Supporting Information: Predicting Pesticide Vapour Pressures: The Power of Functional Groups

Mark Heezen,[†,‡] Manuel Alcamí,[*,†] Clemens Rauer,[†] and Freija De Vleeschouwer[*,‡]

[†]*Departamento de Química, C-13, Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain*

[‡]*Department of General Chemistry: Algemene Chemie (ALGC), Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium*

E-mail: manuel.alcami@uam.es; Freija.De.Vleeschouwer@vub.be

Contents

Pearson coefficient of QM properties

Table S1: Pearson coefficient of the correlation between the quantum mechanically calculated property and the vapour pressure

QM property	Pearson coefficient
Gibbs free energy (w)	0.2889
Entropy (w)	0.2837
Zero point energy (w)	-0.4728
Highest occupied molecular orbital (w)	0.0903
Lowest unoccupied molecular orbital (w)	0.2855
Dipole moment (w)	-0.4893
Dipole moment (g)	-0.4354
Isotropic polarisability (w)	-0.6882
Rotational constant 1 (w)	0.3667
Rotational constant 2 (w)	0.3914
Frequency 1 (w)	0.2203
Frequency 1 (o)	0.1922
Frequency 1 (g)	0.1836
HOMO-LUMO gap (g)	0.2682
ΔGibbs free energy (wg)	0.8550
ΔEntropy (wg)	-0.0204
ΔHeat capacity at constant volume (wg)	-0.3774
ΔZero point energy (wg)	0.5072
ΔHighest occupied molecular orbital (wg)	0.0017
ΔLowest unoccupied molecular orbital (wg)	-0.0655
ΔDipole moment (wg)	-0.3760
ΔNuclear repulsion energy (wg)	-0.0296

Table S1: (continued)

QM property	Pearson coefficient
ΔRotational constant 1 (wg)	-0.0258
ΔRotational constant 2 (wg)	-0.0470
ΔRotational constant 3 (wg)	-0.1033
ΔFrequency 1 (wg)	0.0274
ΔEntropy (og)	-0.0432
ΔHeat capacity at constant volume (og)	-0.3343
ΔElectronic single point energy (og)	0.8300
ΔDipole moment (og)	-0.3408
ΔNuclear repulsion energy (og)	0.1864
ΔRotational constant 1 (og)	-0.0178
ΔRotational constant 2 (og)	0.0036
ΔRotational constant 3 (og)	-0.0261
ΔFrequency 1 (og)	-0.0003
ΔGibbs free energy (wo)	0.5356
ΔEntropy (wo)	0.0209
ΔHeat capacity at constant volume (wo)	-0.1315
ΔZero point energy (wo)	0.2956
ΔHighest occupied molecular orbital (wo)	0.0500
ΔLowest unoccupied molecular orbital (wo)	0.0000
ΔDipole moment (wo)	-0.2089
ΔNuclear repulsion energy (wo)	-0.2653
ΔRotational constant 1 (wo)	-0.0160
ΔRotational constant 2 (wo)	-0.0750
ΔRotational constant 3 (wo)	-0.1210

Table S1: (continued)

QM property	Pearson coefficient
ΔFrequency 1 (wo)	0.0431
ΔCavity surface area (wo)	0.3180
ΔCavity volume (wo)	0.6314

Functional group clustering

The list of functional groups on Checkmol's website[S1] shows dependencies in the functional groups present. For example, the columns *alcohol*, *primary alcohol*, *secondary alcohol*, and *tertiary alcohol* are present. In this case, the column *alcohol* is the sum of the other three mentioned columns. We looked for these dependencies by hand, since they are not described in the Checkmol documentation. In Checkmol some dependencies could be spotted by functional groups that end in *deriv.*, short for derivative, as part of the group name. All the relationships found between the functional groups present in our database are shown in Table S2. Based on our chemical intuition and taking the trade-off between detail in the model and number of features into account, we decided to either cluster the groups together to the main group or drop the main group and keep the subcategories. This is indicated in the clustered column in Table S2 by *yes* or *no*. In some cases, a functional group appears both as a main category and a subcategory. When marked *yes* for both, this functional group is excluded from the final database and is listed only as part of the overarching functional group to prevent double counting.

