Diving Deeper into VOCs:

Predicting Formulation Component GC-MS Response Factor Using Quantitative Structure-activity Relationships Coupled with Artificial Neural Networks

By Jessica Lum, Madeline Schultz, and Erik Sapper,
Department of Chemistry and Biochemistry,
California Polytechnic State University

The identification, measurement, and reduction of volatile organic compounds (VOCs) has been a key motivator in recent coatings research and development efforts. Analytical methods for determining VOC levels in organic coatings continue to improve, as chromatographic and spectroscopic approaches afford a means of quantifying VOC content directly in waterborne as well as solventborne coatings.

Heuristic methods for estimating the volatility of formulation components are common but are not extensively validated using quantitative structure-property relationships. Thus, a clearer link between component transport through an evolving coating matrix during curing processes, the bulk volatility of a compound, and the elution and quantification of compounds in a gas chromatograph (GC) still must be made to promote innovation in this area.

To address these issues, digital tools such as molecular descriptors and machine learning models are being combined with experimental measurements to better understand the time-dependent mechanistic nature of VOCs in coatings and to enable predictive control over the volatility and in-coating behavior of newly developed formulation components.

Here, we present the development and validation of a molecular structure-based neural network for the prediction of response factor for formulation components in a gas chromatography (GC) analysis. This represents an important step in creating large-scale computational design tools that enable in silico formulation, optimization, and end-use property prediction of environmentally benign coatings.

INTRODUCTION

Consumer and market demand within the coatings industry continues to put pressure on formulators to create high-performance coatings that also have adequately low levels of volatile organic compounds. An ongoing challenge is the creation and optimization of important end-use coating properties while still meeting environmental regulation specifications.

As formulators are urged to innovate more quickly, it has become apparent that traditional empirical and Edisonian (guess-and-test) methods, even statistically designed methods of formulation discovery, must be augmented with newer technologies, such as those represented by digitization, automation, machine learning, and artificial intelligence.

There is also an increased emphasis on understanding chemical and physical interactions within the formulation at all stages of the paint production, application, and film-forming process. The growing consumer demand for environmentally benign "green" coatings has led to a push within the paint industry for improved predictive models and developmental workflows that make use of these next generation technologies.

Consider, for example, the recent South Coast Air Quality Management District (SCAQMD) Test Method 319 (Determination of Exclusion Status for Compounds in Film-Forming Coatings), where measurement, estimation, or prediction of the low vapor pressure of a formulation component may lead to its exclusion in VOC calculation and reporting.

Environmentally conscious consumers and regulatory agencies such as the U.S. Environmental Protection Agency (EPA) have continued to drive the paint and coatings industry towards greener formulating methods, such as shifting from solvent-based to water-based coatings as a method of reducing VOCs.

Throughout the late 1960s and 1970s, there was an increased concern regarding air pollution and the detrimental effects to both human and environmental health.

Diving Deeper into VOCs

From this pollution arose the need to define and regulate the effect of paints and coatings on the local environment by limiting the amount of certain additives in paint which are damaging to the environment.

The EPA identified volatile organic compounds as "any compound of carbon, excluding carbon monoxide, carbon dioxide, carbonic acid, metallic carbides or carbonates, and ammonium carbonate, which participates in atmospheric photochemical reactions, except those designated by the EPA as having negligible photochemical reactivity."[1]

The EPA calculates compliance with VOC content regulations (Title 40, Chapter I, Subchapter C, Part 59, Subpart D—National Volatile Organic Compound Emission Standards for Architectural Coatings S: 63 FR 48877, Sept. 11, 1998, §59.406) according to *Equation 1*:

$$\text{VOC content} = \frac{(W_v - W_w - W_{ec})}{(V - V_w - V_{ec})} \qquad (1)$$

In *Equation 1*, VOC content = grams of VOC per liter of coating; W_v = mass of total volatiles, in grams; W_w = mass of water, in grams; W_{ec} = mass of exempt compounds, in grams; V = volume of coating, in liters; V_w = volume of water, in liters; and V_{ec} = volume of exempt compounds, in liters.

In 1984, the EPA introduced Method 24 to quantify the amount of VOCs in coatings and inks sold in the United States. Method 24 is an indirect method of VOC determination, wherein the water content, solids content, and density of the coating are directly measured and used to back-calculate the amount of VOCs by a mass difference approach.

Method 24 is insufficient for waterborne coatings with low VOC content, as the indirect method erroneously determines small mass fractions of VOCs as compared to the much larger water weight percent, with exponentially increasing error below VOC content of approximately 250 g/L.

As coatings shifted from solvent-based formulations to more environmentally friendly water-based formulations, the insufficiencies in this method motivated the need for new standardized regulatory methods and measurement procedures. Despite the need for improved methods, EPA Method 24 is currently the regulatory method federally mandated across the United States.

States and regions throughout the United States have various guidelines that extend beyond federal rules. California, particularly the Los Angeles air basin, has faced, and *continues* to face, high prevalence of air pollution known as "smog," a portmanteau coined in the 1900s to describe the uniquely industrial mixture of smoke and fog becoming increasingly prevalent in large urban areas.

Regulatory agencies such as the California Air Resource Board (CARB), and more specifically SCAQMD, formed the most stringent regulations in the United States to reduce the local effects of this increasing pollution. Method 313 is a direct method for the measurement and quantitation of VOCs using a gas chromatograph with flame ionization detector (GC-FID) applied to samples with less than 150g/L of VOCs.

