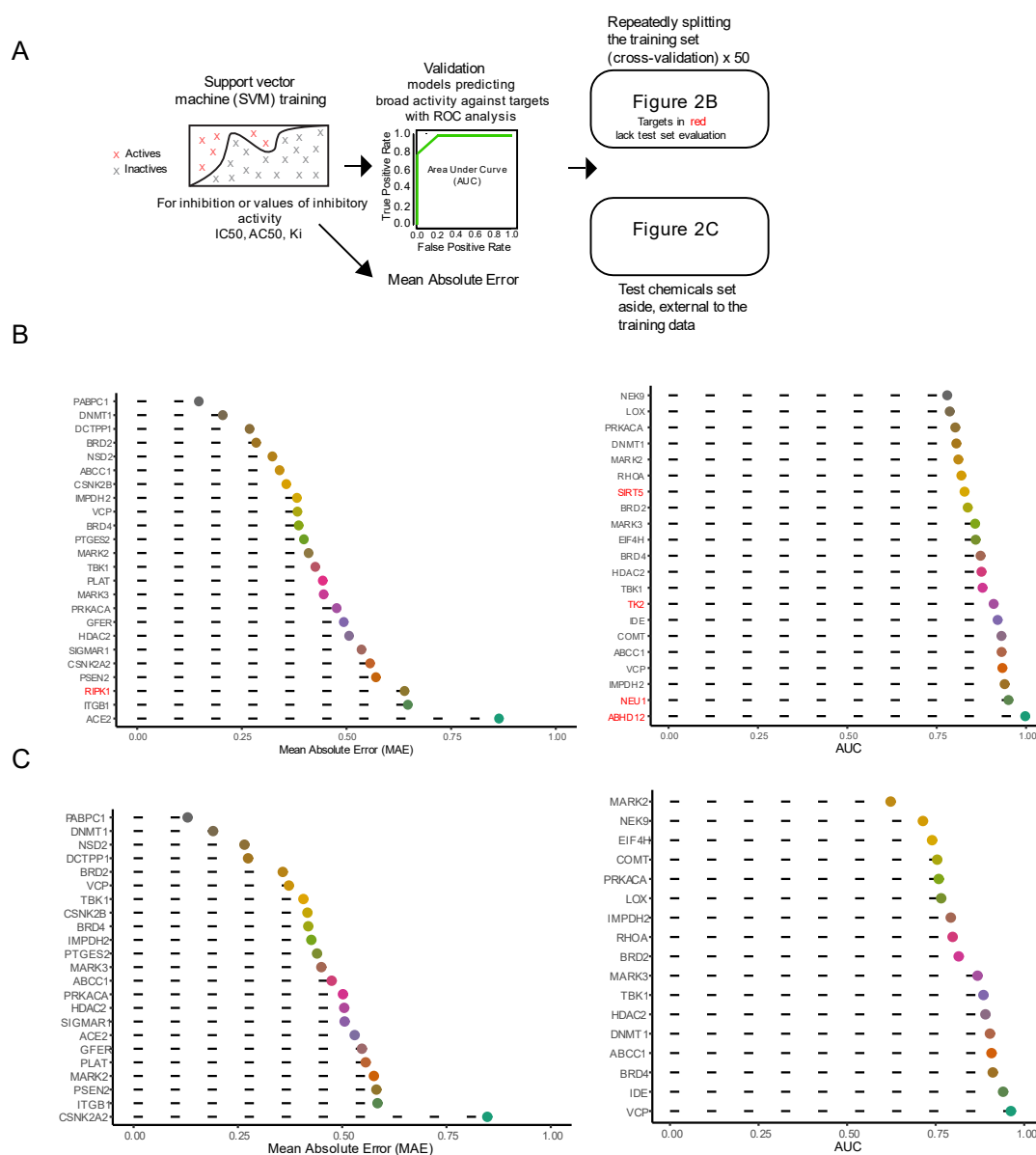


















Figure 7.2. Models of chemical features accurately predict inhibitors of SARS-CoV-2 targets. **A)** Pipeline for fitting and validating models that predict IC_{50} , K_i , or AC_{50} or a classification score, which reflects broad inhibitory activity against the listed viral targets. **B) Left**, mean absolute error (MAE) in predicting the log transformed endpoints (IC_{50} , K_i , AC_{50}). **Right**, classification of broadly inhibiting chemicals using the area under the receiver operating characteristic (ROC) curve (AUC). Plots are for 10-fold cross validation, repeated 5 times. The model predictions are from an ensemble of three support vector machines (SVM), trained on different chemical feature sets. **C) Left**, external test set performance for regression models, where possible. **Right**, external test set performance for classification models, where possible.

A

Image	Viral Target	Chemical	Name	Category	Score	Unit
	HDAC2	D00VUL	Phenazopyridine	approved	0.9650111	Inhibition
	IDE	DB12001	Abemaciclib	approved; investigational	0.9873756	Inhibition
	MARK3	D00NAX	Promazine	approved	0.9765537	Inhibition
	IMPDH2	D0F0ZE	Tyverb/Tykerb	approved	0.9591730	Inhibition
	ABCC1	DB00670	Pirenzepine	approved	0.9752358	Inhibition
	ABHD12	DB11742	Ebastine	approved; investigational	0.9782785	Inhibition
	VCP	D0U3SY	Alectinib	approved	0.9879945	Inhibition
	BRD2	D0V9WF	Lestaurtinib	approved	23.0047892	nM
	BRD4	D0E7PQ	Vorinostat	approved	15.3853689	nM
	CSNK2A2	D06QCC	Cefmenoxime	approved	1.4576031	nM
	ITGB1	DB11611	Lifitegrast	approved	3.9707880	nM
	PSEN2	D0Y9EW	Vemurafenib	approved	4.9246898	nM
	RIPK1	DB11942	Selinexor	approved; investigational	0.6120112	nM
	SIGMAR1	DB09056	Amorolfine	approved; investigational	0.8099480	nM
	TBK1	DB11963	Dacomitinib	approved; investigational	73.5593947	nM
	ACE2	D0N5HJ	Enalaprilat	approved	0.3465582	nM

B

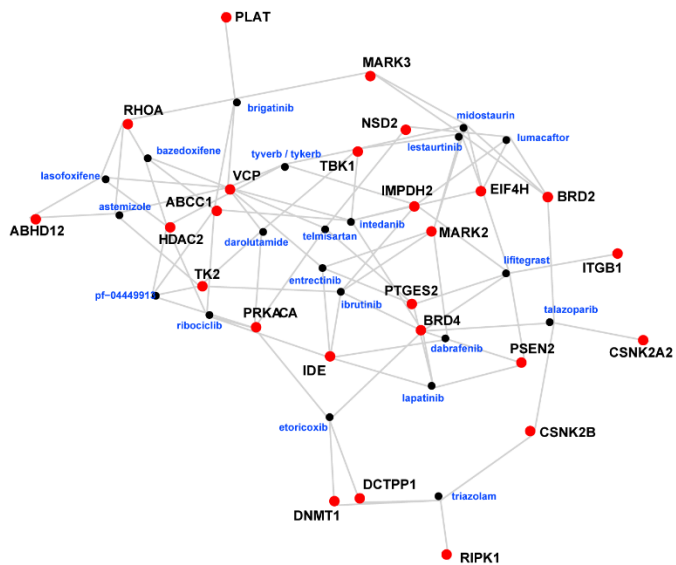
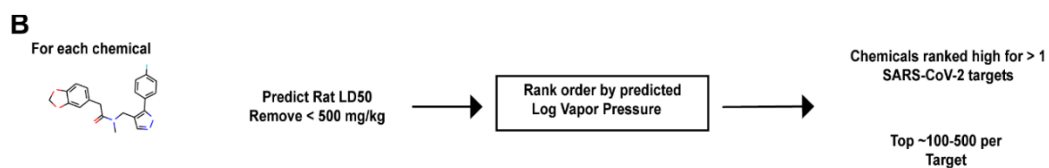
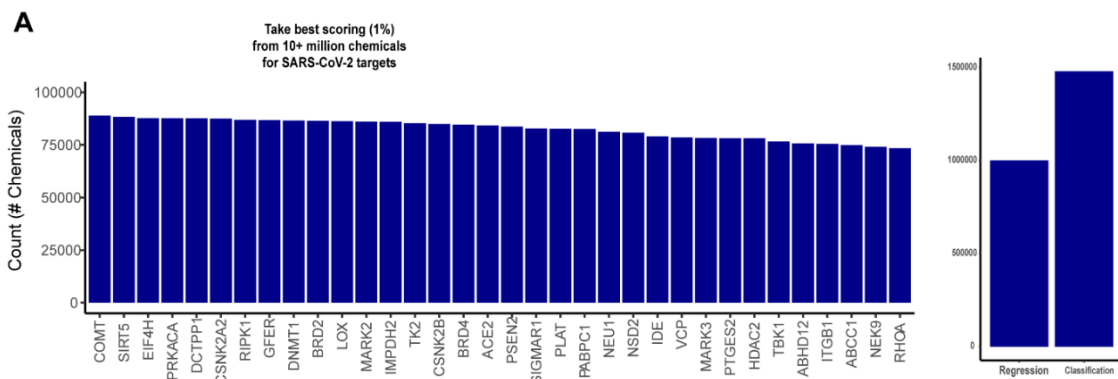


Figure 3

Figure 7.3. Approved drugs with putative activity against SARS-CoV-2 targets. **A)** The best predicted activity against SARS-CoV-2 targets among databases of approved drugs. Viral targets with few promising candidates are omitted. **B)** Network showing drugs that are among the top 25 for multiple viral targets (drugs: black nodes; viral targets: red nodes).



C

Image	Target	Name	Score	Model Type	Predicted Assay Value (nM)	Image	Target	Name	Score	Model Type	Predicted Assay Value (nM)
	ABCC1	ZINC000585123546	0.9847671	IC50	1610.039		LOX	ZINC000040268534	0.9519342	Classification	
	ABHD12	ZINC000019213470	0.9663231	Classification			MARK2	ZINC000044839375	0.8824998	Ki	3230.616
	BRD2	ZINC000013550759	0.7967906	IC50	97.785		MARK3	ZINC000013069764	0.9927328	Ki	5880.926
	BRD4	ZINC000021861821	0.9466124	IC50	80.382		NEK9	ZINC000041149819	0.9674333	Classification	
	COMT	ZINC000169617863	0.9485490	Classification			NEU1	ZINC000012788748	0.7745192	Classification	
	DNMT1	ZINC000072229485	0.9533028	IC50	4328.601		PRKACA	ZINC000000862475	0.8994358	Ki	707.695
	EIF4H	ZINC000002646755	0.8946055	Classification			RHOA	ZINC0000409017970	0.9332619	Classification	
	HDAC2	ZINC000065337200	0.9850594	IC50	4600.933		SIRT5	ZINC000013101063	0.8004927	Classification	
	IDE	ZINC000008609677	0.9984797	Classification			TBK1	ZINC000072125546	0.9699165	IC50	435.007
	IMPDH2	ZINC000004479471	0.9599488	IC50	223.031		VCP	ZINC000035508424	0.9870381	IC50	2071.873
							ACE2	ZINC000000000582		Ki	148.104

Figure 4

Figure 7.4. Predicting activity against SARS-CoV-2 targets among theoretical volatile chemicals. **A) Left**, count of chemicals per target after initially filtering based on predicted scores. **Right**, chemical counts across all viral targets for the models predicting general inhibitory scores (**Classification**) and those for specific inhibitory endpoints (**Regression**) (e.g. IC50). **B)** Pipeline for further prioritizing chemical sets according to estimated vapor pressure and low mammalian toxicity (LD50). **C)** Top ranking predictions of general inhibitory activity (**Score**) and/or specific inhibitory endpoints (**Predicted Assay Value**) against SARS-CoV-2 targets from the ZINC database, filtered to the highest estimated log vapor pressures.

