SUPPORTING INFORMATION

Data-Driven, Explainable Machine Learning Model for Predicting Volatile Organic Compounds' Standard Vaporization Enthalpy

José Ferraz-Caetano[1], Filipe Teixeira[2], M. Natália D. S. Cordeiro[1]

1) LAQV-REQUIMTE – Department of Chemistry and Biochemistry – Faculty of Sciences, University of Porto - Rua do Campo Alegre, S/N, 4169-007 Porto, Portugal
2) CQUM – Centre of Chemistry, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

KEYWORDS: Machine Learning; Supervised Learning; Standard Vaporization Enthalpy; Thermochemical prediction

INDEX

Annex S1 – Repository and Code

Table S1

List of links for the model's database and developed code.

Model	Website
Vaporization Enthalpy Database	https://gitfront.io/r/ESA-Vanadium/HdCi2vFs5msj/VOC-EnthVapML/
Model Development, Statistical Results and External Database[1]	https://gitfront.io/r/ESA-Vanadium/HdCi2vFs5msj/VOC-EnthVapML/

Statistical definitions

- Mathematical definition of the average absolute relative deviation (AARD), root mean square error (RMSE) and mean absolute error (MAE):

$$AARD = \frac{100}{N}\sum_{i}^{N}\frac{|prediction(i) - literature(i)|}{literature(i)}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(prediction(i) - literature(i)\right)^2}{N}}$$

$$MAE = \frac{\sum_{i=0}^{N}|(prediction(i) - literature(i)|}{N}$$

Annex S2 – Model Database Description

The list of compounds used in the model's database and the VOC database[1] validation are presented in the Online Repository in Annex 1. In Figure S1 and S2, we present a graphical representation of different chemical families presented in the model database sorted by the overall number of entries.

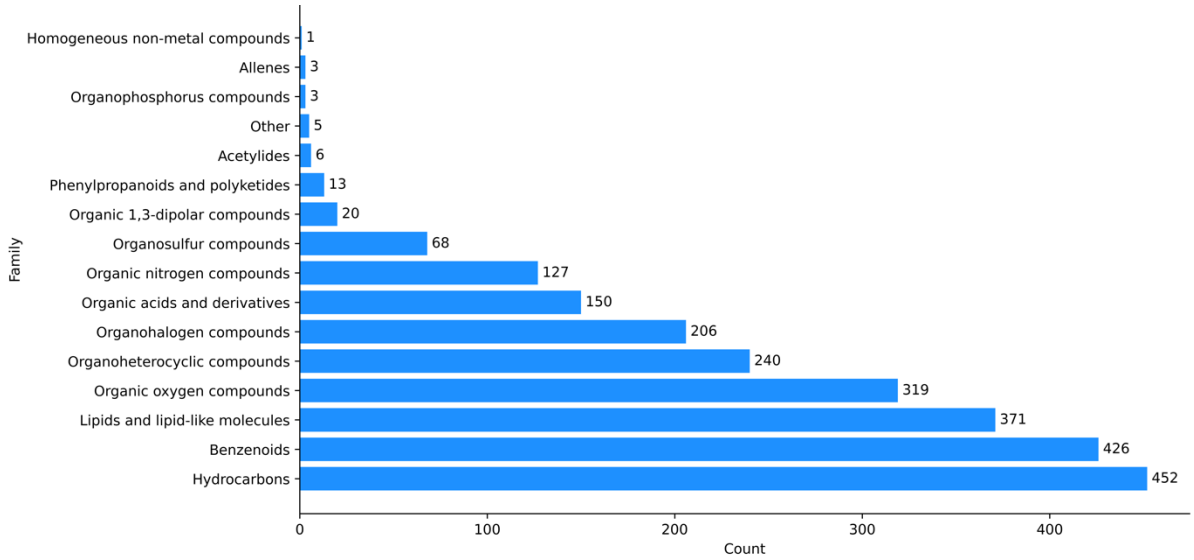

Figure S1. Graphical representation of entries in the Model Database, sorted by molecular families.