The complexity of this method is the main deterrent to its use. VOCs are quantified via multilevel calibration curves generated for each analyte used in the coating formulation.[2] Relative response factors allow for the calculation of volatiles through this direct method. The regulation of VOCs is relative to the retention time of methyl palmitate. Compounds that elute prior to methyl palmitate are not included in the calculation of volatiles per liter coating. The complexity and laborious sample preparation associated with this method render its use undesirable and drove the innovation of a new standard: ASTM D6886.

ASTM D6886 is a non-regulatory analytical method suitable for the analysis of coatings with less than 150g/L of VOCs, which resulted from an in-depth study by California Polytechnic State University for the California Air Resource Board.[3] This method does not define a VOC as Method 313 does, rather it identifies and quantifies *all* volatiles within a formulation. Although it is not regulatory in nature, it has been widely adopted by SCAQMD as it provides for a less labor-intensive direct measurement of VOC content in coatings as compared to Method 313.

Like Method 313, GC-FID is used in ADTM D6886 to quantify the volatile compounds present in the material. This method utilizes an internal standard, ethylene glycol diethyl ether (EGDE), for the calculation of response factors for an analyte of interest, as discussed in subsequent sections of this manuscript. Herein all response factors discussed were collected according to ASTM D6886.

Globally, VOCs are regulated by federal and local governments. Looking beyond the United States, Europe developed ISO 11890, a widely employed direct method for the analysis of samples with expected VOC content between 0.1% and 15% by mass.[4]

While Method 313 defines a VOC as anything that elutes before methyl palmitate, ISO 11890 defines a VOC as compounds with a boiling point below 250 °C. This is dictated by EU Directive 2004/42/EU.[4] ASTM-D6886 and ISO 11890 are very similar in practice, with direct measurements preformed via GC-FID, and primarily differ in the associated VOC determination that follows as dictated by regulatory agencies within relevant regions.

Here, we combine structure-property relationships, neural networks, and gas chromatographic analytical methods to create a digitally enabled workflow that can support the formulator chemist while evolving as quickly as the regulations themselves.

We present a multipronged approach to working with, measuring, and understanding the nature of VOCs in coatings formulations. First, we present a method of improved prediction for quantifying the response factor (RF) of compounds being analyzed by gas chromatography, as a means of augmenting and expediting VOC determination by ASTM D6886 and other chromatographic approaches.

Ongoing work is employing vapor pressure (VP) prediction and measurement to improve the working definition of VOC as it applies to coating production, application, and film-formation processes. Finally, we propose new directions for incorporating these predictive approaches into the formulation development process.

TABLE 1—Calculated molecular descriptor with the largest positive or negative correlation with compound response factor.

Number	Descriptor	Definition	Correlation
1	O%	Percentage of oxygen atoms	-0.8480
2	AAC	Mean information index on atomic composition	-0.8194
3	IC0	Information Content index (neighborhood symmetry of 0-order)	-0.8194
4	MLOGP	Moriguchi octanol-water partition coeff. (logP)	0.8071
5	BLTA96	Verhaar Algae base-line toxicity from MLOGP (mmol/l)	-0.8071
6	BLTD48	Verhaar Daphnia base-line toxicity from MLOGP (mmol/l)	-0.8071
7	BLTF96	Verhaar Fish base-line toxicity from MLOGP (mmol/l)	-0.8071
8	Me	Mean atomic Sanderson electronegativity (scaled on Carbon atom)	-0.7739
9	Mor31s	Signal 31 / weighted by I-state	0.7641
10	SM1_Dz(p)	Spectral moment of order 1 from Barysz matrix weighted by polarizability	0.7614
11	SpMin1_Bh(m)	Smallest eigenvalue n. 1 of Burden matrix weighted by mass	0.7566
12	Psi_i_A	Intrinsic state pseudoconnectivity index - type S average	-0.7494
13	Psi_e_A	Electrotopological state pseudoconnectivity index - type S average	-0.7494
14	BIC0	Bond Information Content index (neighborhood symmetry of 0-order)	-0.7478
15	CATS2D_00_LL	CATS2D Lipophilic-Lipophilic at lag 00	0.7474
16	CATS2D_01_LL	CATS2D Lipophilic-Lipophilic at lag 01	0.7463
17	SM1_Dz(v)	Spectral moment of order 1 from Barysz matrix weighted by van der Waals volume	0.7363
18	SIC0	Structural Information Content index (neighborhood symmetry of 0-order)	-0.7325
19	SM1_Dz(Z)	Spectral moment of order 1 from Barysz matrix weighted by atomic number	-0.7308
20	Eta_alpha_A	Eta average core count	0.7307

MATERIALS AND METHODS

Response factor determination by GC

The response factor, or RF, of an analyte compound is the ratio between the chromatographic signal produced by the compound and the quantity or amount of analyte which produces the signal. Ideally, this ratio is 1.0, or unity, allowing for simple quantification and comparison of analyte composition in a tested mixture, although differences in compound activity within an analytical instrument usually cause deviations from unity.

The role of accurate RF measurements in VOC analysis is critical; with faulty RF information, calculations of VOC content in a tested formulation may not be reliable. Further, newly created compounds or additives must be characterized against an internal standard to empirically determine the RF before analysis of chromatograms may occur.

RF allows for the quantitation of analyte in a mixture as compared to an internal standard, as defined by *Equation 2*.

$$RF = \frac{MA \times AA}{MI \times AI} \quad (2)$$

Standards with equal weights of a chosen analyte and internal standard are used to determine the response factor for analyte of interest. The mass of the analyte added (MA) and the relative peak area (AA) from the FID spectrum are standardized by the mass of an internal standard (MI) and the associated peak area for that internal standard (AI). Ethylene glycol diethyl ether (EGDE) was used as the internal standard.[3]

Quantitative structure activity relationships for identifying molecular features relevant to response factor

The quantitative structure activity relationship (QSAR) approach makes use of large numbers of chemical and topological descriptors that correlate molecular structure to activities or properties of interest.[5] The use of QSAR serves two purposes: 1) discover or validate heuristic relationships and 2) provide a list of relevant features or inputs for use in subsequent modeling and prediction exercises.