A

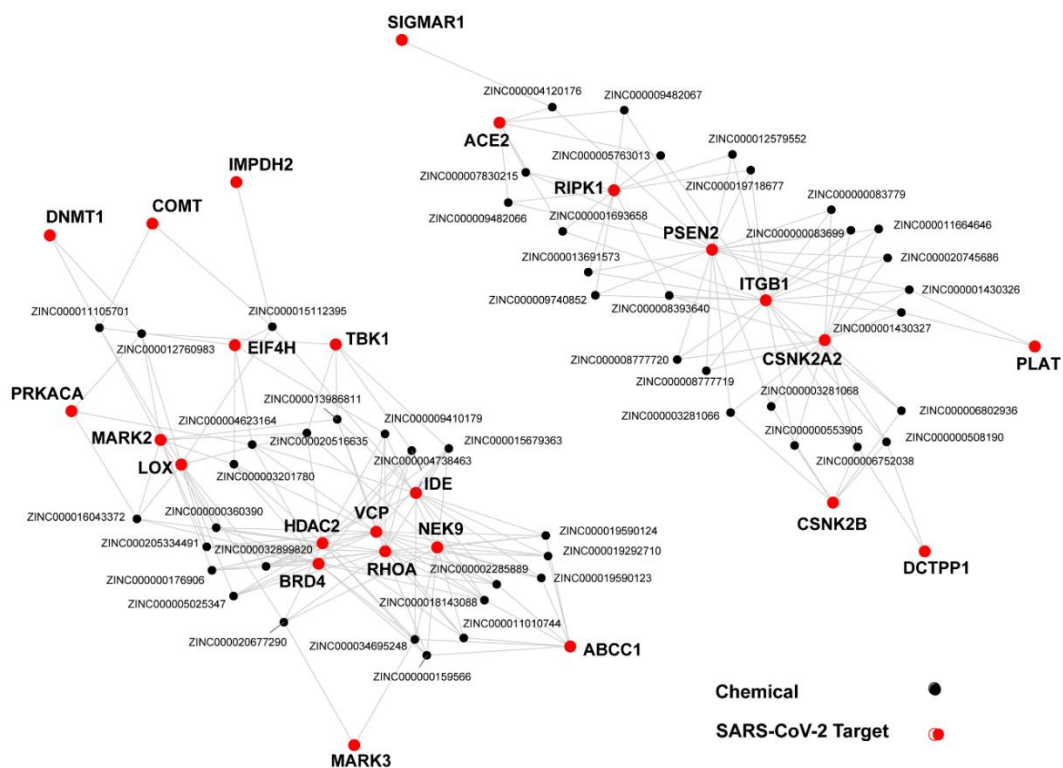


Figure 7.5. Predicted chemicals rank highly for multiple SARS-CoV-2 targets. a) Network of chemicals predicted to have low toxicity that are ranked highly for > 1 viral targets. Chemicals were considered if for multiple viral targets they had > 0.75 inhibitory/class scores or predictions of specific assay measures (K_i , IC_{50} , and AC_{50}) < 100 nM.

A



Image	Target	Chemical	Score	Image	Target	Chemical	Score
	ACE2	ZINC000409143350	0.0364 nM		COMT	ZINC000001230197	0.957286399188097
	CSNK2A2	ZINC000096310808	0.8832 nM		DNMT1	ZINC000004377187	0.9703539081114
	CSNK2B	ZINC000000808028	48.4502 nM		EIF4H	ZINC000016194037	0.933971718889293
	DCTPP1	ZINC000100267962	54.7004 nM		HDAC2	ZINC000020725405	0.992614222252123
	GFER	ZINC000067280191	198.7122 nM		IDE	ZINC000004946116	0.990204127520776
	ITGB1	ZINC000245230693	3.7941 nM		IMPDH2	ZINC000013117452	0.986379822884044
	NSD2	ZINC000004705927	388.8352 nM		LOX	ZINC000013371505	0.97968369430634
	PABPC1	ZINC000096940647	2242.6084 nM		MARK2	ZINC000006548568	0.899013031939661
	PLAT	ZINC000225825359	90.6871 nM		MARK3	ZINC000024471026	0.997455774977146
	PSEN2	ZINC000085393386	0.9381 nM		NEK9	ZINC000023888283	0.977367919972458
	PTGES2	ZINC000023010530	121.9883 nM		NEU1	ZINC000001753421	0.876816009226939
	RIPK1	ZINC000014200189	0.028 nM		PRKACA	ZINC000012630194	0.944270504944003
	SIGMAR1	ZINC000000285272	0.4209 nM		RHOA	ZINC000004865708	0.951689646254862
	ABCC1	ZINC000020150907	0.988250396044501		SIRT5	ZINC000241231017	0.837287352871655
	ABHD12	ZINC000072356259	0.978944577164529		TBK1	ZINC000025249236	0.990012021863476
	BRD2	ZINC000101526703	0.859767323801036		TK2	ZINC000055132982	0.770208109403797
	BRD4	ZINC000027757620	0.981262503922496		VCP	ZINC000019375342	0.992581942027781

Figure 7.6. Predictions of SARS-CoV-2 targets among chemicals lacking odorant properties. A) Sample of ZINC chemicals scoring highly for inhibitory activity against the viral targets.

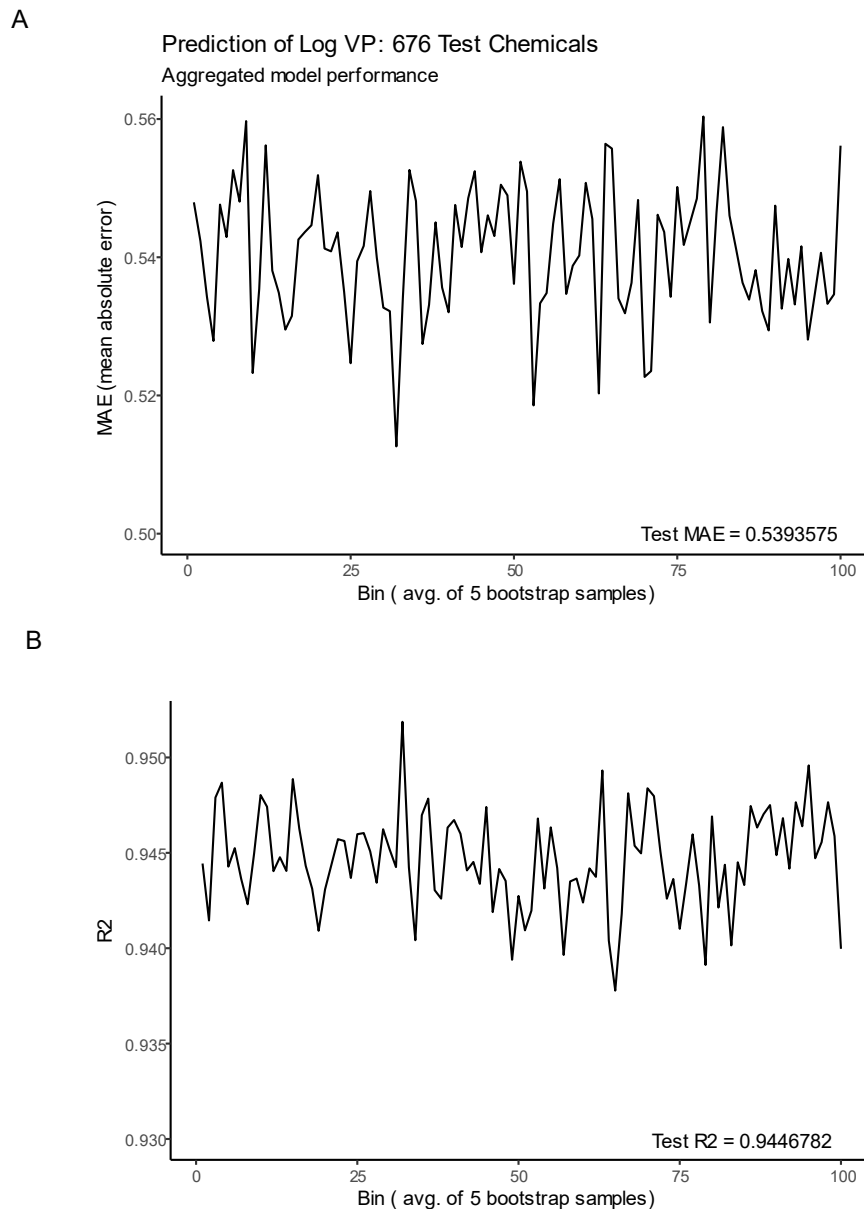


Figure 7.7. Machine learning can successfully model vapor pressure. A) Ensemble model for predicting log vapor pressure is validated on 676 test chemicals. Test set predictions are bootstrapped 500 times, averaged over 100 bins (5 bootstrap samples per bin). Predictive success is quantified as the mean absolute error (MAE); average in plot area. **B)** The test chemical predictions are assessed using the R2 value, bootstrapped 500 times and averaged over 100 bins (5 bootstrap samples per bin). Overall R2 value reported in the plot area. Individual models are trained on different chemical feature sets and predictions are aggregated.

A

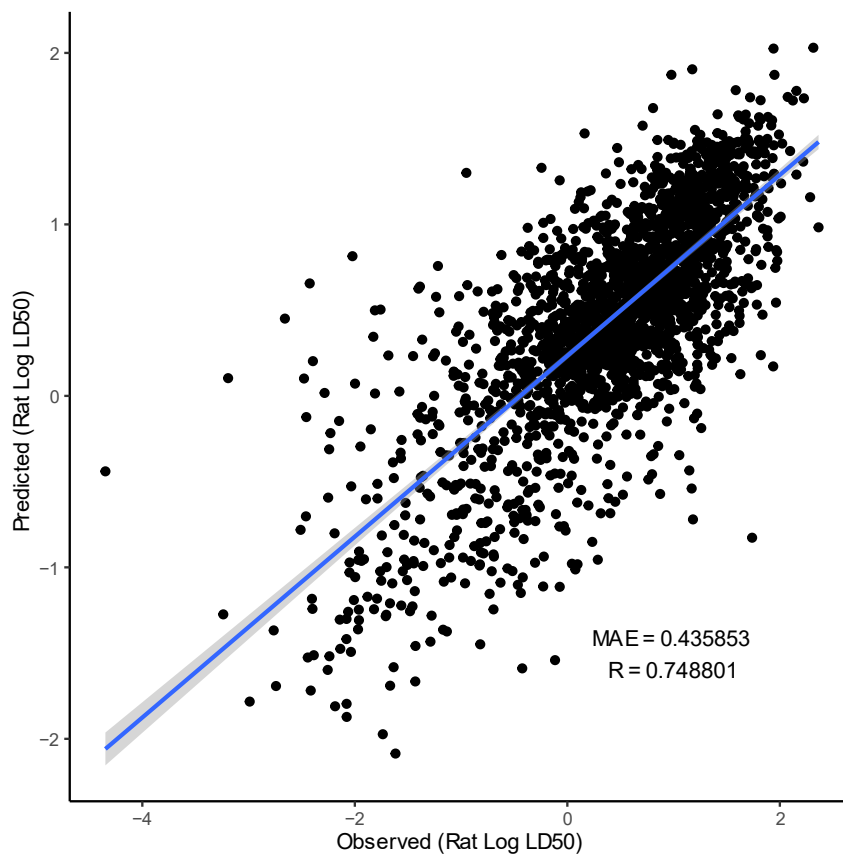
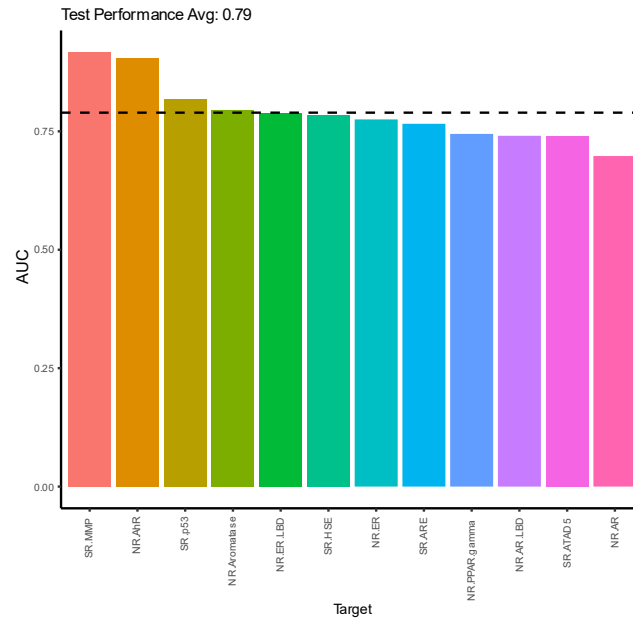


Figure 7.8. Machine learning can accurately predict the toxicity of chemicals. A) Ensemble model prediction of rat log LD₅₀ for 2895 test chemicals. Relationship between predicted and observed log LD₅₀ is quantified as the correlation. Value reported in plot area. The Mean absolute Error (MAE) in the prediction of these test chemicals is also reported. Blue line is the least squares approximation of the plotted (x, y) pairs (black dots); the thin transparent, gray band surrounding the blue line indicates the error in the fit.

A



B

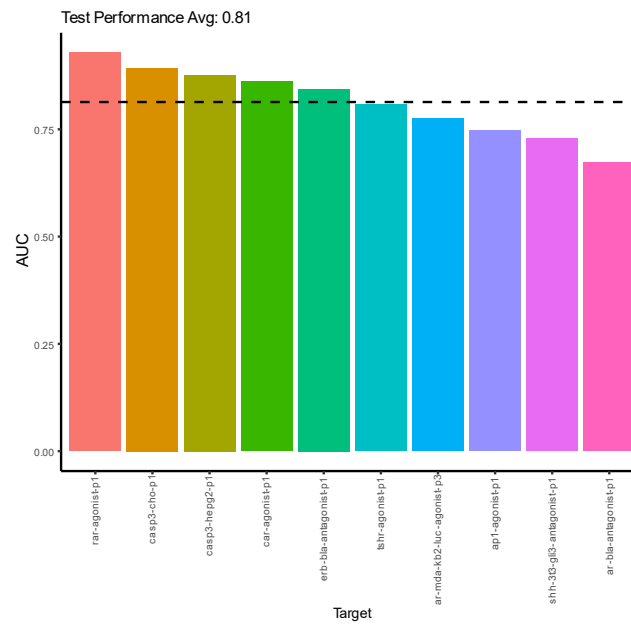


Figure 9

Figure 7.9. Additional toxicity endpoints are also accurately modeled by machine learning. A) Test set performance for 10 well studied toxicity endpoints. These proteins were studied in the Tox21 competition, which provided competitors with common set of chemicals to train their algorithms and then test them to evaluate success. The approach taken here represents the chemicals by binary fingerprints and other physicochemical properties; these features are then used to train a regularized random forest (RRF) as well as a support vector machine with a radial basis function kernel (RBF SVM) (Methods). Predictions of the test chemicals from these two algorithms are subsequently aggregated. The aggregated prediction is evaluated using ROC (Receiver Operator Characteristic) analysis, which defines success according to the area under the ROC curve (AUC). Best performance 1.0, which implies perfect sensitivity (true positive rate) and specificity (1-false positive rate). b) The same approach is next applied to a set of protein targets that were not among the 10 previously modeled with machine learning. As before, machine learning models based on physicochemical features accurately predict most of the toxicity targets.