After the dependencies have been corrected, we checked whether identical columns were present. If this was the case, one of these columns was dropped. The other column got a name consisting of both columns. This is for example the case for the column that looks into charged functional groups. Since all pesticides are in their neutral form in the database, the column *cation* and *anion* are always equal and therefore merged.

Table S2: Dependencies in the functional groups that are found in the SEPIA database. The second column is made by summing the categories mentioned in the third column. The cluster column indicates whether the subcategories are kept (*no* clustering) or the main category is kept (clustering, so *yes*).

Clustered	Main category	Sub categories
yes	carbonyl compound	aldehyde, ketone
no	hydroxy	alcohol, phenol or hydroxyheteroarene
yes	alcohol	primary alcohol, secondary alcohol, tertiary alcohol, 1,2-diol, 1,2-aminoalcohol
no	ether	dialkyl ether, alkyl aryl ether, diaryl ether
no	amine	primary amine, secondary amine, tertiary amine
yes	primary amine	primary aliphatic amine (alkylamine), primary aromatic amine
yes	secondary amine	secondary amine, secondary aliphatic amine (dialkylamine), secondary aliphatic/aromatic amine (alkylarylamine), secondary aromatic amine (diarylamine)
yes	tertiary amine	tertiary aliphatic amine (trialkylamine), tertiary aliphatic/aromatic amine (alkylarylamine)
no	halogen derivative	alkyl fluoride, alkyl chloride, alkyl bromide, aryl fluoride, aryl chloride, aryl bromide, aryl iodide
yes	carboxylic acid derivative	carboxylic acid, carboxylic acid ester, carboxylic acid amide, carboxylic acid hydrazide, carboxylic acid amidine, carboxylic acid imide (N-substituted), carboxylic acid imide (N-unsubstituted)
yes	carboxylic acid amide	primary carboxylic acid amide, secondary carboxylic acid amide, tertiary carboxylic acid amide
yes	thiocarboxylic acid derivative	thiocarboxylic acid ester, thiolactone
yes	sulfuric acid derivative	sulfuric acid amide ester, sulfuric acid diamide
yes	sulfonic acid derivative	sulfonic acid ester, sulfonamide, sulfone

Hyperparameters

Table S3: Tuned hyperparameters of the generated krr models.

Feature	Molecular properties	Functional groups
Kernel	polynomial	polynomial
degree	1	3
α	0.24574538367297724	0.613569814052685
γ	9.996760120209307	0.10104775081786264
coef0	0.10062930271486187	1.4889769857045507

Table S4: Tuned hyperparameters of the generated XGboost models.

Feature	Molecular properties	Functional groups
Estimators	385	135
Maximum depth	2	6
Learning rate	0.09990673202049342	0.2183777170191625
Subsample	0.8276085028744941	0.9582867426167475
Colsample by tree	0.8917249336219539	0.5868131315729668
Reg alpha	0.2178445187703823	0.8183085518481774
Reg lambda	1.436398222412372	2.7960946484534035