A set of 80 compounds commonly seen in VOC analysis of coatings by GC-MS was chosen as a dataset for this study. Molecular structures were represented using simplified molecular-input line-entry system (SMILES) strings, which were generated for all 80 compounds in the dataset on a Dell XPS 13 9360 laptop running Windows 10.

The Avogadro molecular editor[6] was used to create rough three-dimensional geometries for each compound, which were then optimized using a quick energy minimization algorithm. Then, a total of 5,270 descriptors were calculated for each compound using Dragon 7.[7] Of these, 2,130 were constant, showing no change across the entire set of compounds; 2,301 were near constant, showing negligible change across the compound set; 155 had at least one value missing or incalculable due to molecular structure; and 15 had all values missing or incalculable. These descriptors were removed from the analysis.

The resultant set of 669 descriptors for each of the 80 compounds in the dataset was then subjected to correlation analysis to identify the 20 descriptors with the highest positive or negative correlation to RF. These descriptors are listed in *Table 1*.

The data were normalized with by applying *Equation 3*:

$$(RF_{norm})_y = \frac{RF_y - RF_{min}}{RF_{max} - RF_{min}} \quad (3)$$

where RF_{norm} is the normalized RF of compound *y*, RF_y is the response factor of a generic compound *y*, RF_{max} is the highest RF measured, and RF_{min} is the lowest RF measured, as shown in *Table 2*. Normalizing RF constrains all data to a range between 0 and 1, with 0 being the lowest RF and 1 being the highest RF, which allows for greater efficiency and accuracy during training.[8] The same formula was used to normalize all descriptors to values between 0 and 1.

Deep-learning artificial neural networks (DLANN) for creating production-grade predictive models for new compound response factor estimation

After normalization, the data were randomly divided into two groups; 60 molecules (75%) were allocated to a training set to be used to build and teach the machine learning model, while 20 molecules (25%) were withheld for the validation set. The holdout validation set quantifies the ability of the model to generalize its ruleset to compounds that it has never seen before.

A deep-learning artificial neural network as a nonlinear regression model with Adam optimizer was created in Python 3 in a Jupyter Notebook 6.3.0 using TensorFlow and Keras deep learning libraries. Loss was calculated as mean square error (MSE). Hyperparameter tuning revealed that the model achieved optimal performance in the training and holdout validation sets with 19 descriptors, 500 epochs (or cycles of model learning with exposure to the data), and one hidden layer of consisting of three nodes, or perceptrons.

RESULTS

Figure 1 shows the correlation between predicted and actual values of RF during a training and validation of a neural network with up to 1,000 learning cycles through the data.

Figure 1 shows that with repeated learning, the trained neural network performs better, while the unseen data in the validation set is less able to be appropriately captured by these long-trained models. This is an indication of overfitting; the trained neural network is effectively learning how to memorize the data in the training set. A compromise must be selected that balances performance of the trained model against performance of the model when used with new data. Here, hyperparameter selection indicated that 500 epochs of learning were a suitable stopping point during the model build process.

Figure 2 shows the performance of the RF prediction neural network after being trained on the dataset after 500 epochs of model evolution. The trained neural network shows good agreement between predicted and measured (experimental) values of RF, indicated by the close linear fit to the identity (x=y) line, with an R^2 value of 0.90.

FIGURE 1—Epoch selection for a single-layer artificial neural network using molecular 19 descriptors as input nodes.

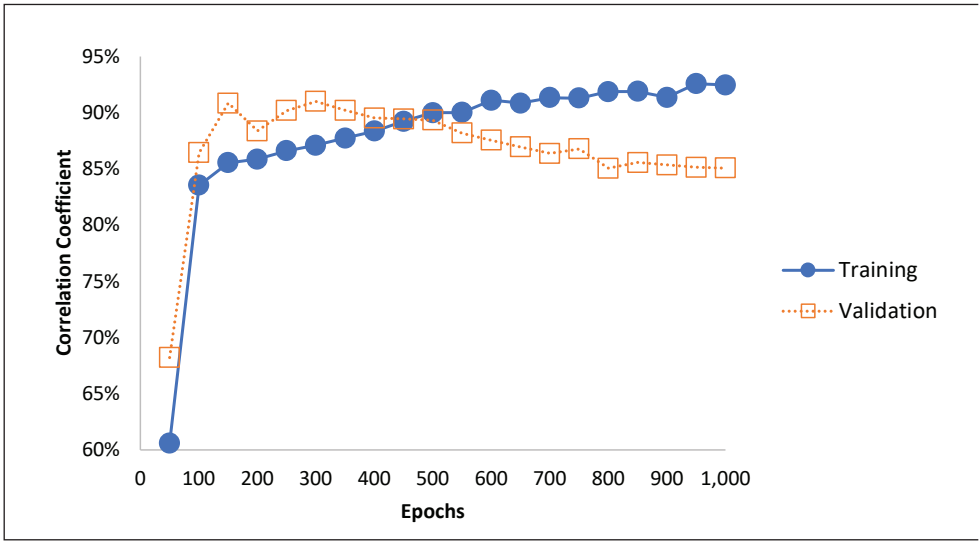

FIGURE 2—Performance of the RF predictive neural network after 500 learning cycles on the dataset. A close linear fit and high R^2 value (0.90) indicates that the model has adequately learned from the dataset.