7.5. Tables

Table 7.1. Top chemical features for regression models

Feature	Target	Description
GATS5s	ABCC1	Geary autocorrelation of lag 5 weighted by I-state
RDF055m	ABCC1	Radial Distribution Function - 055 / weighted by mass
SpMax_B(s)	ABCC1	leading eigenvalue from Burden matrix weighted by I-State
CATS2D_08_AA	BRD2	CATS2D Acceptor-Acceptor at lag 08
RDF035s	BRD2	Radial Distribution Function - 035 / weighted by I-state
SpDiam_X	BRD2	spectral diameter from chi matrix
HATS8p	BRD4	leverage-weighted autocorrelation of lag 8 / weighted by polarizability
R5i+	BRD4	R maximal autocorrelation of lag 5 / weighted by ionization potential
RDF035m	BRD4	Radial Distribution Function - 035 / weighted by mass
Eig02_EA(bo)	CSNK2A2	eigenvalue n. 2 from edge adjacency mat. weighted by bond order
Eig05_EA(bo)	CSNK2A2	eigenvalue n. 5 from edge adjacency mat. weighted by bond order
SpMax2_Bh(m)	CSNK2A2	largest eigenvalue n. 2 of Burden matrix weighted by mass
CATS2D_04_AA	CSNK2B	CATS2D Acceptor-Acceptor at lag 04
SHED_DN	CSNK2B	SHED Donor-Negative
SpMin1_Bh(m)	CSNK2B	smallest eigenvalue n. 1 of Burden matrix weighted by mass
DISPm	DCTPP1	displacement value / weighted by mass
HATS7u	DCTPP1	leverage-weighted autocorrelation of lag 7 / unweighted
Mor31s	DCTPP1	signal 31 / weighted by I-state
MATS1e	DNMT1	Moran autocorrelation of lag 1 weighted by Sanderson electronegativity
Mor23m	DNMT1	signal 23 / weighted by mass
TDB06u	DNMT1	3D Topological distance based descriptors - lag 6 unweighted
GATS4m	GFER	Geary autocorrelation of lag 4 weighted by mass
Mor14m	GFER	signal 14 / weighted by mass
R5i	GFER	R autocorrelation of lag 5 / weighted by ionization potential
DISPp	HDAC2	displacement value / weighted by polarizability
IC2	HDAC2	Information Content index (neighborhood symmetry of 2-order)
P_VSA_MR_5	HDAC2	P_VSA-like on Molar Refractivity, bin 5
F04[C-C]	IMPDH2	Frequency of C - C at topological distance 4
HOMA	IMPDH2	Harmonic Oscillator Model of Aromaticity index
VE1_B(s)	IMPDH2	coefficient sum of the last eigenvector (absolute values) from Burden matrix weighted by I-State
Eig02_AEA(dm)	ITGB1	eigenvalue n. 2 from augmented edge adjacency mat. weighted by dipole moment
SHED_AA	ITGB1	SHED Acceptor-Acceptor
SpMax2_Bh(s)	ITGB1	largest eigenvalue n. 2 of Burden matrix weighted by I-state
F10[C-N]	MARK2	Frequency of C - N at topological distance 10
nPyrroles	MARK2	number of Pyrroles
SaaNH	MARK2	Sum of aaNH E-states

max_conj_path	MARK3	maximum number of atoms that can be in conjugation with each other
SaaNH	MARK3	Sum of aaNH E-states
VE1_H2	MARK3	coefficient sum of the last eigenvector (absolute values) from reciprocal squared distance matrix
GATS3s	NSD2	Geary autocorrelation of lag 3 weighted by I-state
HOMA	NSD2	Harmonic Oscillator Model of Aromaticity index
Mor16s	NSD2	signal 16 / weighted by I-state
H7m	PABPC1	H autocorrelation of lag 7 / weighted by mass
JGI7	PABPC1	mean topological charge index of order 7
P_VSA_MR_2	PABPC1	P_VSA-like on Molar Refractivity, bin 2
GATS4m	PLAT	Geary autocorrelation of lag 4 weighted by mass
Mor04s	PLAT	signal 04 / weighted by I-state
R6p+	PLAT	R maximal autocorrelation of lag 6 / weighted by polarizability
nPyrroles	PRKACA	number of Pyrroles
RDF040v	PRKACA	Radial Distribution Function - 040 / weighted by van der Waals volume
SpMin3_Bh(m)	PRKACA	smallest eigenvalue n. 3 of Burden matrix weighted by mass
Eig02_EA(bo)	PSEN2	eigenvalue n. 2 from edge adjacency mat. weighted by bond order
nArX	PSEN2	number of X on aromatic ring
VE1sign_D/Dt	PSEN2	coefficient sum of the last eigenvector from distance/detour matrix
SHED_DL	PTGES2	SHED Donor-Lipophilic
VE2sign_G	PTGES2	average coefficient of the last eigenvector from geometrical matrix
VE3sign_G	PTGES2	logarithmic coefficient sum of the last eigenvector from geometrical matrix
CATS3D_08_AL	RIPK1	CATS3D Acceptor-Lipophilic BIN 08 (8.000 - 9.000 Å)
MAT55i	RIPK1	Moran autocorrelation of lag 5 weighted by ionization potential
VE3sign_RG	RIPK1	logarithmic coefficient sum of the last eigenvector from reciprocal squared geometrical matrix
BLTA96	SIGMAR1	Verhaar Algae base-line toxicity from MLOGP (mmol/l)
F10[C-C]	SIGMAR1	Frequency of C - C at topological distance 10
TPSA(Tot)	SIGMAR1	topological polar surface area using N,O,S,P polar contributions
Eig01_AEA(dm)	TBK1	eigenvalue n. 1 from augmented edge adjacency mat. weighted by dipole moment
HATS4i	TBK1	leverage-weighted autocorrelation of lag 4 / weighted by ionization potential
SdssC	TBK1	Sum of dssC E-states
AROM	VCP	aromaticity index
E1m	VCP	1st component accessibility directional WHIM index / weighted by mass
MATS5m	VCP	Moran autocorrelation of lag 5 weighted by mass
H5s	ACE2	H autocorrelation of lag 5 / weighted by I-state
Mor10m	ACE2	signal 10 / weighted by mass
Mor17m	ACE2	signal 17 / weighted by mass

Table 7.1. Important chemical features for regression models. Top three chemical features for the viral targets with K_i , IC_{50} , and AC_{50} bioassay activities.

Table 7.2. Top chemical features for classification models

Feature	Target	Description
Mor18s	BRD4	signal 18 / weighted by I-state
SpMAD_G/D	BRD4	spectral mean absolute deviation from distance/distance matrix
SpMax3_Bh(p)	BRD4	largest eigenvalue n. 3 of Burden matrix weighted by polarizability
P_VSA_LogP_3	HDAC2	P_VSA-like on LogP, bin 3
SHED_DA	HDAC2	SHED Donor-Acceptor
SHED_DL	HDAC2	SHED Donor-Lipophilic
G(N..N)	IDE	sum of geometrical distances between N..N
SM1_Dz(i)	IDE	spectral moment of order 1 from Barysz matrix weighted by ionization potential
Wap	IDE	all-path Wiener index
CATS2D_08_DA	TBK1	CATS2D Donor-Acceptor at lag 08
F08[N-N]	TBK1	Frequency of N - N at topological distance 8
P_VSA_e_3	TBK1	P_VSA-like on Sanderson electronegativity, bin 3
H7m	PRKACA	H autocorrelation of lag 7 / weighted by mass
H7s	PRKACA	H autocorrelation of lag 7 / weighted by I-state
RDF060m	PRKACA	Radial Distribution Function - 060 / weighted by mass
GATS6e	MARK3	Geary autocorrelation of lag 6 weighted by Sanderson electronegativity
GATS6m	MARK3	Geary autocorrelation of lag 6 weighted by mass
Mor02m	MARK3	signal 02 / weighted by mass
CATS2D_02_DL	IMPDH2	CATS2D Donor-Lipophilic at lag 02
CATS3D_07_DL	IMPDH2	CATS3D Donor-Lipophilic BIN 07 (7.000 - 8.000 Å)
NaasC	IMPDH2	Number of atoms of type aasC
C-039	ABCC1	Ar-C(=X)-R
VE2sign_Dz(p)	ABCC1	average coefficient of the last eigenvector from Barysz matrix weighted by polarizability
VE3sign_Dz(v)	ABCC1	logarithmic coefficient sum of the last eigenvector from Barysz matrix weighted by van der Waals volume
Mor31s	ABHD12	signal 31 / weighted by I-state
RTi+	ABHD12	R maximal index / weighted by ionization potential
VE3sign_Dz(p)	ABHD12	logarithmic coefficient sum of the last eigenvector from Barysz matrix weighted by polarizability
E2m	BRD2	2nd component accessibility directional WHIM index / weighted by mass
GATS2m	BRD2	Geary autocorrelation of lag 2 weighted by mass
TDB03i	BRD2	3D Topological distance based descriptors - lag 3 weighted by ionization potential
MAXDP	COMT	maximal electrotopological positive variation
nDB	COMT	number of double bonds
P_VSA_MR_2	COMT	P_VSA-like on Molar Refractivity, bin 2
CATS2D_02_AL	DNMT1	CATS2D Acceptor-Lipophilic at lag 02
Mor04s	DNMT1	signal 04 / weighted by I-state
VE3sign_Dt	DNMT1	logarithmic coefficient sum of the last eigenvector from detour matrix
ChiA_B(i)	EIF4H	average Randic-like index from Burden matrix weighted by ionization potential
F05[C-O]	EIF4H	Frequency of C - O at topological distance 5
NaasC	EIF4H	Number of atoms of type aasC
CENT	LOX	centralization
EE_G	LOX	Estrada-like index (log function) from geometrical matrix
VE2_D/Dt	LOX	average coefficient of the last eigenvector (absolute values) from distance/detour matrix
Eta_D_beta	MARK2	eta measure of electronic features
Mor29v	MARK2	signal 29 / weighted by van der Waals volume
SpPosA_B(i)	MARK2	normalized spectral positive sum from Burden matrix weighted by ionization potential
CATS2D_07_AL	NEK9	CATS2D Acceptor-Lipophilic at lag 07
CATS2D_08_AL	NEK9	CATS2D Acceptor-Lipophilic at lag 08
TDB05p	NEK9	3D Topological distance based descriptors - lag 5 weighted by polarizability
CATS2D_06_DL	NEU1	CATS2D Donor-Lipophilic at lag 06

TDB04i	NEU1	3D Topological distance based descriptors - lag 4 weighted by ionization potential
X3A	NEU1	average connectivity index of order 3
nR06	RHOA	number of 6-membered rings
R8s+	RHOA	R maximal autocorrelation of lag 8 / weighted by I-state
SpMin1_Bh(m)	RHOA	smallest eigenvalue n. 1 of Burden matrix weighted by mass
CATS3D_08_NL	SIRT5	CATS3D Negative-Lipophilic BIN 08 (8.000 - 9.000 Å)
O-057	SIRT5	phenol, enol, carboxyl OH
SpMax2_Bh(s)	SIRT5	largest eigenvalue n. 2 of Burden matrix weighted by I-state
CATS2D_04_AL	TK2	CATS2D Acceptor-Lipophilic at lag 04
JGI3	TK2	mean topological charge index of order 3
MATS1i	TK2	Moran autocorrelation of lag 1 weighted by ionization potential
P_VSA_e_3	VCP	P_VSA-like on Sanderson electronegativity, bin 3
RDF020p	VCP	Radial Distribution Function - 020 / weighted by polarizability
SpMaxA_AEA(dm)	VCP	normalized leading eigenvalue from augmented edge adjacency mat. weighted by dipole moment

Table 7.2. Important chemical features for classification models. Top three chemical features for viral targets where the models classified chemicals as active vs inactive relative to broad inhibition rather than a specific assay value (e.g. K_i , IC_{50} , and AC_{50}).