XGBoost SHAP

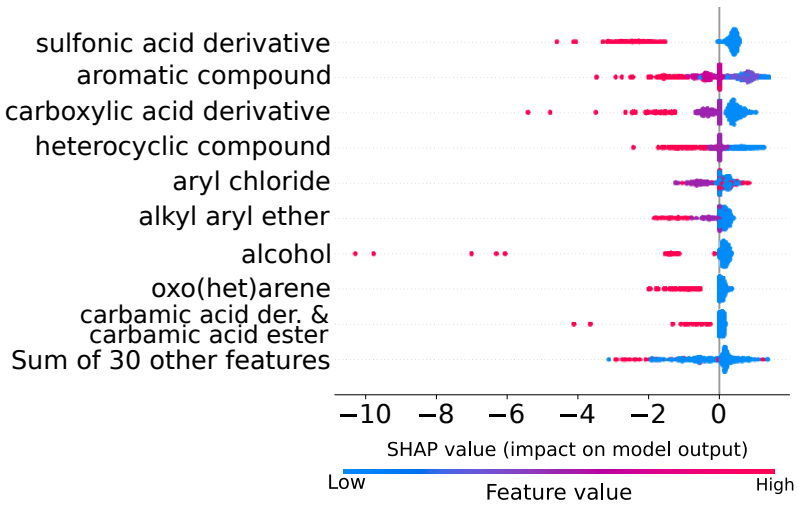


Figure S1: SHAP beeswarm plot for the XGBoost model based on 29 functional groups

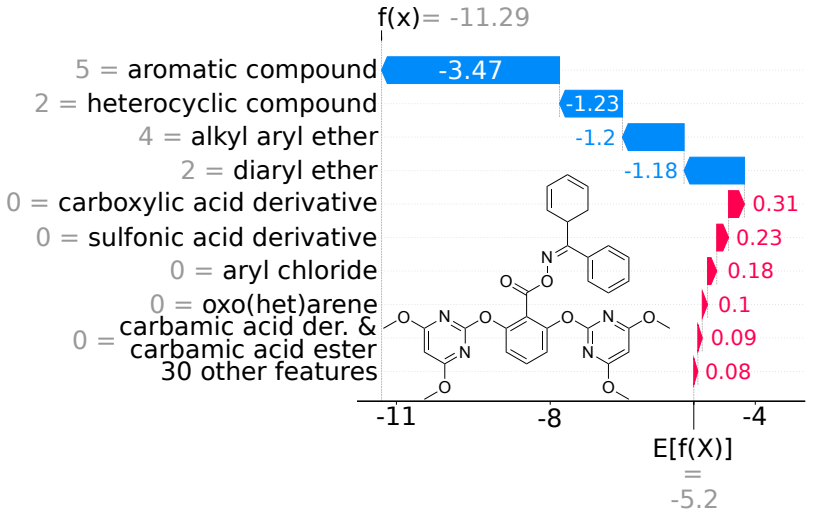
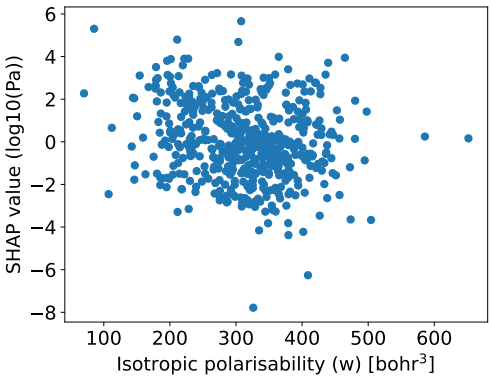
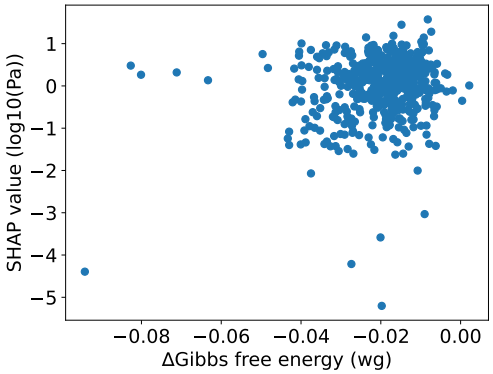


Figure S2: SHAP waterfall plot for pyribenzoxim from the XGBoost model based on 29 functional groups