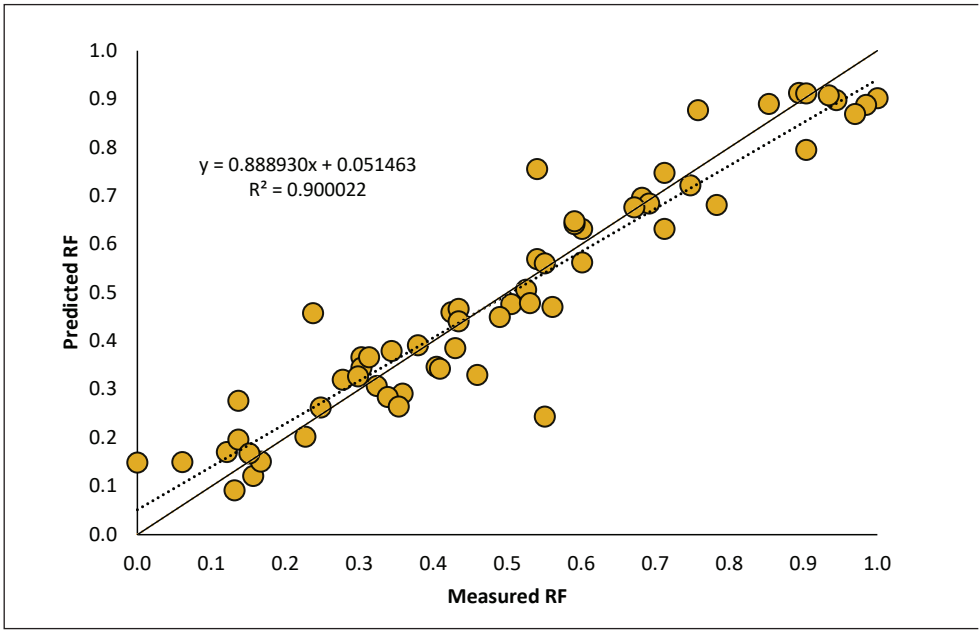

TABLE 2—Response factors and retention times for 80 compound dataset for VOC analysis, with absolute and normalized values provided.