Table 7.3

Teams	NR.Ah R	NR.A R	NR.AR.L BD	NR.Aromat ase	NR.E R	NR.ER.L BD	NR.PPAR.ga mma	SR.A RE	SR.ATA D5	SR.H SE	SR.M MP	SR.p 53
DeepT ox	0.928	0.807	0.879	0.834	0.81	0.814	0.861	0.84	0.793	0.865	0.942	0.862
AMA- ZIZ T	0.913	0.77	0.846	0.819	0.806	0.806	0.83	0.805	0.828	0.842	0.95	0.843
	0.913	0.676	0.848	0.825	0.784	0.805	0.822	0.801	0.814	0.811	0.937	0.847
This Study	0.903	0.701	0.747	0.79	0.78	0.794	0.751	0.758	0.733	0.784	0.913	0.817
Micro- somes Charite	0.901				0.785	0.827	0.717	0.804	0.812			0.826
	0.896	0.688	0.789	0.781	0.707	0.798	0.7	0.739	0.751	0.852	0.88	0.834
fillips- PL RCC	0.893	0.736	0.743	0.776	0.771			0.758		0.766	0.928	0.815
	0.872	0.763	0.747	0.792	0.781	0.762	0.637	0.761	0.673	0.755	0.92	0.795
MML	0.871	0.693	0.66	0.709	0.75	0.71	0.645	0.701	0.749	0.647	0.854	0.815
CGL	0.866	0.742	0.566	0.749	0.759	0.727	0.738	0.747	0.737	0.775	0.88	0.817
Froze- narm kibutz	0.865	0.744	0.722	0.74	0.745	0.79	0.803	0.7	0.726	0.752	0.859	0.803
	0.865	0.75	0.694	0.729	0.757	0.779	0.666	0.708	0.737	0.587	0.838	0.787
ToxFit	0.862	0.744	0.757	0.738	0.729	0.752	0.791	0.697	0.729	0.689	0.862	0.803
Super- Tox VIF	0.854		0.56	0.742				0.711			0.862	0.732
	0.827	0.797	0.61	0.671	0.732	0.735	0.666	0.636	0.656	0.723	0.796	0.648
NCI	0.812	0.628	0.592	0.698	0.483	0.703	0.736	0.783	0.714	0.858	0.851	0.747
dmlab	0.781	0.828	0.819	0.838	0.766	0.772	0.831	0.768	0.8	0.855	0.946	0.88
Toxic Avg	0.715	0.721	0.611	0.671	0.646	0.64	0.682	0.633	0.593	0.465	0.732	0.614
Swam- idass	0.353	0.571	0.748	0.274	0.68	0.738	0.585	0.372	0.391	0.711	0.828	0.661

Table 7.3. Test performance for this study compared to the top prior efforts from 400 total. Test set chemicals that the machine learning model has not seen provide a true evaluation of the prediction success. Compared to 400 previous efforts to predict the same test chemicals for the 10 toxicity endpoints the current study places among the best of these efforts.

Table 7.4

Protocol Name	Assay Target	Target Category	Cell Line	Cell Type
tox21-ahr-p1	AhR	NR	HepG2	Liver
tox21-ap1-agonist-p1	AP-1 agonist	SR	ME-180	Cervical Cancer
tox21-ar-bla-agonist-p1	AR-BLA agonist	NR	HEK293	Kidney
tox21-ar-bla-antagonist-p1	AR-BLA antagonist	NR	HEK293	Kidney
tox21-ar-mda-kb2-luc-agonist-p1	AR-MDA agonist	NR	MDA-MB-453	Breast Cancer
tox21-ar-mda-kb2-luc-agonist-p3	AR-MDA agonist (with antagonist)	NR	MDA-MB-453	Breast Cancer
tox21-ar-mda-kb2-luc-antagonist-p1	AR-MDA antagonist	NR	MDA-MB-453	Breast Cancer
tox21-ar-mda-kb2-luc-antagonist-p2	AR-MDA antagonist (lower agonist)	NR	MDA-MB-453	Breast Cancer
tox21-are-bla-p1	ARE	SR	HepG2	Liver
tox21-aromatase-p1	Aromatase	SR	MCF-7	Breast Cancer
tox21-car-agonist-p1	CAR agonist	NR	HepG2	Liver
tox21-car-antagonist-p1	CAR antagonist	NR	HepG2	Liver
tox21-casp3-cho-p1	Caspase-3/7	Cytotoxicity	CHO	Hamster
tox21-casp3-hepg2-p1	Caspase-3/7	Cytotoxicity	HepG2	Liver
tox21-dt40-p1	Cell viability	Gene Tox	DT40	Chicken
tox21-elg1-luc-agonist-p1	ATAD5	Gene Tox	HEK293	Kidney
tox21-er-bla-agonist-p2	ER-BLA agonist	NR	HEK293	Kidney
tox21-er-bla-antagonist-p1	ER-BLA antagonist	NR	HEK293	Kidney
tox21-er-luc-bg1-4e2-agonist-p2	ER-BG1 agonist	NR	BG1	Ovarian
tox21-er-luc-bg1-4e2-agonist-p4	ER-BG1 agonist (with antagonist)	NR	BG1	Ovarian
tox21-er-luc-bg1-4e2-antagonist-p1	ER-BG1 antagonist	NR	BG1	Ovarian
tox21-er-luc-bg1-4e2-antagonist-p2	ER-BG1 antagonist (lower agonist)	NR	BG1	Ovarian
tox21-erb-bla-antagonist-p1	ER-beta antagonist	NR	HEK293	Kidney
tox21-erb-bla-p1	ER-beta agonist	NR	HEK293	Kidney
tox21-err-p1	ERR	NR	HEK293	Kidney
tox21-esre-bla-p1	ER stress	SR	HeLa	Cervical Cancer
tox21-fxr-bla-agonist-p2	FXR-BLA agonist	NR	HEK293	Kidney
tox21-fxr-bla-antagoist-p1	FXR-BLA antagonist	NR	HEK293	Kidney
tox21-gh3-tre-agonist-p1	TR-beta agonist	NR	GH3	Rat pituitary
tox21-gh3-tre-antagonist-p1	TR-beta antagonist	NR	GH3	Rat pituitary
tox21-gr-hela-bla-agonist-p1	GR-BLA agonist	NR	HeLa	Cervical Cancer
tox21-gr-hela-bla-antagonist-p1	GR-BLA antagonist	NR	HeLa	Cervical Cancer
tox21-h2ax-cho-p2	H2AX	Gene Tox	CHO	Hamster

Protocol Name	Assay Target	Target Category	Cell Line	Cell Type
tox21-hdac-p1	HDAC	Gene Tox	HCT-116	Colon Cancer
tox21-hre-bla-agonist-p1	HRE-BLA agonist	SR	ME-180	Cervical Cancer
tox21-hse-bla-p1	HSE-BLA	SR	HeLa	Cervical Cancer
tox21-luc-biochem-p1	Luciferase, biochemical	Counter Screen	N/A	Biochemical
tox21-mitotox-p1	Mitochondria toxicity	SR	HepG2	Liver
tox21-nfkb-bla-agonist-p1	NFkB agonist	SR	ME-180	Cervical Cancer
tox21-p53-bla-p1	P53	Gene Tox	HCT-116	Colon Cancer
tox21-pgc-err-p1	PGC-ERR	NR	HEK293	Kidney
tox21-ppard-bla-agonist-p1	PPAR-delta-BLA agonist	NR	HEK293	Kidney
tox21-ppard-bla-antagonist-p1	PPAR-delta-BLA antagonist	NR	HEK293	Kidney
tox21-pparg-bla-agonist-p1	PPAR-gamma agonist	NR	HEK293	Kidney
tox21-pparg-bla-antagonist-p1	PPAR-gamma antagonist	NR	HEK293	Kidney
tox21-pr-bla-agonist-p1	PR-BLA agonist	NR	HEK293	Kidney
tox21-pr-bla-antagonist-p1	PR-BLA antagonist	NR	HEK293	Kidney
tox21-pxr-p1	PXR agonist	NR	HepG2	Liver
tox21-rar-agonist-p1	RAR agonist	NR	C3H10T1/2	Murine embryo fibroblast
tox21-rar-antagonist-p2	RAR antagonist	NR	C3H10T1/2	Murine embryo fibroblast
tox21-rar-viability-p2	RAR viability	Cytotoxicity	C3H10T1/2	Murine embryo fibroblast
tox21-ror-cho-antagonist-p1	ROR antagonist	NR	CHO	Hamster
tox21-ror-cho-viability-p1	ROR viability	Cytotoxicity	CHO	Hamster
tox21-rt-viability-hek293-p1	Cell viability	Cytotoxicity	HEK293	Kidney
tox21-rt-viability-hepg2-p1	Cell viability	Cytotoxicity	HepG2	Liver
tox21-rxr-bla-agonist-p1	RXR-BLA	NR	HEK293	Kidney
tox21-sbe-bla-agonist-p1	SBE-BLA (TGF-beta) agonist	Developmental Toxicity	HEK293	Kidney
tox21-sbe-bla-antagonist-p1	SBE-BLA (TGF-beta) antagonist	Developmental Toxicity	HEK293	Kidney
tox21-shh-3t3-gli3-agonist-p1	Hedgehog agonist	Developmental Toxicity	NIH/3T3	Murine embryo fibroblast
tox21-shh-3t3-gli3-antagonist-p1	Hedgehog antagonist	Developmental Toxicity	NIH/3T3	Murine embryo fibroblast
tox21-spec-hek293-p1	Auto fluorescence	Counter Screen	HEK293	Kidney
tox21-spec-hepg2-p1	Auto fluorescence	Counter Screen	HepG2	Liver
tox21-trhr-hek293-p1	TRHR agonist and antagonist	GPCR	HEK293	Kidney
tox21-tshr-agonist-p1	TSHR agonist	GPCR	HEK293	Kidney
tox21-tshr-antagonist-p1	TSHR antagonist	GPCR	HEK293	Kidney
tox21-tshr-wt-p1	TSHR wild type	GPCR	HEK293	Kidney
tox21-vdr-bla-agonist-p1	VDR-BLA agonist	NR	HEK293	Kidney
tox21-vdr-bla-antagonist-p1	VDR-BLA antagonist	NR	HEK293	Kidney

Table 7.4. Details for the Tox21 assays. The terminology for the Tox21 assays is clarified. The protocol name is the shorthand descriptor that used throughout to refer to characteristics of the assay. The subsequent columns provide the detail behind the shorthand or abbreviated description. Tox21 competition protein targets highlighted in yellow. NR: Nuclear response; SR: stress response

7.6. Methods

7.6.1. Data Sources for machine learning

7.6.1.2. ZINC

ZINC is a free database comprised of 230 million chemicals for in silico analyses. It was developed as a resource for non-commercial research. Chemicals predicted here are from a purchasable subset; however, availability is subject to change and pricing may vary widely (Sterling & Irwin, 2015).

7.6.1.3. Bioassay data

Bioassay data was retrieved from ChEMBL 25 using the associated Python module, which enables access to the API services via Python (EMBL-EBI, 2011; Mendez et al., 2019). The various inhibitory measures/endpoints, wherever possible, are standardized to nM units; the logarithm of the standardized values was used for machine learning. Regression models were fit for a single endpoint. For classification machine learning models, however, ‘active’ class chemicals were defined using the activity comments, endpoints with values up to 10,000 nM (K_i and IC_{50}) and for the semi-quantitative % inhibition, greater than 10%. The majority class was downsampled during the training and model tuning phases to adjust for possible class imbalances. Training for the regression and classification approaches was done on 85% of the total data. Notably, in a small number of cases the remaining 15% was insufficient to effectively estimate performance using an external test set. To reduce bias, feature selection (recursive feature elimination (RFE) algorithm) was always run on 85% of the data over 250-300 different

partitions (iteratively running the 10-fold cross validation 25-30 times). However, for these cases, the held-out portion (15%) was then incorporated back into the dataset to better estimate performance of the trained model by 10-fold cross-validation (repeated 5 times). We also fit 3 different radial basis function (RBF) support vector machine (SVM) models, wherein the chemical features (predictors) were randomly sampled (50%) from the top 70. This makes the performance estimates more conservative.

7.6.1.3. Toxicity data

Training and testing data are curated by various government agencies and provided freely to the general public as databases (see Key Resources Table) (Fonger, Hakkinen, Jordan, & Publicker, 2014; Kinsner-Ovaskainen et al., 2009; Richard & Williams, 2002).

7.6.1.3. Vapor Pressure data

Training and testing data are from EPI Suite (EPA, 2015), which is developed and maintained by the Environmental Protection Agency (EPA). Methods for fitting these models are as outlined in the Figure 1 pipeline. To compare the vapor pressure model predictions with respect to different machine learning methods as well as EPI suite, data were split into train/test partitions as defined in a previous study (Zang et al., 2017).

7.6.2. Selecting optimally predictive chemical features

7.6.2.1, Optimizing chemical structures

Chemical features were computed with ~5300 AlvaDesc descriptors, from the developers of DRAGON software, and 3D coordinates and optimization performed using RDKit in Python (Landrum, 2006).

7.6.2.2 Chemical feature ranking and importance

Recursive feature elimination iteratively selects subsets of features to identify optimal sets. The algorithm is a “wrapper” and therefore relies on an additional algorithm to supply predictions and quantify importance. We used two different algorithms, depending on the size and composition of data: (1) Random Forest and (2) Support Vector Machine (SVM). Random forest determines the importance in relation to the % increase in error when permuting a feature or predictor. There is no equivalent method for computing importance with the SVM. Accordingly, the importance is based on fitting a model between the response and each predictor or feature as compared to null. If the response is numeric, importance is derived from the pseudo R^2 (non-linear regression). If, however, the response is binary, the AUC is instead computed for each predictor or feature (see Key Resources Table for algorithm source files).

Including cross-validation with the recursive feature elimination (RFE) partitions the training data into multiple folds. This step avoids biasing performance estimates but results in lists of top predictors over the cross-validation folds such that importance of a predictor is based on a selection rate.