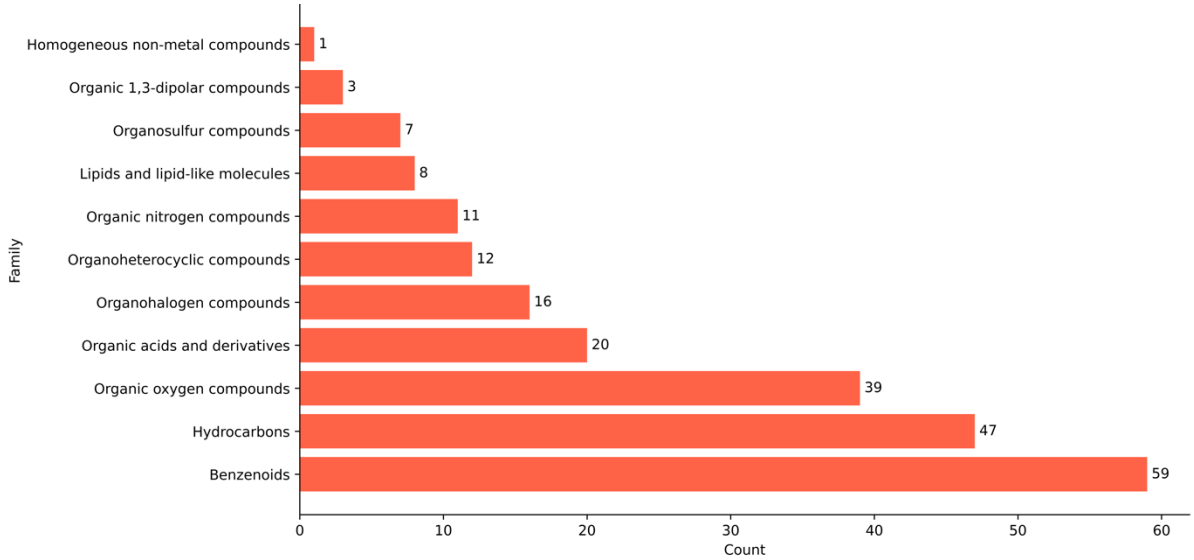

Figure S2. Graphical representation of entries in the VOC Database, sorted by molecular families.

Annex S3 – RDKit Full Descriptor List

Descriptors used in Machine Learning model development (Table S2). The respective bibliographical foundation of each descriptor is described in the RDKit WebBook Documentation[2]. The descriptor calculations were made by converting each SMILES string (representing each molecular entry), running on top of Python version 3.9[3] and using the RDKit package, version 2022.09.4[4].

Table S2

List of Descriptors Used in Machine Learning model development

Descriptor	Type	Meaning
BalabanJ	Structural	Distance sum of the two end-vertex for each edge. BalabanJ index has been proven to be relevant to network branching.
BertzCT	Structural	Complexity index, considering both the variety of kinds of bond connectivities and atom types; information contents related to bond connectivity and atom type diversity.
0χ , 1χ	Structural	This descriptor signifies a retention index (zero order) derived directly from gradient retention times.
0χn – 4χn	Structural	This descriptor signifies a retention index (zero order) derived directly from gradient retention times.
0χv – 4χv	Structural	This descriptor signifies atomic valence connectivity index (order 0).
VSA_EState1 – VSA_EState11	Surface Area	MOE-type (QSAR model) descriptors using EState indices and surface area contributions.