#.	Compound	SMILES	RF	Norm RF	RT	Norm RT
			RF=Response Factor; Norm RF= Normalized Response Factor; RT= Retention Time; Norm RT=Normalized Retention Time			
1.	(3-Hydroxy-2,2,4-trimethylpentyl) 2-methylpropanoate	CC(C)C(C(C)(C)COC(=O)C(C)C)O	1.31	0.5354	856	0.66589327
2.	[2,2,4-Trimethyl-1-(2-methylpropanoyloxy)pentyl] 2-methylpropanoate	CC(C)CC(C)(C)C(OC(=O)C(C)C)OC(=O)C(C)C	1.32	0.5404	961	0.76798144
3.	1,2-Diethoxyethane	CCOCCOCC	1.00	0.3788	520	0.34106729
4.	1,3,5-Trimethylbenzene	CC1=CC(=CC(=C1)C)C	2.23	1.0000	657	0.47389791
5.	1,4-Xylene	CC1=CC=C(C=C1)C	2.20	0.9848	592	0.41067285
6.	1-[2-(2-Methoxypropoxy)propoxy]butane	CCCCOCC(C)OCC(C)OC	1.07	0.4141	791	0.60382831
7.	1-Butoxybutane	CCCCOCCCC	1.60	0.6818	590	0.40893271
8.	1-Butoxypropan-2-ol	CCCCOCC(C)O	1.15	0.4545	630	0.44779582
9.	1-Chloro-4-(trifluoromethyl)benzene	C1=CC(=CC=C1C(F)(F)F)Cl	1.10	0.4293	574	0.39327146
10.	1-Methoxy-2-[2-(2-methoxyethoxy)ethoxy]ethane	COCCOCCOCCOC	0.77	0.2626	781	0.59338747
11.	1-Methoxypropyl acetate	CCC(OC)OC(=O)C	0.88	0.3182	572	0.39153132
12.	1-Methylpyrrolidin-2-one	CN1CCCC1=O	0.96	0.3586	692	0.50812065
13.	1-Phenoxypropan-2-ol	CC(COC1=CC=CC=C1)O	1.29	0.5253	804	0.61600928
14.	2-(2-Butoxyethoxy)ethanol	CCCCOCCOCCO	1.06	0.4091	769	0.58178654
15.	2-(2-Hydroxyethoxy)ethanol	C(COCCO)O	0.49	0.1212	641	0.45823666
16.	2-(2-Methoxyethoxy)ethanol	COCCOCCO	0.70	0.2273	624	0.44199536
17.	2-(2-Methoxypropoxy)propan-1-ol	CC(CO)OCC(C)OC	0.89	0.3232	663	0.47969838
18.	2-(2-Propoxyethoxy)ethanol	CCCOCCOCCO	0.80	0.2778	718	0.53248260
19.	2,4,7,9-Tetramethyldec-5-yne-4,7-diol	CC(C)CC(C)(C#CC(C)(CC(C)C)O)O	1.62	0.6919	867	0.67691415
20.	2-[2-(2-Hydroxyethoxy)ethoxy]ethanol	C(COCCOCCO)O	0.37	0.0606	755	0.56844548
21.	2-[2-(2-Methoxypropoxy)propoxy]propan-1-ol	CC(CO)OCC(C)OCC(C)OC	0.92	0.3384	813	0.62470998
22.	2-[2-[2-(2-Hydroxyethoxy)ethoxy]ethoxy]ethanol	C(COCCOCCOCCO)O	0.56	0.1566	917	0.72563805
23.	2-[2-[2-[2-(2-Hydroxyethoxy)ethoxy]ethoxy]ethoxy]ethanol	C(COCCOCCOCCOCCO)O	0.51	0.1313	1088	0.89095128
24.	2-[Butyl(2-hydroxyethyl)amino]ethanol	CCCCN(CCO)CCO	1.10	0.4293	836	0.64733179
25.	2-Amino-2-ethylpropane-1,3-diol	CCC(CO)(CO)N	0.52	0.1364	737	0.55162413
26.	2-Amino-2-methylpropan-1-ol	CC(C)(CO)N	0.85	0.3030	488	0.31032483
27.	2-Benzofuran-1,3-dione	C1=CC=C2C(=C1)C(=O)OC2=O	0.46	0.1061	845	0.65545244
28.	2-Butoxyethanol	CCCCOCCO	0.72	0.2374	607	0.42575406
29.	2-Butoxyethanol	CCCCOCCO	1.15	0.4545	607	0.42575406
30.	2-Ethoxyethyl acetate	CCOCCOC(=O)C	1.34	0.5505	599	0.41763341
31.	2-Ethyl-2-(hydroxymethyl)propane-1,3-diol	CCC(CO)(CO)CO	1.05	0.4040	817	0.62819026
32.	2-Ethylhexanal	CCCCC(CC)C=O	1.73	0.7475	683	0.49941995
33.	2-Ethylhexyl benzoate	CCCCC(CC)COC(=O)C1=CC=CC=C1	1.58	0.6717	1041	0.84512761
34.	2-Methylpentane-2,4-diol	CC(CC(C)(C)O)O	1.09	0.4242	616	0.43387471
35.	2-Methylprop-2-enoic acid	CC(=C)C(=O)O	0.95	0.3535	505	0.32656613
36.	2-Methylpropan-1-ol	CC(C)CO	1.44	0.6010	328	0.15545244
37.	2-*tert*-Butylphenol	CC(C)(C)C1=CC=CC=C1O	1.66	0.7121	818	0.62935035
38.	3-iodoprop-2-ynyl N-butylcarbamate	CCCCNC(=O)OCC#CI	0.25	0.0000	1007	0.81264501
39.	4-Methyl-1,3-dioxolan-2-one	CC1COC(=O)O1	0.59	0.1717	659	0.47621810
40.	4-Methylpentan-2-one	CC(C)CC(=O)C	1.44	0.6010	473	0.29640371
41.	5-Isocyanato-1-(isocyanatomethyl)-1,3,3-trimethylcyclohexane	CC1(CC(CC(C1)(C)CN=C=O)N=C=O)C	1.25	0.5051	960	0.76682135
42.	Benzoic acid	C1=CC=C(C=C1)C(=O)O	1.11	0.4343	753	0.56670534
43.	Bis(2-methylpropyl) hexanedioate	CC(C)COC(=O)CCCCC(=O)OCC(C)C	1.30	0.5303	1003	0.80858469
44.	Butan-1-ol	CCCCO	1.34	0.5505	395	0.22041763
45.	Butan-2-one	CCC(=O)C	0.87	0.3131	328	0.15603248
46.	Butanal	CCCC=O	1.20	0.4798	319	0.14733179
47.	Butyl acetate	CCCCOC(=O)C	1.22	0.4899	535	0.35614849
48.	Butyl prop-2-enoate	CCCCOC(=O)C=C	1.36	0.5606	598	0.41705336
49.	Decane	CCCCCCCCCC	2.12	0.9444	667	0.48375870
50.	Diethyl hexanedioate	CCOC(=O)CCCCC(=O)OCC	1.06	0.4091	851	0.66183295
51.	Diphenylmethanone	C1=CC=C(C=C1)C(=O)C2=CC=CC=C2	2.04	0.9040	1012	0.81728538
52.	Dodecane	CCCCCCCCCCCC	2.10	0.9343	773	0.58584687
53.	Ethane-1,2-diol	C(CO)O	0.52	0.1364	422	0.24651972
54.	Ethanol	CCO	0.85	0.3030	193	0.02552204
55.	Ethenyl acetate	CC(=O)OC=C	0.52	0.1364	309	0.13747100
56.	Heptan-2-one	CCCCCC(=O)C	1.58	0.6717	595	0.41415313
57.	Heptane	CCCCCCC	1.94	0.8535	435	0.25928074
58.	Hexanal	CCCCCC=O	1.42	0.5909	528	0.34918794
59.	Hexane	CCCCCC	1.75	0.7576	325	0.15255220
60.	Methanol	CO	0.58	0.1667	167	0.00000000
61.	Methyl acetate	CC(=O)OC	0.55	0.1515	249	0.07946636
62.	Methyl hexadecanoate	CCCCCCCCCCCCCCCC(=O)OC	1.32	0.5404	1201	1.00000000
63.	Methyl nonanoate	CCCCCCCCC(=O)OC	1.42	0.5909	778	0.59048724
64.	*N,N*-diethylethanamine	CCN(CC)CC	1.80	0.7828	419	0.24361949
65.	*N*-butan-2-ylidenehydroxylamine	CCC(=NO)C	1.16	0.4596	504	0.32598608
66.	*N*-butylbutan-1-amine	CCCCNCCCC	1.70	0.7323	641	0.45881671
67.	*N*-methyl-sarcosinol	CN(C)CCO	0.84	0.2980	460	0.28364269
68.	Nonane	CCCCCCCCC	2.11	0.9394	604	0.42227378
69.	Octane	CCCCCCCC	2.02	0.8939	527	0.34860789
70.	Oxolane	C1CCOC1	1.11	0.4343	369	0.19547564
71.	Pentadecane	CCCCCCCCCCCCCCC	2.04	0.9040	909	0.71751740
72.	Phenylmethanol	C1=CC=C(C=C1)CO	1.66	0.7121	691	0.50696056
73.	Propan-2-ol	C[C-](C)O	0.93	0.3434	232	0.06264501
74.	Propan-2-one	CC(=O)C	0.87	0.3131	218	0.04930394
75.	Propane-1,2,3-triol	C(C(CO)O)O	0.55	0.1515	652	0.46867749
76.	Propane-1,2-diol	CC(CO)O	0.74	0.2475	467	0.29002320
77.	*tert*-Butyl acetate	CC(=O)OC(C)(C)C	1.24	0.5000	435	0.25928074
78.	Toluene	CC1=CC=CC=C1	2.17	0.9697	507	0.32888631
79.	Tridecane	CCCCCCCCCCCCC	2.10	0.9343	819	0.63051044
80.	Undecane	CCCCCCCCCCC	2.13	0.9495	722	0.53712297