7.6.2.2.1 Selection Bias

Selecting features or predictors on the same dataset used for cross validation results in models that have already “seen” possible partitions of the data and therefore performance metrics will be biased. Selection bias (Ambroise & McLachlan, 2002) was addressed by bootstrapping and cross validation, which ensure some separation between predictor/feature selection and model-fitting/validation. In addition to these methods, we used hidden test sets.

7.6.3. Selecting optimal machine learning algorithms

The support vector machine (SVM) with the radial basis function kernel (RBF) outperformed regularized Random Forest (regRF) or performed comparably. Rather than utilize many different approaches, we aggregated multiple SVM models to improve generalizability. However, in the case of the classification model for EIF4H, we included the regularized random forest algorithm, as the aggregated prediction (SVM and regRF) was clearly optimal on the test data. Algorithm selection and training was done using the classification and regression training package in R (R Development Core Team, 2016), caret (Kuhn, 2008), and the implementation of the Support Vector Machine (SVM) algorithm in Kernlab (Karatzoglou et al., 2004).

7.6.4. Enriched Substructures/Cores

Enriched cores were analyzed using RDKit through Python (Landrum, 2006). The algorithm performs an exhaustive search for maximum a common substructure among a set of chemicals. In practice, larger sets often yield fewer substantive cores. To remedy this, the algorithm includes a threshold parameter that relaxes the proportion of chemicals

containing the core. We used a threshold of 0.55, which ensures that the majority of the chemicals contained the core.

7.6.5. Chemical Fingerprinting

Extended Connectivity Fingerprints (ECFP) are a class of cheminformatic algorithms that iteratively combine chemical features that are present within a predefined radius/diameter, representing them by set of integer values. Typically, the fingerprint is converted into a binary string of fixed length using a hash function. Here, the bit length was set at 1024 and a radius of 2 (diameter = 4 or ECFP4). This structural representation was preferred as it is strongly associated with activity (Rogers & Hahn, 2010){Formatting Citation}. Accordingly, it is a suitable alternative to identify drug candidates in the absence of machine learning models. We used the ECFP algorithm in RDKit (Morgan or circular fingerprint) (Landrum, 2006; Morgan, 1965). The similarity between the fingerprints of chemicals with known activity against the SARS-CoV-2 targets and prospective chemicals was computed using the Tanimoto index. This index is a similarity coefficient (0-1; 1 = max similarity). It is the overlap of the “on-bits” divided by the sum of the unique “on-bits. Notably, coefficients of 1 need not imply identical chemicals.

$$sim(AB) = \frac{c}{a+b-c} \text{ where } c = \text{overlapping "on-bits"}; a = \text{"on bits" in } A; b = \text{"on-bits" in } B$$

7.6.6. Support Vector Machine (SVM)

Training the support vector machine (SVM) involves identifying a set of parameters that optimize a cost function, where cost 1 and cost 0 correspond to training chemicals labeled as “Active” and “Inactive,” respectively. θ^T is the scoring function or output of the support vector machine. If the output is ≥ 0 , the prediction is “Active.” The function (f) is a kernel function.

$$SVM \text{ Cost} = \min_{\theta} C \sum_{i=1}^m y^{(i)} \text{cost}_1(\theta^T f^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T f^{(i)}) + \frac{1}{2} \sum_{j=1}^n \theta_j^2$$

The kernel determines the shape of the decision boundary between the active and inactive chemicals from the training set. The radial basis function (RBF) or Gaussian kernel enables the learning of more complex, non-linear boundaries. It is therefore well suited for problems in which the biologically active chemicals cannot be properly classified as a linear function of physicochemical properties. This kernel computes the similarity for each chemical (x) and a set of landmarks (l), where σ^2 is a tunable parameter determined by the problem and data. The similarity with respect to these landmarks is used to predict new chemicals (“Active” vs. “Inactive”).

$$\text{Gaussian Kernel} = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$$

7.6.7. Model Performance Metrics

The Area under the ROC Curve (AUC) assesses the true positive rate (TPR or sensitivity) as a function of the false positive rate (FPR or 1-specificity) while varying the probability threshold (T) for a label (Active/Inactive). If the computed probability score (x) is greater

than the threshold (T), the observation is assigned to the active class. Integrating the curve provides an estimate of classifier performance, with the top left corner giving an AUC of 1.0 denoting maximum sensitivity to detect all targets or actives in the data without any false positives. The theoretical random classifier is reported at $AUC = 0.5$.

$$TPR(T) = \int_T^{\infty} f_1(x) dx$$

$$FPR(T) = \int_T^{\infty} f_0(x) dx$$

Where T is a variable threshold and x is a probability score.

However, we generated classifiers that are more authentic than theoretical random classification, shuffling the chemical feature values in the models and statistically comparing the mean AUCs across multiple partitions of the data. This controls against optimally tuned algorithms predicting well simply because of specific predictor attributes (e.g. range, mean, median, and variance) or models that are of a specific size (number of predictors) performing well even with shuffled values. Additionally, biological data sets are often small, with stimuli or chemicals that—rather than random selection—reflect research biases, possibly leading to optimistic validation estimates without the proper controls.

We used the AUC for evaluating classification models. For the classification-based training, we initially converted the inhibitory data into a binary label (Active/Inactive). For predictions of quantitative bioassay measures (e.g. K_i , IC_{50} , AC_{50} , Log LD_{50}), we computed the mean absolute error (MAE), the correlation coefficient (R) and the squared correlation coefficient (R²). **MAE:** Mean absolute error is the mean of

the absolute difference between predicted and observed (% usage). It therefore assigns equal weight to all prediction errors, whether large or small.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y - \hat{y}|; \text{ where, } \hat{y} = \text{predicted and } y = \text{observed}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}; \text{ where, } TP = \text{True Positive and } FN = \text{False Negative}$$

$$\text{Specificity} = \frac{TN}{TN+FP}; \text{ where, } TN = \text{True Negative and } FP = \text{False Positive}$$

References

- Acree, F., Turner, R. B., Gouck, H. K., Beroza, M., & Smith, N. (1968). l-lactic acid: A mosquito attractant isolated from humans. *Science*.
<https://doi.org/10.1126/science.161.3848.1346>
- Adipietro, K. A., Mainland, J. D., & Matsunami, H. (2012). Functional evolution of mammalian odorant receptors. *PLoS Genet*.
<https://doi.org/10.1371/journal.pgen.1002821>
- Ahmed, L., Zhang, Y., Block, E., Buehl, M., Corr, M. J., Cormanich, R. A., ... Zhuang, H. (2018). Molecular mechanism of activation of human musk receptors OR5AN1 and OR1A1 by (R)-muscone and diverse other musk-smelling compounds. *Proceedings of the National Academy of Sciences*. Retrieved from
<http://www.pnas.org/content/early/2018/04/05/1713026115.abstract>
- Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), 6562–6566.
<https://doi.org/10.1073/pnas.102102699>
- Badel, L., Ohta, K., Tsuchimoto, Y., & Kazama, H. (2017). Decoding of Context-Dependent Olfactory Behavior in *Drosophila*. *Neuron*, 91(1), 155–167.
<https://doi.org/10.1016/j.neuron.2016.05.022>
- Bagheri, S. H. R., Asghari, A. M., Farhadi, M., Shamshiri, A. R., Kabir, A., Kamrava, S. K., ... Salimi, A. (2020). Coincidence of COVID-19 epidemic and olfactory dysfunction outbreak. *MedRxiv*. <https://doi.org/10.1101/2020.03.23.20041889>
- Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D. Y., Chen, L., & Wang, M. (2020). Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA - Journal of the American Medical Association*. <https://doi.org/10.1001/jama.2020.2565>
- Bak, J. H., Jang, S. J., & Hyeon, C. (2019). Modular structure of human olfactory receptor codes reflects the bases of odor perception. *BioRxiv*, 525287.
<https://doi.org/10.1101/525287>
- Bell, J. S., & Wilson, R. I. (2016). Behavior Reveals Selective Summation and Max Pooling among Olfactory Processing Channels. *Neuron*, 91(2), 425–438.
<https://doi.org/http://dx.doi.org/10.1016/j.neuron.2016.06.011>
- Besansky, N. J., Hill, C. A., & Costantini, C. (2004). No accounting for taste: Host preference in malaria vectors. *Trends in Parasitology*.
<https://doi.org/10.1016/j.pt.2004.03.007>
- Bewick, S., Gurarie, E., Weissman, J. L., Beattie, J., Davati, C., Flint, R., ... Fagan, W. F. (2019). Trait-based analysis of the human skin microbiome. *Microbiome*.
<https://doi.org/10.1186/s40168-019-0698-2>

- Boyle, S. M., Guda, T., Pham, C. K., Tharadra, S. K., Dahanukar, A., & Ray, A. (2016). Natural DEET substitutes that are strong olfactory repellents of mosquitoes and flies. In *bioRxiv*. <https://doi.org/10.1101/060178>
- Boyle, S. M., McInally, S., Tharadra, S., & Ray, A. (2016). Short-term memory trace mediated by termination kinetics of olfactory receptor. *Scientific Reports*, 6. <https://doi.org/10.1038/srep19863>
- Braks, M. A. H., & Takken, W. (1999). Incubated human sweat but not fresh sweat attracts the malaria mosquito *Anopheles gambiae sensu stricto*. *Journal of Chemical Ecology*. <https://doi.org/10.1023/A:1020970307748>
- Braun, T., Voland, P., Kunz, L., Prinz, C., & Gratzl, M. (2007). Enterochromaffin Cells of the Human Gut: Sensors for Spices and Odorants. *Gastroenterology*. <https://doi.org/10.1053/j.gastro.2007.02.036>
- Buck L and Axel, R. (1991). A Novel Multigene Family May Encode Odorant Receptors: A Molecular Basis. *Cell*, 65, 175–187.
- Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*. <https://doi.org/10.1038/nrn2575>
- Cao, Y., Charisi, A., Cheng, L. C., Jiang, T., & Girke, T. (2008). ChemmineR: A compound mining framework for R. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btn307>
- Cardé, R. T., & Willis, M. A. (2008). Navigational strategies used by insects to find distant, wind-borne sources of odor. *Journal of Chemical Ecology*. <https://doi.org/10.1007/s10886-008-9484-5>
- Carey, A. F., Wang, G., Su, C. Y., Zwiebel, L. J., & Carlson, J. R. (2010). Odorant reception in the malaria mosquito *Anopheles gambiae*. *Nature*. <https://doi.org/10.1038/nature08834>
- Castro, J. B., Ramanathan, A., & Chennubhotla, C. S. (2013). Categorical Dimensions of Human Odor Descriptor Space Revealed by Non-Negative Matrix Factorization. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0073289>
- Charlier, L., Topin, J., Ronin, C., Kim, S. K., Goddard, W. A., Efremov, R., & Golebiowski, J. (2012). How broadly tuned olfactory receptors equally recognize their agonists. Human OR1G1 as a test case. *Cellular and Molecular Life Sciences*. <https://doi.org/10.1007/s00018-012-1116-0>
- Chen, C.-F. F., Zou, D.-J., Altomare, C. G., Xu, L., Greer, C. A., & Firestein, S. J. (2014). Nonsensory target-dependent organization of piriform cortex. *Proceedings of the National Academy of Sciences*, 111(47), 16931–16936. <https://doi.org/10.1073/pnas.1411266111>

- Chen, X. (2002). TTD: Therapeutic Target Database. *Nucleic Acids Research*.
<https://doi.org/10.1093/nar/30.1.412>
- Chen, Z., Hu, J., Zhang, Z., Jiang, S., Han, S., Yan, D., ... Zhang, Z. (2020). Efficacy of hydroxychloroquine in patients with COVID-19: results of a randomized clinical trial. *MedRxiv*. <https://doi.org/10.1101/2020.03.22.20040758>
- Chihara, T., Kitabayashi, A., Morimoto, M., Takeuchi, K. ichi, Masuyama, K., Tonoki, A., ... Miura, M. (2014). Caspase Inhibition in Select Olfactory Neurons Restores Innate Attraction Behavior in Aged *Drosophila*. *PLoS Genetics*.
<https://doi.org/10.1371/journal.pgen.1004437>
- Consonni, V., Todeschini, R., & Pavan, M. (2002). Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *Journal of Chemical Information and Computer Sciences*, 42(3), 682–692. <https://doi.org/10.1021/ci015504a>
- Cook, B. L., Steuerwald, D., Kaiser, L., Graveland-Bikker, J., Vanberghem, M., Berke, A. P., ... Zhang, S. (2009). Large-scale production and study of a synthetic G protein-coupled receptor: Human olfactory receptor 17-4. *Proceedings of the National Academy of Sciences*. <https://doi.org/10.1073/pnas.0811089106>
- Cooperband, M. F., & Cardé, R. T. (2006). Orientation of *Culex* mosquitoes to carbon dioxide-baited traps: Flight manoeuvres and trapping efficiency. *Medical and Veterinary Entomology*. <https://doi.org/10.1111/j.1365-2915.2006.00613.x>
- Cork, A., & Park, K. C. (1996). Identification of electrophysiologically-active compounds for the malaria mosquito, *Anopheles gambiae*, in human sweat extracts. *Medical and Veterinary Entomology*. <https://doi.org/10.1111/j.1365-2915.1996.tb00742.x>
- Coutinho-Abreu, I. V., Sharma, K., Cui, L., Yan, G., & Ray, A. (2019). Odorant ligands for the CO₂ receptor in two *Anopheles* vectors of malaria. *Scientific Reports*.
<https://doi.org/10.1038/s41598-019-39099-0>
- Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*. <https://doi.org/citeulike-article-id:3443126>
- De Lacy Costello, B., Amann, A., Al-Kateb, H., Flynn, C., Filipiak, W., Khalid, T., ... Ratcliffe, N. M. (2014). A review of the volatiles from the healthy human body. *Journal of Breath Research*. <https://doi.org/10.1088/1752-7155/8/1/014001>
- de March, C. A., Titlow, W. B., Sengoku, T., Breheny, P., Matsunami, H., & McClintock, T. S. (2020). Modulation of the combinatorial code of odorant receptor response patterns in odorant mixtures. *Molecular and Cellular Neurosciences*, 104, 103469. <https://doi.org/10.1016/j.mcn.2020.103469>
- Debnath, T., Prasetyawan, D., & Nakamoto, T. (2019, September 5). *Prediction of Odor*