Hall Kier α	Structural	Descriptor fingerprint that displays the difference between active and inactive molecules.
HeavyAtomCount	Structural	The number of heavy atoms in the molecule.
HeavyAtomMolWt	Structural	The average molecular weight of the molecule ignoring hydrogens.
κ_1 , κ_2 , κ_3	Structural	This descriptor signifies # κ shape index: $(n-1) \times 2 / m^2$
Max and Min AbsEStateIndex	Electronic	Maximum and minimum Absolute E-State
Max and Min AbsPartialCharge	Electronic	Maximum and minimum Absolute Partial Charge
Max and Min EStateIndex	Electronic	Maximum and minimum E-State
Max and Min PartialCharge	Electronic	Maximum and minimum Partial Charge
Mol logP	Structural	Wildman-Crippen logP value.
MolMR	Electronic	Wildman-Crippen molar refractivity.
MolWt	Structural	The average molecular weight of the molecule.
NHOH Count	Structural	The number of NHs or OHs.
NO Count	Structural	The number of Nitrogens and Oxygens.
$n_{alicarb}$	Structural	The number of aliphatic carbocycles.
n_{alihet}	Structural	The number of aliphatic heterocycles.
n_{alirig}	Structural	The number of aliphatic rings.

$n_{arocarb}$	Structural	The number of aromatic carbocycles.
n_{arohet}	Structural	The number of aromatic heterocycles.
n_{arorig}	Structural	The number of aromatic rings.
n_{Ha}	Structural	The number of Hydrogen Bond Acceptors.
n_{Hd}	Structural	The number of Hydrogen Bond Donors.
n_{het}	Structural	The number of Heteroatoms.
n_{radele}	Structural	The number of radical electrons.
n_{rot}	Structural	The number of Rotatable Bonds.
n_{satrig}	Structural	The number of saturated rings.
n_{ele}	Structural	The number of valence electrons.
PEOE_VSA1 – PEOE_VSA14	Surface Area	MOE-type (QSAR model) descriptors using partial charges and surface area contributions.
SMR_VSA1 – SMR_VSA10	Surface Area	MOE-type (QSAR model) descriptors using MR contributions and surface area contributions.
SlogP_VSA1 – SlogP_VSA12	Surface Area	MOE-type (QSAR model) descriptors using logP contributions and surface area contributions.
TPSA	Surface Area	The total polar surface area of a molecule based upon fragment calculations.
VSA_EState1 – VSA_EState10	Surface Area	MOE-type (QSAR model) descriptors using EState indices and surface area contributions.

Annex S4 – ML Model Development

Initial model setup for $\Delta_{vap}H_m°$ prediction was assessed with increasing training loads to determine a train/test split. Considering the trade-off between result accuracy, model over-fitting and data representability, we have selected a 60:40 train/test split for model optimization. We thus calculated the statistical performance of the Random Forest Regressor, Gradient Boosting Regressor, Support Vector Machines and MLP Neural Network algorithms, comparing with Linear Regressor, as presented in Table S3.

Table S3

Statistical results for model performance for each algorithm for a 60:40 train/test split: prediction scores, mean absolute error, average absolute relative deviation, random mean square error and coefficient of determination.

Regressor Algorithm	R^2 Test	R^2 Train	Score Test	MAE / kJ mol^{-1}	RMSE / kJ mol^{-1}	AARD / %
Random Forest	0.969 ± 0.004	0.9951 ± 0.0007	0.968 ± 0.005	3.5 ± 0.2	5.8 ± 0.4	5.8 ± 0.2
Gradient Boosting	0.969 ± 0.005	0.988 ± 0.001	0.968 ± 0.005	3.8 ± 0.1	5.8 ± 0.4	6.3 ± 0.2
MLP Neural Network	0.96 ± 0.02	0.98 ± 0.02	0.95 ± 0.03	3.7 ± 0.5	6.8 ± 2.0	6.7 ± 1.9
Linear Regression	0.93 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	6.1 ± 0.3	8.9 ± 0.5	10.3 ± 0.4
SVM	0.55 ± 0.05	0.54 ± 0.03	0.51 ± 0.05	10.0 ± 0.5	22.1 ± 2.5	16.4 ± 0.7

The RF, GB, MLP algorithms compare favorably in overall accuracy with reported models for thermodynamic predictions, which average MAE values[5-7] under 4 kJ mol^{-1} (even models using a single family dataset[8]). Our results strongly suggest that all three models benefit from a wide and diverse set of chemical species, while RF, GB and MLP present more accurate predictions than the SVM algorithm and linear regression.