FIGURE 3—Using the trained neural network to predict the RF for new compounds. A close linear fit between predicted and actual RF, along with a high R2 value (0.89), all indicate that the model has adequately learned from the dataset and is able to make reasonable predictions on new compounds.

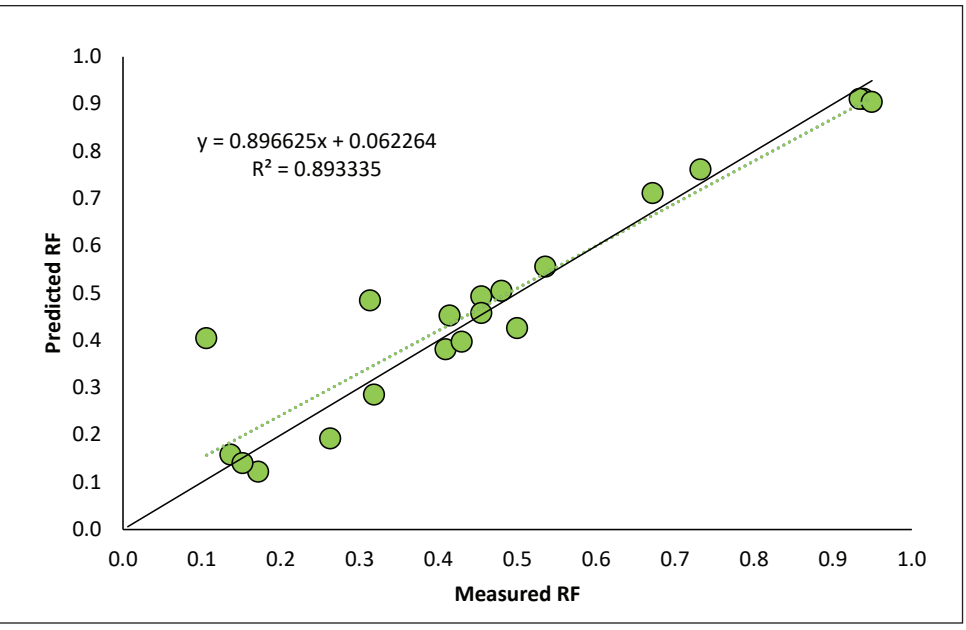

FIGURE 4—The trained artificial neural network with 19 descriptors as inputs, one hidden layer with three nodes, and one output layer (predicted RF). Image generated using the online tool NN-SVG.[9]

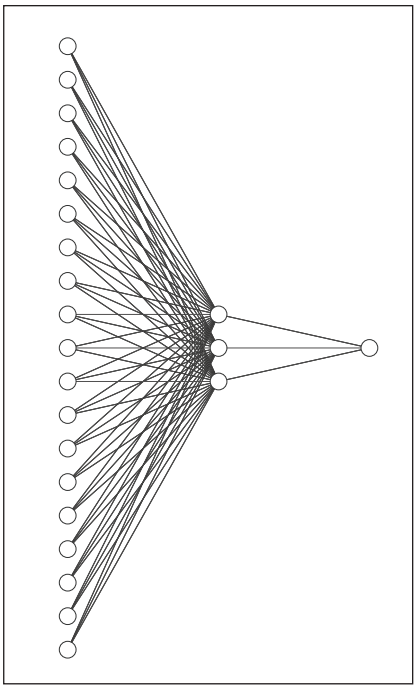

Figure 3 shows the performance of the RF prediction neural network after being trained on the dataset after 500 epochs of model evolution and after being predicting the RF of a validation set of test compounds that were not included in the training of the model.

The neural network shows good agreement during this validation stage, as indicated by the predicted and measured (experimental) values of RF. As with the training of the model in *Figure 2*, the close linear fit to the identity (x=y) line, and an R^2 value of 0.89 indicate a suitably high level of model performance and predictivity.

The final, production-ready neural network had 19 input nodes (the list of most-correlated descriptors), one hidden layer with a modest number of three perceptron or computing nodes, and a singular output—the predicted response factor. The architecture of the resultant neural network is shown in *Figure 4*.

DISCUSSION

For the first time, a quantitative structure-activity relationship approach was combined with neural networks to create a machine learning model custom-built for the performance prediction of formulation components being subjected to VOC identification and quantification.