- Descriptor Group of Essential Oils from Mass Spectra using Machine Learning*. 1–3. <https://doi.org/10.1109/isoen.2019.8823226>
- Dekker, T., Geier, M., & Cardé, R. T. (2005). Carbon dioxide instantly sensitizes female yellow fever mosquitoes to human skin odours. *Journal of Experimental Biology*. <https://doi.org/10.1242/jeb.01736>
- Dekker, T., Takken, W., & Braks, M. A. H. (2001). Innate preference for host-odor blends modulates degree of anthropophagy of *Anopheles gambiae* sensu lato (Diptera: Culicidae). *Journal of Medical Entomology*. <https://doi.org/10.1603/0022-2585-38.6.868>
- Dennis, E. J., Dobosiewicz, M., Jin, X., Duvall, L. B., Hartman, P. S., Bargmann, C. I., & Vosshall, L. B. (2018). A natural variant and engineered mutation in a GPCR promote DEET resistance in *C. elegans*. *Nature*. <https://doi.org/10.1038/s41586-018-0546-8>
- Dewan, A., Cichy, A., Zhang, J., Miguel, K., Feinstein, P., Rinberg, D., & Bozza, T. (2018). Single olfactory receptors set odor detection thresholds. *BioRxiv*. Retrieved from <http://biorxiv.org/content/early/2018/06/07/341099.abstract>
- Dravnieks, A. (1985). *Atlas of Odor Character Profiles*. <https://doi.org/10.1520/DS61-EB>
- Dubois, D., & Rouby, C. (2002). Names and Categories for Odors: The Veridical Label. In C. Rouby, B. Schaal, D. Dubois, R. Gervais, & A. Holley (Eds.), *Olfaction, Taste, and Cognition* (pp. 47–66). <https://doi.org/10.1017/CBO9780511546389.009>
- Dweck, H. K. M., Ebrahim, S. A. M., Kromann, S., Bown, D., Hillbur, Y., Sachse, S., ... Stensmyr, M. C. (2013). Olfactory preference for egg laying on citrus substrates in *Drosophila*. *Current Biology*. <https://doi.org/10.1016/j.cub.2013.10.047>
- EMBL-EBI. (2011). ChEMBL. *ChEMBL*.
- EPA, U. S. (2015). Estimation Programs Interface Suite™ for Microsoft® Windows. *United States Environmental Protection Agency, Washington, DC, USA*.
- Fonger, G. C., Hakkinen, P., Jordan, S., & Publicker, S. (2014). The National Library of Medicine's (NLM) Hazardous Substances Data Bank (HSDB): Background, recent enhancements and future plans. *Toxicology*. <https://doi.org/10.1016/j.tox.2014.09.003>
- Fujita, Y., Takahashi, T., Suzuki, A., Kawashima, K., Nara, F., & Koishi, R. (2007). Deorphanization of dresden G protein-coupled receptor for an odorant receptor. *Journal of Receptors and Signal Transduction*. <https://doi.org/10.1080/10799890701534180>
- Gautret, P., Lagier, J.-C., Parola, P., Hoang, V. T., Meddeb, L., Mailhe, M., ... Raoult, D. (2020). Hydroxychloroquine and azithromycin as a treatment of COVID-19: results

- of an open-label non-randomized clinical trial. *International Journal of Antimicrobial Agents*. <https://doi.org/10.1016/j.ijantimicag.2020.105949>
- Geithe, C., Noe, F., Kreissl, J., & Krautwurst, D. (2017). The Broadly Tuned Odorant Receptor OR1A1 is Highly Selective for 3-Methyl-2,4-nonanedione, a Key Food Odorant in Aged Wines, Tea, and Other Foods. *Chemical Senses*. <https://doi.org/10.1093/chemse/bjw117>
- Gonzalez-Kristeller, D. C., do Nascimento, J. B. P., Galante, P. A. F., & Malnic, B. (2015). Identification of agonists for a group of human odorant receptors. *Frontiers in Pharmacology*. <https://doi.org/10.3389/fphar.2015.00035>
- Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., ... Krogan, N. J. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*. <https://doi.org/10.1038/s41586-020-2286-9>
- Grant, A. J., Aghajanian, J. G., O'Connell, R. J., & Wigton, B. E. (1995). Electrophysiological responses of receptor neurons in mosquito maxillary palp sensilla to carbon dioxide. *Journal of Comparative Physiology A*. <https://doi.org/10.1007/BF00187475>
- Grant, G. G., & Hall, A. C. (2019). Interactions of N, N- diethyl-meta-toluamide (DEET) and Novel Insect Repellents with Mammalian GABAA and Glycine Receptors. *The FASEB Journal*, 33(1_supplement), 813.13-813.13. https://doi.org/10.1096/fasebj.2019.33.1_supplement.813.13
- Guha, R. (2007). Chemical informatics functionality in R. *Journal of Statistical Software*. <https://doi.org/10.18637/jss.v018.i05>
- Gupta, R. K., & Bhattacharjee, A. (2006). *Discovery and Design of New Arthropod/Insect Repellents by Computer-Aided Molecular Modeling*. <https://doi.org/10.1201/9781420006650.ch10>
- Gutiérrez, E. D., Dhurandhar, A., Keller, A., Meyer, P., & Cecchi, G. A. (2018). Predicting natural language descriptions of mono-molecular odorants. *Nature Communications*. <https://doi.org/10.1038/s41467-018-07439-9>
- Haddad, R., Medhanie, A., Roth, Y., Harel, D., & Sobel, N. (2010). Predicting odor pleasantness with an electronic nose. *PLoS Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1000740>
- Hallem, E. A., & Carlson, J. R. (2006). Coding of Odors by a Receptor Repertoire. *Cell*, 125(1), 143–160. <https://doi.org/http://dx.doi.org/10.1016/j.cell.2006.01.050>
- Healy, T. P., Copland, M. J. W., Cork, A., Przyborowska, A., & Halket, J. M. (2002). Landing responses of *Anopheles gambiae* elicited by oxocarboxylic acids. *Medical and Veterinary Entomology*. <https://doi.org/10.1046/j.1365-2915.2002.00353.x>
- Hill, C. A., Fox, A. N., Pitts, R. J., Kent, L. B., Tan, P. L., Chrystal, M. A., ... Zwiebel,

- L. J. (2002). G protein-coupled receptors in *Anopheles gambiae*. *Science*.
<https://doi.org/10.1126/science.1076196>
- Hu, X. S., Ikegami, K., Vihani, A., Zhu, K. W., Zapata, M., de March, C. A., ... Matsunami, H. (2020). Concentration-Dependent Recruitment of Mammalian Odorant Receptors. *ENeuro*, 7(2). <https://doi.org/10.1523/ENEURO.0103-19.2019>
- Hug, J. J., Krug, D., & Müller, R. (2020). Bacteria as genetically programmable producers of bioactive natural products. *Nature Reviews Chemistry*.
<https://doi.org/10.1038/s41570-020-0176-1>
- Jacquier, V., Pick, H., & Vogel, H. (2006). Characterization of an extended receptive ligand repertoire of the human olfactory receptor OR17-40 comprising structurally related compounds. *Journal of Neurochemistry*. <https://doi.org/10.1111/j.1471-4159.2006.03771.x>
- Jaeger, S. R., McRae, J. F., Bava, C. M., Beresford, M. K., Hunter, D., Jia, Y., ... Newcomb, R. D. (2013). A mendelian trait for olfactory sensitivity affects odor experience and food selection. *Current Biology*.
<https://doi.org/10.1016/j.cub.2013.07.030>
- Jones, W. D., Cayirlioglu, P., Grunwald Kadow, I., & Vosshall, L. B. (2007). Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. *Nature*. <https://doi.org/10.1038/nature05466>
- Kalbe, B., Knobloch, J., Schulz, V. M., Wecker, C., Schlimm, M., Scholz, P., ... Osterloh, S. (2016). Olfactory receptors modulate physiological processes in human airway smooth muscle cells. *Frontiers in Physiology*, 7(AUG).
<https://doi.org/10.3389/fphys.2016.00339>
- Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11(9), 1–20.
<https://doi.org/10.1016/j.csda.2009.09.023>
- Katritzky, A. R., Wang, Z., Slavov, S., Tsikolia, M., Dobchev, D., Akhmedov, N. G., ... Linthicum, K. J. (2008). Synthesis and bioassay of improved mosquito repellents predicted from chemical structure. *Proceedings of the National Academy of Sciences of the United States of America*. <https://doi.org/10.1073/pnas.0800571105>
- Keller, A., Gerkin, R. C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., ... Meyer, P. (2017). Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327), 820–826. <https://doi.org/10.1126/science.aal2014>
- Keller, A., & Vosshall, L. B. (2016). Olfactory perception of chemically diverse molecules. *BMC Neuroscience*, 17(1), 55. <https://doi.org/10.1186/s12868-016-0287-2>
- Keller, A., Zhuang, H., Chi, Q., Vosshall, L. B., & Matsunami, H. (2007). Genetic

- variation in a human odorant receptor alters odour perception. *Nature*, 449(7161), 468–472. <https://doi.org/10.1038/nature06162>
- Kent, L. B., Walden, K. K. O., & Robertson, H. M. (2008). The Gr family of candidate gustatory and olfactory receptors in the yellow-fever mosquito *Aedes aegypti*. *Chemical Senses*. <https://doi.org/10.1093/chemse/bjm067>
- Kepple, D., & Koulakov, A. (2017). *Constructing an olfactory perceptual space and predicting percepts from molecular structure*. Retrieved from <http://arxiv.org/abs/1708.05774>
- Khan, R. M., Luk, C.-H., Flinker, A., Aggarwal, A., Lapid, H., Haddad, R., & Sobel, N. (2007). Predicting Odor Pleasantness from Odorant Structure: Pleasantness as a Reflection of the Physical World. *Journal of Neuroscience*. <https://doi.org/10.1523/JNEUROSCI.1158-07.2007>
- Kinsner-Ovaskainen, A., Rzepka, R., Rudowski, R., Coecke, S., Cole, T., & Prieto, P. (2009). Acutoxbase, an innovative database for in vitro acute toxicity studies. *Toxicology in Vitro*. <https://doi.org/10.1016/j.tiv.2008.12.019>
- Klekota, J., & Roth, F. P. (2008). Chemical substructures that enrich for biological activity. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btn479>
- Klocke, J. A., Darlington, M. V., & Balandrin, M. F. (1987). 1,8-Cineole (Eucalyptol), a mosquito feeding and ovipositional repellent from volatile oil of *Hemizonia fitchii* (Asteraceae). *Journal of Chemical Ecology*. <https://doi.org/10.1007/BF01012562>
- Knaden, M., & Hansson, B. S. (2014). Mapping odor valence in the brain of flies and mice. *Current Opinion in Neurobiology*. <https://doi.org/10.1016/j.conb.2013.08.010>
- Knaden, M., Strutz, A., Ahsan, J., Sachse, S., & Hansson, B. S. (2012). Spatial Representation of Odorant Valence in an Insect Brain. *Cell Reports*. <https://doi.org/10.1016/j.celrep.2012.03.002>
- Koulakov, A. A. (2011). In search of the structure of human olfactory space. *Frontiers in Systems Neuroscience*, 5. <https://doi.org/10.3389/fnsys.2011.00065>
- Kowalewski, J., & Ray, A. (2020a). Predicting Human Olfactory Perception from Activities of Odorant Receptors. *IScience*, 23(8), 101361. <https://doi.org/10.1016/j.isci.2020.101361>
- Kowalewski, J., & Ray, A. (2020b). Predicting novel drugs for SARS-CoV-2 using machine learning from a >10 million chemical space. *Heliyon*. <https://doi.org/10.1016/j.heliyon.2020.e04639>
- Kreher, S. A., Mathew, D., Kim, J., & Carlson, J. R. (2008). Translation of Sensory Input into Behavioral Output via an Olfactory System. *Neuron*, 59(1), 110–124. <https://doi.org/http://dx.doi.org/10.1016/j.neuron.2008.06.010>