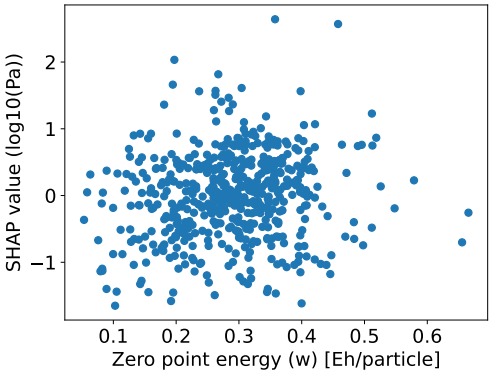
Correlation between QM properties and SHAP values



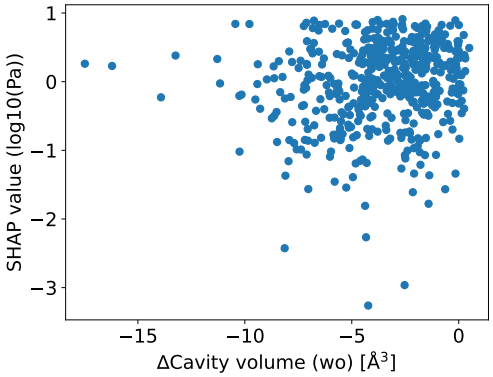
(a) Isotropic polarisability (w)



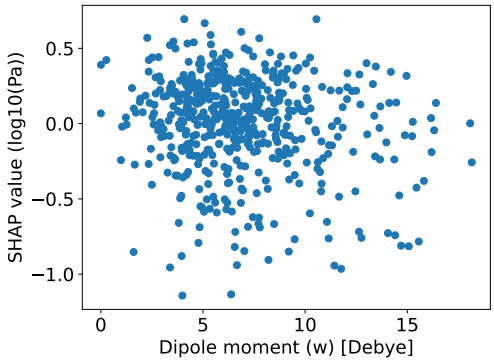
(b) ΔGibbs free energy (wg)



(c) Zero point energy (w)



(d) ΔCavity volume (wo)



(e) Dipole moment (w)

Figure S3: The correlation between the value of the property and its respective SHAP value for the 5 most important QM properties. G, w, and o indicate if the property was calculated without solvation or solvated in water or octanol, respectively.

Functional group slopes

Table S5: Slopes for the number of a particular functional groups present vs the SHAP value based on the training set for the krr model

Functional group	slope	R^2	max. FG	Δ SHAP
Cation & Anion	-0.648373668	0.894	2	1.395
Carbonyl ompound	-0.874390722	0.828	2	3.294
Oxime ether	-0.887470008	0.884	4	3.320
Ketene acetal or derivative	-1.261285091	0.959	1	1.527
Acetal	-0.846935919	0.555	2	3.016
Hemiaminal	-0.863020671	0.824	2	2.504
Enamine	-0.342168544	0.775	4	1.397
Enol	-1.967633684	0.993	2	4.214
Enol ether	-0.523128315	0.988	1	0.634
Alcohol	-0.772416646	0.947	18	14.245
Phenol or Hydroxyheteroarene	0.544151528	0.619	2	2.610
Dialkyl ether	-0.173065397	0.239	2	1.345
Alkyl aryl ether	-0.389781294	0.626	4	2.489
Diaryl ether	-0.450452128	0.754	2	1.799
Hydrazine derivative	-0.892741785	0.905	1	1.574
Primary amine	-2.015409203	0.896	2	6.578
Tertiary amine	0.317113242	0.628	1	0.997
Alkyl fluoride	-0.453115602	0.738	2	2.409
Alkyl chloride	-0.648718161	0.914	6	4.792
Aryl fluoride	0.189519368	0.528	4	1.434
Aryl chloride	-0.050947804	0.029	6	3.657
Aryl bromide	0.166433277	0.279	2	1.012

Table S5: (continued)