Review of the correlated descriptors (*Table 1*) indicates that the chemical descriptor most correlated with experimental response factor is 0%, the percentage of oxygen atoms in the molecule. This confirms a commonly stated heuristic (or rule of thumb) regarding off-the-cuff estimation of response factors: as the ratio of elemental oxygen to carbon in the molecule increases, the observed response factor will decrease as the flame ionization detector will oxidize proportionally less of an oxygen-rich molecule as compared to a molecule containing less or no elemental oxygen. Implications for the design of new formulation additives based on the remaining descriptors is forthcoming in a manuscript being prepared by the authors.

The trained neural network model predicted 90.0% of the variance in the actual RF in the training set as shown in *Table 3*, with 3.89% mean squared error (MSE) and 89.3% in the validation set with 6.16% mean squared error MSE as shown in *Table 4*, indicating a high degree of accuracy and flexibility across many different chemical functionalities and a variety of GC column retention time behavior.

The neural network produced here may be implemented in predictive, digital lab workflows that are focused on in silico or virtual formulation and coating property prediction. Predictive tools that are derived from chemical structure and empirical measurements will be critical for moving research and development efforts into more accelerated, digitally enabled regimes.

As more predictive tools become available, the hope is that a common set of predictive tools may be used by both formulators as well as those concerned with the end-use properties and environmental impacts of new products as they enter the market. The goal is to bridge the divide between regulatory agencies and coatings formulators, provide a science-backed means of prediction and regulation that enables innovation, and facilitate the free market design of new products while respecting the product life cycle and the best practices of corporate stewardship.

FUTURE WORK

This first successful implementation of a neural network applied to VOC analysis workflows opens the door for further development and integration of machine learning tools for formulation research, optimization, and characterization. Current work in progress is employing similar approaches to the estimation of compound retention time and vapor pressure; both properties may be used in an inverse design, genetic algorithm-enabled workflow for the discovery of new molecular formulation components.

Future implementations of these models will be able to predict a

TABLE 3—Predicted and experimental response factors (absolute and normalized) in the training set.

Compound	RF	Pred RF	Norm RF	Norm Pred RF
	RF=Response Factor; Pred RF=Predicted Response Factor; Norm RF=Normalized Response Factor; Norm Pred RF=Normalized Predicted Response Factor			
[2,2,4-Trimethyl-1-(2-methylpropanoyloxy)pentyl] 2-methylpropanoate	1.32	1.376	0.540	0.569
1,2-Diethoxyethane	1.00	1.023	0.379	0.391
1,3,5-Trimethylbenzene	2.23	2.034	1.000	0.901
1,4-Xylene	2.20	2.006	0.985	0.887
1-Butoxybutane	1.60	1.627	0.682	0.696
1-Chloro-4-(trifluoromethyl)benzene	1.10	1.012	0.429	0.385
1-Methylpyrrolidin-2-one	0.96	0.826	0.359	0.291
1-Phenoxypropan-2-ol	1.29	1.251	0.525	0.506
2-(2-Hydroxyethoxy)ethanol	0.49	0.587	0.121	0.170
2-(2-Methoxyethoxy)ethanol	0.70	0.650	0.227	0.202
2-(2-Methoxypropoxy)propan-1-ol	0.89	0.858	0.323	0.307
2-(2-Propoxyethoxy)ethanol	0.80	0.883	0.278	0.320
2,4,7,9-Tetramethyldec-5-yne-4,7-diol	1.62	1.604	0.692	0.684
2-[2-(2-Hydroxyethoxy)ethoxy]ethanol	0.37	0.546	0.061	0.150
2-[2-(2-Methoxypropoxy)propoxy]propan-1-ol	0.92	0.812	0.338	0.284
2-[2-[2-(2-Hydroxyethoxy)ethoxy]ethoxy]ethanol	0.56	0.490	0.157	0.121
2-[2-[2-[2-(2-Hydroxyethoxy)ethoxy]ethoxy]ethoxy]ethanol	0.51	0.431	0.131	0.091
2-Amino-2-ethylpropane-1,3-diol	0.52	0.797	0.136	0.276
2-Amino-2-methylpropan-1-ol	0.85	0.976	0.303	0.366
2-Butoxyethanol	0.72	1.156	0.237	0.458
2-Ethoxyethyl acetate	1.34	0.732	0.551	0.243
2-Ethyl-2-(hydroxymethyl)propane-1,3-diol	1.05	0.935	0.404	0.346
2-Ethylhexanal	1.73	1.678	0.747	0.721
2-Methylpentane-2,4-diol	1.09	1.160	0.424	0.459
2-Methylprop-2-enoic acid	0.95	0.773	0.354	0.264
2-Methylpropan-1-ol	1.44	1.362	0.601	0.562
2-Tert-butylphenol	1.66	1.729	0.712	0.747
3-Iodoprop-2-ynyl N-butylcarbamate	0.25	0.544	0.000	0.149
4-Methylpentan-2-one	1.44	1.500	0.601	0.631
5-Isocyanato-1-(isocyanatomethyl)-1,3,3-trimethylcyclohexane	1.25	1.193	0.505	0.476
Benzoic acid	1.11	1.173	0.434	0.466
Bis(2-methylpropyl) hexanedioate	1.30	1.197	0.530	0.478
Butan-1-ol	1.34	1.359	0.551	0.560
Butyl acetate	1.22	1.139	0.490	0.449
Butyl prop-2-enoate	1.36	1.181	0.561	0.470
Decane	2.12	2.028	0.944	0.898
Diethyl hexanedioate	1.06	0.928	0.409	0.342
Diphenylmethanone	2.04	1.823	0.904	0.794
Dodecane	2.10	2.047	0.934	0.908
Ethanol	0.85	0.932	0.303	0.344
Ethenyl acetate	0.52	0.639	0.136	0.196
Heptan-2-one	1.58	1.589	0.672	0.676
Heptane	1.94	2.011	0.854	0.889
Hexanal	1.42	1.519	0.591	0.641
Hexane	1.75	1.987	0.758	0.877
Methanol	0.58	0.547	0.167	0.150
Methyl acetate	0.55	0.582	0.152	0.168
Methyl hexadecanoate	1.32	1.744	0.540	0.754
Methyl nonanoate	1.42	1.531	0.591	0.647
N,N-diethylethanamine	1.80	1.599	0.783	0.681
N-butan-2-ylidenehydroxylamine	1.16	0.903	0.460	0.330
N-methyl-sarcosinol	0.84	0.897	0.298	0.327
Octane	2.02	2.056	0.894	0.912
Oxolane	1.11	1.123	0.434	0.441
Pentadecane	2.04	2.054	0.904	0.911
Phenylmethanol	1.66	1.501	0.712	0.632
Propan-2-ol	0.93	1.001	0.343	0.380
Propan-2-one	0.87	0.975	0.313	0.366
Propane-1,2-diol	0.74	0.770	0.247	0.262
Toluene	2.17	1.970	0.970	0.869