- Kuhn, M. (2008). caret Package. *Journal Of Statistical Software*, 28(5), 1–26. Retrieved from <http://www.jstatsoft.org/v28/i05/paper>
- Kumar, A., Tauxe, G. M., Perry, S., Scott, C. A., Dahanukar, A., & Ray, A. (2020). Contributions of the Conserved Insect Carbon Dioxide Receptor Subunits to Odor Detection. *Cell Reports*. <https://doi.org/10.1016/j.celrep.2020.03.074>
- Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta - A system for feature selection. *Fundamenta Informaticae*, 101(4), 271–285. <https://doi.org/10.3233/FI-2010-288>
- Kurtovic, A., Widmer, A., & Dickson, B. J. (2007). A single class of olfactory neurons mediates behavioural responses to a *Drosophila* sex pheromone. *Nature*. <https://doi.org/10.1038/nature05672>
- Landrum, G. (2006). RDKit: Open-source cheminformatics. *Online*. <Http://Www.Rdkit.Org>. Accessed.
- Lawless, H. T. (1984). Flavor Description of White Wine by “Expert” and Nonexpert Wine Consumers. *Journal of Food Science*. <https://doi.org/10.1111/j.1365-2621.1984.tb13686.x>
- Li, Y. C., Bai, W. Z., & Hashikawa, T. (2020). The neuroinvasive potential of SARS-CoV2 may be at least partially responsible for the respiratory failure of COVID-19 patients. *Journal of Medical Virology*. <https://doi.org/10.1002/jmv.25728>
- Licon, C. C., Bosc, G., Sabri, M., Mantel, M., Fournel, A., Bushdid, C., ... Bensafi, M. (2019). Chemical features mining provides new descriptive structure-odor relationships. *PLOS Computational Biology*, 15(4), e1006945. Retrieved from <https://doi.org/10.1371/journal.pcbi.1006945>
- Logan, J. G., Birkett, M. A., Clark, S. J., Powers, S., Seal, N. J., Wadhams, L. J., ... Pickett, J. A. (2008). Identification of human-derived volatile chemicals that interfere with attraction of *Aedes aegypti* mosquitoes. *Journal of Chemical Ecology*. <https://doi.org/10.1007/s10886-008-9436-0>
- Lu, T., Qiu, Y. T., Wang, G., Kwon, J. Y., Rutzler, M., Kwon, H. W., ... Zwiebel, L. J. (2007). Odor Coding in the Maxillary Palp of the Malaria Vector Mosquito *Anopheles gambiae*. *Current Biology*. <https://doi.org/10.1016/j.cub.2007.07.062>
- MacWilliam, D., Kowalewski, J., Kumar, A., Pontrello, C., & Ray, A. (2018). Signaling Mode of the Broad-Spectrum Conserved CO2Receptor Is One of the Important Determinants of Odor Valence in *Drosophila*. *Neuron*, 97(5), 1153-1167.e4. <https://doi.org/10.1016/j.neuron.2018.01.028>
- Mahé, P., Ralaivola, L., Stoven, V., & Vert, J. P. (2006). The pharmacophore kernel for virtual screening with support vector machines. *Journal of Chemical Information and Modeling*, 46(5), 2003–2014. <https://doi.org/10.1021/ci060138m>

- Mahevas, M., Tran, V.-T., Roumier, M., Chabrol, A., Paule, R., Guillaud, C., ... Costedoat, N. (2020). No evidence of clinical efficacy of hydroxychloroquine in patients hospitalized for COVID-19 infection with oxygen requirement: results of a study using routinely collected data to emulate a target trial. *MedRxiv*, 2020.04.10.20060699. <https://doi.org/10.1101/2020.04.10.20060699>
- Mainland, J. D., Keller, A., Li, Y. R., Zhou, T., Trimmer, C., Snyder, L. L., ... Matsunami, H. (2014). The missense of smell: Functional variability in the human odorant receptor repertoire. *Nature Neuroscience*, 17(1), 114–120. <https://doi.org/10.1038/nn.3598>
- Majid, A., & Kruspe, N. (2018). Hunter-Gatherer Olfaction Is Special. *Current Biology*, 28(3), 409–413.e2. <https://doi.org/10.1016/j.cub.2017.12.014>
- Maldonado, A. G., Doucet, J. P., Petitjean, M., & Fan, B. T. (2006). Molecular similarity and diversity in chemoinformatics: From theory to applications. *Molecular Diversity*. <https://doi.org/10.1007/s11030-006-8697-1>
- Mao, L., Wang, M., Chen, S., He, Q., Chang, J., Hong, C., ... Hu, B. (2020). Neurological manifestations of COVID-19. *MedRxiv*. <https://doi.org/10.1101/2020.02.22.20026500>
- Mashukova, A., Spehr, M., Hatt, H., & Neuhaus, E. M. (2006). beta-Arrestin2-Mediated Internalization of Mammalian Odorant Receptors. *Journal of Neuroscience*. <https://doi.org/10.1523/JNEUROSCI.2897-06.2006>
- Maßberg, D., Jovancevic, N., Offermann, A., Simon, A., Baniahmad, A., Perner, S., ... Hatt, H. (2016). The activation of OR51E1 causes growth suppression of human prostate cancer cells. *Oncotarget*, 7(30), 48231–48249. <https://doi.org/10.18632/oncotarget.10197>
- Maßberg, D., Simon, A., Häussinger, D., Keitel, V., Gisselmann, G., Conrad, H., & Hatt, H. (2015). Monoterpene (-)-citronellal affects hepatocarcinoma cell signaling via an olfactory receptor. *Archives of Biochemistry and Biophysics*. <https://doi.org/10.1016/j.abb.2014.12.004>
- Matarazzo, V., Clot-Faybesse, O., Marcet, B., Guiraudie-Capraz, G., Atanasova, B., Devauchelle, G., ... Ronin, C. (2005). Functional characterization of two human olfactory receptors expressed in the baculovirus Sf9 insect cell system. *Chemical Senses*. <https://doi.org/10.1093/chemse/bji015>
- Mboera, L. E. G., Knols, B. G. J., Takken, W., & Della Torre, A. (1997). The response of *Anopheles gambiae* s.l. and *A. funestus* (Diptera: Culicidae) to tents baited with human odour or carbon dioxide in Tanzania. *Bulletin of Entomological Research*. <https://doi.org/10.1017/s0007485300027322>
- McClintock, T. S., Khan, N., Alimova, Y., Aulisio, M., Han, D. Y., & Breheny, P. (2020). Encoding the Odor of Cigarette Smoke. *The Journal of Neuroscience : The*

- Official Journal of the Society for Neuroscience*, 40(37), 7043–7053.
<https://doi.org/10.1523/JNEUROSCI.1144-20.2020>
- McClintock, T. S., Wang, Q., Sengoku, T., Titlow, W. B., & Breheny, P. (2020). Mixture and concentration effects on odorant receptor response patterns in vivo. *Chemical Senses*. <https://doi.org/10.1093/chemse/bjaa032>
- McGann, J. P. (2017). Poor human olfaction is a 19th-century myth. *Science*, Vol. 356. <https://doi.org/10.1126/science.aam7263>
- McRae, J. F., Mainland, J. D., Jaeger, S. R., Adipietro, K. A., Matsunami, H., & Newcomb, R. D. (2012). Genetic variation in the odorant receptor OR2J3 is associated with the ability to detect the “grassy” smelling odor, cis-3-hexen-1-ol. *Chemical Senses*. <https://doi.org/10.1093/chemse/bjs049>
- Meijerink, J., & Van Loon, J. J. A. (1999). Sensitivities of antennal olfactory neurons of the malaria mosquito, *Anopheles gambiae*, to carboxylic acids. *Journal of Insect Physiology*. [https://doi.org/10.1016/S0022-1910\(98\)00135-8](https://doi.org/10.1016/S0022-1910(98)00135-8)
- Menashe, I., Abaffy, T., Hasin, Y., Goshen, S., Yahalom, V., Luetje, C. W., & Lancet, D. (2007). Genetic elucidation of human hyperosmia to isovaleric acid. *PLoS Biology*. <https://doi.org/10.1371/journal.pbio.0050284>
- Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., ... Leach, A. R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gky1075>
- Meunier, D., Lambiotte, R., & Bullmore, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in Neuroscience*. <https://doi.org/10.3389/fnins.2010.00200>
- Mombaerts, P. (1999). Molecular biology of odorant receptors in vertebrates. *Annual Review of Neuroscience*, 22, 487–509. <https://doi.org/10.1146/annurev.neuro.22.1.487>
- Mombaerts, P., Wang, F., Dulac, C., Vassar, R., Chao, S. K., Nemes, A., ... Axel, R. (1996). The molecular biology of olfactory perception. *Cold Spring Harbor Symposia on Quantitative Biology*, 61, 135–145.
- Mombaerts, Peter. (2001). THE HUMAN REPERTOIRE OF ODORANT RECEPTOR GENES AND PSEUDOGENES. *Annual Review of Genomics and Human Genetics*, 2(1), 493–510. <https://doi.org/10.1146/annurev.genom.2.1.493>
- Morgan, H. L. (1965). The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2), 107–113. <https://doi.org/10.1021/c160017a018>
- Moriwaki, H., Tian, Y. S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*. <https://doi.org/10.1186/s13321->

018-0258-y

- Neuhaus, E. M., Mashukova, A., Zhang, W., Barbour, J., & Hatt, H. (2006). A specific heat shock protein enhances the expression of mammalian olfactory receptor proteins. *Chemical Senses*. <https://doi.org/10.1093/chemse/bjj049>
- Nisius, B., & Bajorath, J. (2010). Reduction and recombination of fingerprints of different design increase compound recall and the structural diversity of hits. *Chemical Biology and Drug Design*. <https://doi.org/10.1111/j.1747-0285.2009.00930.x>
- Noe, F., Polster, J., Geithe, C., Kotthoff, M., Schieberle, P., & Krautwurst, D. (2017). OR2M3: A highly specific and narrowly tuned human odorant receptor for the sensitive detection of onion key food odorant 3-mercapto-2-methylpentan-1-ol. *Chemical Senses*, 42(3), 195–210. <https://doi.org/10.1093/chemse/bjw118>
- Noronha, A., Modamio, J., Jarosz, Y., Guerard, E., Sompairac, N., Preciat, G., ... Thiele, I. (2019). The Virtual Metabolic Human database: Integrating human and gut microbiome metabolism with nutrition and disease. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gky992>
- Nozaki, Y., & Nakamoto, T. (2016). Odor impression prediction from mass spectra. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0157030>
- Olofsson, J. K., & Gottfried, J. A. (2015). The muted sense: Neurocognitive limitations of olfactory language. *Trends in Cognitive Sciences*, Vol. 19, pp. 314–321. <https://doi.org/10.1016/j.tics.2015.04.007>
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. <https://doi.org/10.3115/v1/d14-1162>
- Pfister, P., Smith, B. C., Evans, B. J., Brann, J. H., Trimmer, C., Sheikh, M., ... Rogers, M. E. (2020). Odorant Receptor Inhibition Is Fundamental to Odor Encoding. *Current Biology : CB*, 30(13), 2574-2587.e6. <https://doi.org/10.1016/j.cub.2020.04.086>
- Python Core Team. (2015). Python: A dynamic, open source programming language. <https://doi.org/10.1109/8.121596>
- Qiu, Y. T., Smallegange, R. C., Hoppe, S., Van Loon, J. J. A., Bakker, E. J., & Takken, W. (2004). Behavioural and electrophysiological responses of the malaria mosquito *Anopheles gambiae* Giles sensu stricto (Diptera: Culicidae) to human skin emanations. *Medical and Veterinary Entomology*. <https://doi.org/10.1111/j.0269-283X.2004.00534.x>
- R Development Core Team. (2016). R: A Language and Environment for Statistical

- Computing. *R Foundation for Statistical Computing Vienna Austria, 0*, {ISBN} 3-900051-07-0. <https://doi.org/10.1038/sj.hdy.6800737>
- R Development Core Team, R. (2011). R: A Language and Environment for Statistical Computing. In *R Foundation for Statistical Computing*. <https://doi.org/10.1007/978-3-540-74686-7>
- Raiche, G. (2010). nFactors: an R package for parallel analysis and non graphical solutions to the Cattell scree test.
- Reeder, N. L., Ganz, P. J., Carlson, J. R., & Saunders, C. W. (2009). Isolation of a Deet-Insensitive Mutant of *Drosophila melanogaster* (Diptera: Drosophilidae). *Journal of Economic Entomology*. <https://doi.org/10.1603/0022-0493-94.6.1584>
- Richard, A. M., & Williams, C. L. R. (2002). Distributed structure-searchable toxicity (DSSTox) public database network: A proposal. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*. [https://doi.org/10.1016/S0027-5107\(01\)00289-5](https://doi.org/10.1016/S0027-5107(01)00289-5)
- Riva, L., Yuan, S., Yin, X., Martin-Sancho, L., Matsunaga, N., Burgstaller-Muehlbacher, S., ... Chanda, S. K. (2020). A Large-scale Drug Repositioning Survey for SARS-CoV-2 Antivirals. *BioRxiv*, 2020.04.16.044016. <https://doi.org/10.1101/2020.04.16.044016>
- Robertson, H. M., & Kent, L. B. (2009). Evolution of the gene lineage encoding the carbon dioxide receptor in insects. *Journal of Insect Science*. <https://doi.org/10.1673/031.009.1901>
- Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*. <https://doi.org/10.1021/ci100050t>
- Roland, B., Deneux, T., Franks, K. M., Bathellier, B., & Fleischmann, A. (2017). Odor identity coding by distributed ensembles of neurons in the mouse olfactory cortex. *ELife*, 6. <https://doi.org/10.7554/eLife.26337>
- Rosenberg, R., Lindsey, N. P., Fischer, M., Gregory, C. J., Hinckley, A. F., Mead, P. S., ... Petersen, L. R. (2018). Vital signs: Trends in reported vectorborne disease cases — United States and Territories, 2004-2016. *Morbidity and Mortality Weekly Report*. <https://doi.org/10.15585/mmwr.mm6717e1>
- Rossiter, K. J. (1996). Structure–Odor Relationships. *Chemical Reviews*, 96(8), 3201–3240. <https://doi.org/10.1021/cr950068a>
- Saito, H., Chi, Q., Zhuang, H., Matsunami, H., & Mainland, J. D. (2009). Odor coding by a mammalian receptor repertoire. *Science Signaling*, 2(60). <https://doi.org/10.1126/scisignal.2000016>
- Saito, H., Kubota, M., Roberts, R. W., Chi, Q., & Matsunami, H. (2004). RTP family members induce functional expression of mammalian odorant receptors. *Cell*,