Considering our best supervised algorithms (GB, RG and MLP), we argue a positive outlook on further model development using these three algorithms. Key difference between a Neural Network and our best supervised algorithms stems from a wide variation of RMSE, calculated with a high prediction error, not compatible with sustained recurrent calculations. Hence a sustained perspective on model optimization using the ensemble algorithms.

Annex S5 – Model Validation

To evaluate the wide scope of this model, we have predicted the standard vaporization enthalpy of a known database used for non-ML prediction methods[1]. This external database is composed of several hydrocarbon-based molecules, not drawn from our original model database. Table S4 and Figure S10 display an evaluation of our model against previously reported studies.

Table S4

Statistical results for model performance of each optimized algorithm for $\Delta_{vap}H_m°$ prediction of the external dataset.

Regressor Algorithm	R^2 Test	MAE / kJ mol^{-1}	RMSE / kJ mol^{-1}	AARD / %
Random Forest	0.964	2.815	3.860	4.748
Gradient Boosting	0.950	3.531	4.643	6.030
Linear Regressor	0.965	2.144	2.700	4.240
Support Vector Machine	0.953	2.669	3.567	4.889
MLP Neural Network	**0.982**	**1.379**	**2.124**	**2.480**

Table S5

Ranking of method performance for $\Delta_{vap}H_m°$ predictions across the same external dataset.

Model	AARD / %
Guthrie and Taylor[11]	10.6
Ducros[12]	7.0
Domalski and Hearing[13]	6.4
Constantinou and Gani[14]	5.5
Chickos[15]	4.6
Kolská[16]	4.1
Gharagheizi[1]	3.1
Santos and Leal[17]	2.6
Our ML Model	**2.5**

Our ML model has predicted the standard vaporization enthalpies of the selected molecules in the database with an AARD in line with the best previous attempts. The low AARD displayed by our model is even more significant if we consider the background of the compared studies. For example, more accurate models can be only applied for predictions of hydrocarbons (as Santos and Leal's model), which relinquishes any attempt to apply such model to any generic

molecule. Also, some models only work if specific types of chemical descriptors are known (such as Abraham's solvation features), or depend on functional/fragment group values, while our proposed model has a faster, wider scope of chemical diversity in the original dataset. For example, while Gharagheizi presents one of the most diverse initial datasets, it does not include representative counts of key chemical families like Lipids, Salts, Organometallic and Organic acid compounds, as we include in our database.

Annex S6 – ML VOC Model Optimization

Prior to model development we evaluated each tested algorithms performance for predicting VOCs $\Delta_{vap}H_m°$. We present the results in Table S6, where we see that the RF yielded the best results. We then used a pipeline strategy to automatically select and tune each algorithms' best features for $\Delta_{vap}H_m°$ determinations, scoring with the lowest prediction MAE possible. Sample code is presented in the Annex S1. Using the selected hyper-parameters, the model is optimized by removing offsetting descriptors, which do not contribute positively towards model prediction (PI > 0%). Our models were recalculated accordingly and statistical results with value evaluations from the firsts iterations are presented in the manuscript's Table 1.

Table S6

Statistical results for model performance of each optimized algorithm for VOC $\Delta_{vap}H_m°$ prediction.

Regressor Algorithm	R^2 Test	MAE / kJ mol^{-1}	RMSE / kJ mol^{-1}	AARD / %
Random Forest	**0.946**	**3.072**	**4.785**	**6.362**
Gradient Boosting	0.943	3.754	5.116	7.885
Linear Regressor	0.933	3.636	5.315	8.049
Support Vector Machine	0.651	8.005	12.085	17.561
MLP Neural Network	0.937	3.167	5.286	6.322

List of optimal hyperparameters for RF algorithm for model optimization:

Random Forest Regressor : n_estimators= 300, min_samples_split= 2, min_samples_leaf= 1, max_features='sqrt', max_depth= 20, bootstrap= False, random_state=47

Bibliography

1. Gharagheizi, F.; Ilani-Kashkouli, P.; Acree, W. E.; Mohammadi, A. H.; Ramjugernath, D., A group contribution model for determining the vaporization enthalpy of organic compounds at the standard reference temperature of 298K. *Fluid Phase Equilibria* **2013,** *360*, 279-292.