TABLE 4—Predicted and experimental response factors (absolute and normalized) in the validation set.

Compound	RF	Pred RF	Norm RF	Pred Norm RF
	RF=Response Factor; Pred RF=Predicted Response Factor; Norm RF=Normalized Response Factor; Pred Norm RF= Predicted Normalized Response Factor			
(3-Hydroxy-2,2,4-trimethylpentyl) 2-methylpropanoate	1.31	1.39	0.535	0.555
1-[2-(2-Methoxypropoxy)propoxy]butane	1.07	1.21	0.414	0.452
1-Butoxypropan-2-ol	1.15	1.28	0.455	0.493
1-Methoxy-2-[2-(2-methoxyethoxy)ethoxy]ethane	0.77	0.78	0.263	0.192
1-Methoxypropyl acetate	0.88	0.94	0.318	0.285
2-(2-Butoxyethoxy)ethanol	1.06	1.10	0.409	0.380
2-[Butyl(2-hydroxyethyl)amino]ethanol	1.10	1.12	0.429	0.396
2-Benzofuran-1,3-dione	0.46	1.13	0.106	0.404
2-Butoxyethanol	1.15	1.22	0.455	0.458
2-Ethylhexyl benzoate	1.58	1.65	0.672	0.711
4-Methyl-1,3-dioxolan-2-one	0.59	0.66	0.172	0.122
Butan-2-one	0.87	1.27	0.313	0.484
Butanal	1.20	1.30	0.480	0.504
Ethane-1,2-diol	0.52	0.72	0.136	0.158
N-butylbutan-1-amine	1.70	1.73	0.732	0.761
Nonane	2.11	1.98	0.939	0.910
Propane-1,2,3-triol	0.55	0.69	0.152	0.140
t-Butyl acetate	1.24	1.17	0.500	0.426
Tridecane	2.10	1.98	0.934	0.910
Undecane	2.13	1.97	0.949	0.903

quantified VOC content profile for a proposed formulation, even if certain components of the formulation are novel and have not been fully characterized by laboratory methods. This computational approach to additive and formulation design will assist the next generation of formulation scientists tasked with quickly and efficiently formulating and optimizing environmentally benign high-performance coatings. ❊

ACKNOWLEDGMENTS

The authors would like to thank Professor Dane Jones for countless fruitful and informative discussions about the field of VOC quantification.

References

1. Protection of Environment. Title 40, Chapter 1, Subchapter C, Part 51, Subpart F, 51.100. *Code of Federal Regulations*. *https://www.ecfr.gov/* (accessed Feb 28, 2022).
2. Method 313—Determination of Volatile Organic Compounds (VOC) by Gas Chromatography/ Mass Spectrometry/ Flame Ionization Detection. South Coast AQMD. *www.aqmd.gov/docs/default-source/laboratory-procedures/methods-procedures/313-91.pdf* (accessed Mar 2, 2022).
3. ASTM D6886: Standard Test Method for Determination of the Weight Percent Individual Volatile Organic Compounds in Waterborne Air-Dry Coatings by Gas Chromatography.
4. ISO 11890-2:2020. Paints and Varnishes—Determination of Volatile Organic Compounds (VOC) and/or Semi Volatile Organic Compounds (SVOC) Content—Part 2: Gas-Chromatographic Method.
5. Todeschini, R., and Consonni, V. *Molecular Descriptors for Chemoinformatics*, 2nd ed., Wiley-VCH, Weinheim, Germany, 2009.
6. Hanwell, M.D., Curtis, D.E., Lonie, D.C., Vandermeersch, T., Zurek, E., and Hutchison, G.R. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics*, **2012,** 4:17.
7. Dragon 7.0, Kode Chemoinformatics srl, Pisa, Italy.
8. Chollet, F. *Deep Learning with Python*, Manning, Shelter Island, New York, 2018.
9. LeNail, A. NN-SVG: Publication-Ready NN-Architecture Schematics. *https://alexlenail.me/NN-SVG/index.html* (accessed Feb 28, 2022).

JESSICA LUM and **MADELINE SCHULTZ** are recent graduates of the Department of Chemistry and Biochemistry at California Polytechnic State University, San Luis Obispo, CA 93407.
ERIK SAPPER is an assistant professor in the same department. Email: esapper@calpoly.edu.