Functional group	slope	R^2	max. FG	Δ SHAP
Carboxylic acid derivative	-1.361323191	0.859	5	12.001
Lactam	0.315500889	0.609	1	0.908
Carbonitrile	-1.334673399	0.882	2	4.371
Oxo(het)arene	-0.73534747	0.812	2	2.147
Thioxo(het)arene	0.057822369	0.028	2	1.586
Orthocarboxylic acid derivative	0.255828151	0.280	2	1.487
Carbamic acid derivative & Carbamic acid ester (Urethane)	-1.410363777	0.865	2	4.426
Thiocarbamic acid derivative	-0.207981271	0.733	1	0.449
Thiocarbamic acid	0.370398128	0.876	1	0.721
Nitro compound	-0.719626005	0.877	2	2.515
Sulfuric acid derivative	-2.204580328	0.953	1	3.087
Sulfonic acid derivative	-3.342756898	0.978	2	8.683
Thiophosphoric acid derivative & Thiophosphoric acid ester	-1.349203033	0.964	2	3.607
Alkene	-0.428175965	0.390	2	2.963
Alkyne	-0.262078189	0.789	2	0.844
Aromatic compound	-1.469477984	0.878	5	15.546
Heterocyclic compound	-0.767741679	0.816	3	4.573

Table S6: Slopes for the number of a particular functional groups present vs the SHAP value based on the training set for the XGBoost model

Functional group	slope	R^2	max. FG	Δ SHAP
Cation & Anion	-0.098534500	0.205	2	0.709
Carbonyl ompound	-0.785091004	0.662	2	2.031
Oxime ether	-0.319649505	0.580	4	1.392
Ketene acetal or derivative	-0.856720000	0.993	1	1.011
Acetal	-0.106090702	0.184	2	1.228
Hemiaminal	-0.236111525	0.583	2	0.701
Enamine	0.144524181	0.765	4	0.577
Enol	-0.318487868	0.637	2	0.677
Enol ether	-0.335051043	0.963	1	0.447
Alcohol	-0.743179220	0.881	18	10.564
Phenol or Hydroxyheteroarene	0.101097077	0.426	2	0.382
Dialkyl ether	0.019990731	0.005	2	1.505
Alkyl aryl ether	-0.477384318	0.656	4	2.250
Diaryl ether	-0.140197415	0.176	2	1.647
Hydrazine derivative	-0.736427746	0.833	1	1.169
Primary amine	-1.143238922	0.909	2	2.524
Tertiary amine	0.383752920	0.885	1	0.672
Alkyl fluoride	-0.494573027	0.935	2	1.225
Alkyl chloride	-0.089784324	0.371	6	1.141
Aryl fluoride	0.176252119	0.515	4	1.492
Aryl chloride	-0.128231275	0.097	6	2.076
Aryl bromide	0.030147923	0.160	2	0.224
Carboxylic acid derivative	-0.953588977	0.815	5	6.313

Table S6: (continued)

Functional group	slope	R^2	max. FG	Δ SHAP
Lactam	0.218086651	0.565	1	0.446
Carbonitrile	-0.713590181	0.789	2	1.491
Oxo(het)arene	-0.804565939	0.780	2	2.197
Thioxo(het)arene	0.039256823	0.020	2	1.105
Orthocarboxylic acid derivative	0.281045382	0.483	2	0.939
Carbamic acid derivative & Carbamic acid ester (Urethane)	-1.057390689	0.637	2	4.274
Thiocarbamic acid derivative	0.283824509	0.731	1	0.503
Thiocarbamic acid	0.236863882	0.783	1	0.643
Nitro compound	-0.306187744	0.680	2	1.110
Sulfuric acid derivative	-1.608049591	0.949	1	1.942
Sulfonic acid derivative	-2.740670903	0.968	2	5.165
Thiophosphoric acid derivative & Thiophosphoric acid ester	-0.677722118	0.827	2	2.028
Alkene	-0.136939744	0.280	2	1.288
Alkyne	0.010914247	0.081	2	0.121
Aromatic compound	-0.770697016	0.691	5	4.865
Heterocyclic compound	-0.617153813	0.704	3	3.692

References

(S1) Haider, N. Checkmol/Matchmol Homepage. 2018; https://homepage.univie.ac.at/norbert.haider/cheminf/cmmm.html.