2. RDKit: Open-source cheminformatics - Descritptor Webbook. https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors (accessed March 1, 2022).

3. Python Software Foundation - Python Language Reference, version 3.9.8. http://www.python.org.

4. Landrum, G. RDKit: Open-source cheminformatics 2022_09_4 (Q3 2022) Release - January 16, 2023. http://www.rdkit.org/ (accessed January 18, 2023).

5. Dobbelaere, M. R.; Plehiers, P. P.; Van de Vijver, R.; Stevens, C. V.; Van Geem, K. M., Learning Molecular Representations for Thermochemistry Prediction of Cyclic Hydrocarbons and Oxygenates. *The Journal of Physical Chemistry A* **2021,** *125* (23), 5166-5179.

6. Yalamanchi, K. K.; van Oudenhoven, V. C. O.; Tutino, F.; Monge-Palacios, M.; Alshehri, A.; Gao, X.; Sarathy, S. M., Machine Learning To Predict Standard Enthalpy of Formation of Hydrocarbons. *The Journal of Physical Chemistry A* **2019,** *123* (38), 8305-8313.

7. Aldosari, M. N.; Yalamanchi, K. K.; Gao, X.; Sarathy, S. M., Predicting entropy and heat capacity of hydrocarbons using machine learning. *Energy and AI* **2021,** *4*, 100054.

8. Benson, S. W.; Buss, J. H., Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *The Journal of Chemical Physics* **1958,** *29* (3), 546-572.

9. Churchill, B.; Acree, W. E.; Abraham, M. H., Development of Abraham model expressions for predicting the standard molar enthalpies of vaporization of organic compounds at 298.15 K. *Thermochimica Acta* **2019,** *681*, 178372.

10. Low, K.; Coote, M. L.; Izgorodina, E. I., Explainable Solvation Free Energy Prediction Combining Graph Neural Networks with Chemical Intuition. *Journal of Chemical Information and Modeling* **2022,** *62* (22), 5457-5470.

11. Guthrie, J. P.; Taylor, K. F., Additivity methods for estimating heats of vaporization of organic compounds. *Canadian Journal of Chemistry* **1983,** *61* (3), 602-607.

12. Ducros, M.; Gruson, J. F.; Sannier, H., Estimation des enthalpies de vaporisation des composes organiques liquides. Partie 1. Applications aux alcanes, cycloalcanes, alcenes, hydrocarbures benzeniques, alcools, alcanes thiols, chloro et bromoalcanes, nitriles, esters, acides et aldehydes. *Thermochimica Acta* **1980,** *36* (1), 39-65.

13. Domalski, E. S.; Hearing, E. D., Estimation of the Thermodynamic Properties of C-H-N-O-S-Halogen Compounds at 298.15 K. *Journal of Physical and Chemical Reference Data* **1993,** *22* (4), 805-1159.

14. Constantinou, L.; Gani, R., New group contribution method for estimating properties of pure compounds. *AIChE Journal* **1994,** *40* (10), 1697-1710.

15. Chickos, J. S.; Hesse, D. G.; Liebman, J. F., Estimating vaporization enthalpies of organic compounds with single and multiple substitution. *The Journal of Organic Chemistry* **1989,** *54* (22), 5250-5256.

16. Kolská, Z.; Růžička, V.; Gani, R., Estimation of the Enthalpy of Vaporization and the Entropy of Vaporization for Pure Organic Compounds at 298.15 K and at Normal Boiling Temperature by a Group Contribution Method. *Industrial & Engineering Chemistry Research* **2005,** *44* (22), 8436-8454.

17. Santos, R. C.; Leal, J. P., A Review on Prediction Methods for Molar Enthalpies of Vaporization of Hydrocarbons: The ELBA Method as the Best Answer. *Journal of Physical and Chemical Reference Data* **2012,** *41* (4), 043101.