- 119(5), 679–691. <https://doi.org/10.1016/j.cell.2004.11.021>
- Sanche, S., Lin, Y. T., Xu, C., Romero-Severson, E., Hengartner, N., & Ke, R. (2020). High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. *Emerging Infectious Disease Journal*, 26(7). <https://doi.org/10.3201/eid2607.200282>
- Sanchez-Lengeling, B., Wei, J. N., Lee, B. K., Gerkin, R. C., Aspuru-Guzik, A., & Wiltschko, A. B. (2019). *Machine Learning for Scent: Learning Generalizable Perceptual Representations of Small Molecules*. Retrieved from <http://arxiv.org/abs/1910.10685>
- Santosh, K. C. (2020). AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data. *Journal of Medical Systems*. <https://doi.org/10.1007/s10916-020-01562-1>
- Sanz, G., Schlegel, C., Pernollet, J. C., & Briand, L. (2005). Comparison of odorant specificity of two human olfactory receptors from different phylogenetic classes and evidence for antagonism. *Chemical Senses*. <https://doi.org/10.1093/chemse/bji002>
- Schaffer, E. S., Stettler, D. D., Kato, D., Choi, G. B., Axel, R., & Abbott, L. F. (2018). Odor Perception on the Two Sides of the Brain: Consistency Despite Randomness. *Neuron*, 98(4), 736–742.e3. <https://doi.org/10.1016/j.neuron.2018.04.004>
- Schmiedeberg, K., Shirokova, E., Weber, H. P., Schilling, B., Meyerhof, W., & Krautwurst, D. (2007). Structural determinants of odorant recognition by the human olfactory receptors OR1A1 and OR1A2. *Journal of Structural Biology*. <https://doi.org/10.1016/j.jsb.2007.04.013>
- Schuur, J. H., Selzer, P., & Gasteiger, J. (1996). The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *Journal of Chemical Information and Computer Sciences*, 36(2), 334–344. <https://doi.org/10.1021/ci950164c>
- Sheahan, T. P., Sims, A. C., Zhou, S., Graham, R. L., Pruijssers, A. J., Agostini, M. L., ... Baric, R. S. (2020). An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 in human airway epithelial cell cultures and multiple coronaviruses in mice. *Science Translational Medicine*. <https://doi.org/10.1126/scitranslmed.abb5883>
- Shirasu, M., Yoshikawa, K., Takai, Y., Nakashima, A., Takeuchi, H., Sakano, H., & Touhara, K. (2014). Olfactory receptor and neural pathway responsible for highly selective sensing of musk odors. *Neuron*. <https://doi.org/10.1016/j.neuron.2013.10.021>
- Smallegange, R. C., Qiu, Y. T., van Loon, J. A., & Takken, W. (2005). Synergism between ammonia, lactic acid and carboxylic acids as kairomones in the host-seeking behaviour of the malaria mosquito *Anopheles gambiae sensu stricto*

- (Diptera: Culicidae). *Chemical Senses*. <https://doi.org/10.1093/chemse/bji010>
- Snitz, K., Yablonka, A., Weiss, T., Frumin, I., Khan, R. M., & Sobel, N. (2013). Predicting Odor Perceptual Similarity from Odor Structure. *PLoS Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1003184>
- spaCy. (2016). spaCy API Documentation.
- Spehr, M., Gisselmann, G., Poplawski, A., Riffell, J. A., Wetzel, C. H., Zimmer, R. K., & Hatt, H. (2003). Identification of a testicular odorant receptor mediating human sperm chemotaxis. *Science*. <https://doi.org/10.1126/science.1080376>
- Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *Journal of Chemical Information and Computer Sciences*. <https://doi.org/10.1021/ci025584y>
- Stensmyr, M. C., Dweck, H. K. M., Farhan, A., Ibba, I., Strutz, A., Mukunda, L., ... Hansson, B. S. (2012). A conserved dedicated olfactory circuit for detecting harmful microbes in drosophila. *Cell*. <https://doi.org/10.1016/j.cell.2012.09.046>
- Sterling, T., & Irwin, J. J. (2015). ZINC 15 - Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*. <https://doi.org/10.1021/acs.jcim.5b00559>
- Stettler, D. D., & Axel, R. (2009). Representations of Odor in the Piriform Cortex. *Neuron*, 63(6), 854–864. <https://doi.org/10.1016/j.neuron.2009.09.005>
- Strutz, A., Soelter, J., Baschwitz, A., Farhan, A., Grabe, V., Rybak, J., ... Sachse, S. (2014). Decoding odor quality and intensity in the Drosophila brain. *ELife*. <https://doi.org/10.7554/eLife.04147>
- Suh, G. S. B., Wong, A. M., Hergarden, A. C., Wang, J. W., Simon, A. F., Benzer, S., ... Anderson, D. J. (2004). A single population of olfactory sensory neurons mediates an innate avoidance behaviour in Drosophila. *Nature*. <https://doi.org/10.1038/nature02980>
- Sungnak, W., Huang, N., Bécavin, C., Berg, M., Queen, R., Litvinukova, M., ... Network, H. C. A. L. B. (2020). SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nature Medicine*. <https://doi.org/10.1038/s41591-020-0868-6>
- Swale, D. R., Sun, B., Tong, F., & Bloomquist, J. R. (2014). Neurotoxicity and mode of action of N, N-diethyl-Meta-toluamide (DEET). *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0103713>
- Syed, Z., & Leal, W. S. (2007). Maxillary palps are broad spectrum odorant detectors in culex quinquefasciatus. *Chemical Senses*. <https://doi.org/10.1093/chemse/bjm040>
- Syed, Z., & Leal, W. S. (2008). Mosquitoes smell and avoid the insect repellent DEET.

- Proceedings of the National Academy of Sciences of the United States of America*.
<https://doi.org/10.1073/pnas.0805312105>
- Syed, Z., Pelletier, J., Flounders, E., Chitolina, R. F., & Leal, W. S. (2011). Generic insect repellent detector from the fruit fly *Drosophila melanogaster*. *PLoS ONE*.
<https://doi.org/10.1371/journal.pone.0017705>
- Takken, W., & Knols, B. G. J. (1999). Odor-mediated behavior of Afrotropical malaria mosquitoes. *Annual Review of Entomology*.
<https://doi.org/10.1146/annurev.ento.44.1.131>
- Tauxe, G. M., Macwilliam, D., Boyle, S. M., Guda, T., & Ray, A. (2013). Targeting a dual detector of skin and CO₂ to modify mosquito host seeking. *Cell*.
<https://doi.org/10.1016/j.cell.2013.11.013>
- Tham, E. H., Dyjack, N., Kim, B. E., Rios, C., Seibold, M. A., Leung, D. Y. M., & Goleva, E. (2019). Expression and function of the ectopic olfactory receptor OR10G7 in patients with atopic dermatitis. *Journal of Allergy and Clinical Immunology*, 143(5), 1838-1848.e4. <https://doi.org/10.1016/j.jaci.2018.11.004>
- Topin, J., Demarch, C. A., Charlier, L., Ronin, C., Antonczak, S., & Golebiowski, J. (2014). Discrimination between olfactory receptor agonists and non-agonists. *Chemistry - A European Journal*. <https://doi.org/10.1002/chem.201402486>
- Tran, N. B., Kepple, D. R., Shuvaev, S. A., & Koulakov, A. A. (2019). Deepnose: Using artificial neural networks to represent the space of odorants. *36th International Conference on Machine Learning, ICML 2019*. <https://doi.org/10.1101/464735>
- Trimmer, C., Keller, A., Murphy, N. R., Snyder, L. L., Willer, J. R., Nagai, M. H., ... Mainland, J. D. (2019). Genetic variation across the human olfactory receptor repertoire alters odor perception. *Proceedings of the National Academy of Sciences*, 116(19), 9475 LP – 9480. <https://doi.org/10.1073/pnas.1804106115>
- Turner, S. L., Li, N., Guda, T., Githure, J., Cardé, R. T., & Ray, A. (2011). Ultra-prolonged activation of CO₂-sensing neurons disorients mosquitoes. *Nature*.
<https://doi.org/10.1038/nature10081>
- Turner, S. L., & Ray, A. (2009). Modification of CO₂ avoidance behaviour in *Drosophila* by inhibitory odorants. *Nature*. <https://doi.org/10.1038/nature08295>
- Van Den Broek, I. V. F., & Den Otter, C. J. (1999). Olfactory sensitivities of mosquitoes with different host preferences (*Anopheles gambiae* s.s., *An. arabiensis*, *An. quadriannulatus*, *An. m. atroparvus*) to synthetic host odours. *Journal of Insect Physiology*. [https://doi.org/10.1016/S0022-1910\(99\)00081-5](https://doi.org/10.1016/S0022-1910(99)00081-5)
- Vassar, R., Ngai, J., & Axel, R. (1993). Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. *Cell*, 74(2), 309–318.
[https://doi.org/0092-8674\(93\)90422-M](https://doi.org/0092-8674(93)90422-M) [pii]

- Verhulst, N. O., Andriessen, R., Groenhagen, U., Kiss, G. B., Schulz, S., Takken, W., ... Smallegange, R. C. (2010). Differential attraction of malaria mosquitoes to volatile blends produced by human skin bacteria. *PLoS ONE*.
<https://doi.org/10.1371/journal.pone.0015829>
- Verhulst, N. O., Qiu, Y. T., Beijleveld, H., Maliepaard, C., Knights, D., Schulz, S., ... Smallegange, R. C. (2011). Composition of human skin microbiota affects attractiveness to malaria mosquitoes. *PLoS ONE*.
<https://doi.org/10.1371/journal.pone.0028991>
- Walker, J. D., Rodford, R., & Patlewicz, G. (2003). Quantitative structure-activity relationships for predicting percutaneous absorption rates. *Environmental Toxicology and Chemistry*. <https://doi.org/10.1897/01-454>
- Wan, Y., Shang, J., Graham, R., Baric, R. S., & Li, F. (2020). Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *Journal of Virology*. <https://doi.org/10.1128/jvi.00127-20>
- Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., ... Xiao, G. (2020). Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Research*. <https://doi.org/10.1038/s41422-020-0282-0>
- Weiss, T., Snitz, K., Yablonka, A., Khan, R. M., Gafsou, D., Schneidman, E., & Sobel, N. (2012). Perceptual convergence of multi-component mixtures in olfaction implies an olfactory white. *Proceedings of the National Academy of Sciences*, 109(49), 19959–19964. <https://doi.org/10.1073/pnas.1208110109>
- Wickham, H., & Chang, W. (2016). Package ‘ggplot2’ Create Elegant Data Visualisations Using the Grammar of Graphics Description.
<https://doi.org/10.1093/bioinformatics/btr406>
- Williamson, B., Feldmann, F., Schwarz, B., Meade-White, K., Porter, D., Schulz, J., ... de Wit, E. (2020). Clinical benefit of remdesivir in rhesus macaques infected with SARS-CoV-2. *BioRxiv*, 2020.04.15.043166.
<https://doi.org/10.1101/2020.04.15.043166>
- Wilson, C. D., Serrano, G. O., Koulakov, A. A., & Rinberg, D. (2017). A primacy code for odor identity. *Nature Communications*, 8(1). <https://doi.org/10.1038/s41467-017-01432-4>
- Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... Wilson, M. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkx1037>
- Xue, R. De, Doyle, M. A., & Kline, D. L. (2008). Field evaluation of CDC and mosquito magnet® X traps baited with dry ice, CO2 sachet, and octenol against mosquitoes. *Journal of the American Mosquito Control Association*.
<https://doi.org/10.2987/5701.1>

- Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., & Zhou, Q. (2020). Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science (New York, N.Y.)*. <https://doi.org/10.1126/science.abb2762>
- Zang, Q., Mansouri, K., Williams, A. J., Judson, R. S., Allen, D. G., Casey, W. M., & Kleinstreuer, N. C. (2017). In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. *Journal of Chemical Information and Modeling*. <https://doi.org/10.1021/acs.jcim.6b00625>
- Zhou, Y., Smith, B. H., & Sharpee, T. O. (2018). Hyperbolic geometry of the olfactory space. *Science Advances*. <https://doi.org/10.1126/sciadv.aaq1458>
- Zhu, F., Han, B. C., Kumar, P., Liu, X. H., Ma, X. H., Wei, X. N., ... Chen, Y. Z. (2009). Update of TTD: Therapeutic Target Database. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkp1014>