CAIRRE Supporting Information: Description of Features

Burns, T.D.*; Inglis, C.; Beking, M.; Shen, J; Provost, N; Tigner, J

November 3, 2022

Contents

S1 List of CAIRRE_v1 Features

S1.1 Classification Model

Table S1: Table showing features used by the CAIRRE_v1 classification model, with the feature labels used by CAIRRE, the software used to generate the feature, and the DOI or ISBN of the reference for that feature.

Feature Label	Feature Description	Type	Software	DOI/ISBN
BP_pred	Boiling point	Predicted (kNN)	OPERA	https://doi.org/10.1186/s13321-018-0263-1
LogVP_pred	Log (vapour pressure)	Predicted (kNN)	OPERA	https://doi.org/10.1186/s13321-018-0263-1
CATMoS_NT	Collaborative acute toxicity modeling suite non-toxic (NT)	Predicted (kNN)	OPERA	https://doi.org/10.1289/EHP8495
CATMoS_LD50	Collaborative acute toxicity modeling suite median lethal dose (LD50)	Predicted (kNN)	OPERA	https://doi.org/10.1289/EHP8495
ETA_Beta	Extended topochemical atom (ETA) β	2D	PaDEL	http://dx.doi.org/10.1021/ci0342066
MLFER_L	Molecular linear free energy relation solute gas-hexadecane partition coefficient (L)	2D	PaDEL	http://dx.doi.org/10.1021/ci980339t
piPC1	Path counts: conventional bond order ID number of order 1 ($ln(1+x)$)	2D	PaDEL	https://doi.org/10.1002/9783527628766
minwHBa	Atom type electrotopological state: Minimum e-states for weak hydrogen bond acceptors	2D	PaDEL	https://doi.org/10.1021/ci00028a014
GATS1e	Geary autocorrelation: lag 1 weighted by sanderson electronegativities	2D	PaDEL	ISBN: 978-3-527-31852-0 https://doi.org/10.2307/2986645
GATS1m	Geary autocorrelation: lag 1 weighted by mass	2D	PaDEL	ISBN: 978-3-527-31852-0 https://doi.org/10.2307/2986645
SpMax1_Bhs	Largest absolute eigenvalue of Burden modified matrix (n_1) weighted by relative I-state	2D	PaDEL	ISBN: 978-3-527-31852-0
VR3_Dze	Logarithmic Randic-like eigenvector-based index from Barysz matrix weighted by Sanderson electronegativities	2D	PaDEL	ISBN: 978-3-527-31852-0
MAXDN2	Atom type electrotopological state: Maximum negative intrinsic state difference in the molecule	2D	PaDEL	https://doi.org/10.1016/S0045-6535(99)00463-4
ATSC2v	Centered Broto-Moreau autocorrelation: lag 2 weighted by van der Waals volumes	2D	PaDEL	ISBN: 978-3-527-31852-0
AATSC1m	Average centered Broto-Moreau autocorrelation: lag 1 weighted by mass	2D	PaDEL	ISBN: 978-3-527-31852-0
VPC_4	Chi path cluster: valence path cluster, order 4	2D	PaDEL	ISBN: 9780323158312

S1.2 Regression Features

Table S2: Table (1 of 2) showing features used by the CAIRRE_v1 regression model, with the feature labels used by CAIRRE, the software used to generate the feature, and the DOI or ISBN of the reference for that feature.

Feature Label	Feature Description	Type	Software	DOI/ISBN
LogHL_pred	The Henry's Law constant at 25C	Predicted (kNN)	OPERA	https://doi.org/10.1186/s13321-018-0263-1
LogD55_pred	Octanol-water distribution constant	Predicted (kNN)	OPERA	https://doi.org/10.1186/s13321-018-0263-1
LogKoc_pred	The soil adsorption coefficient of organic compounds	Predicted (kNN)	OPERA	https://doi.org/10.1186/s13321-018-0263-1
FUB_pred	Plasma fraction unbound (human, fraction)	Predicted (kNN)	OPERA	
CATMoS_LD50	Collaborative acute toxicity modeling suite median lethal dose (LD50)	Predicted (kNN)	OPERA	https://doi.org/10.1289/EHP8495
MAXDP	Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule)	2D	PaDEL	https://doi.org/10.1016/S0045-6535(99)00463-4
MAXDP2	Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule)	2D	PaDEL	https://doi.org/10.1016/S0045-6535(99)00463-4
SpMax1_Bhs	Largest absolute eigenvalue of Burden modified matrix (n_1) weighted by relative I-state	2D	PaDEL	ISBN: 978-3-527-31852-0
naasC	Atom type electrotopological state: Count of atom-type E-State :C:-	2D	PaDEL	https://doi.org/10.1021/ci00028a014
minsOH	Atom type electrotopological state: Minimum atom-type E-State: -OH	2D	PaDEL	https://doi.org/10.1021/ci00028a014
minaaCH	Atom type electrotopological state: Minimum atom-type E-State: :CH:	2D	PaDEL	https://doi.org/10.1021/ci00028a014
XLogP	Atom-additive method for log P calculation	2D	PaDEL	https://doi.org/10.1023/A:1008763405023
MLFER_L	Molecular linear free energy relation: solute gas-hexadecane partition coefficient (L)	2D	PaDEL	http://dx.doi.org/10.1021/ci980339t
MLFER_E	Molecular linear free energy relation: Excessive molar refraction (E)	2D	PaDEL	http://dx.doi.org/10.1021/ci980339t
CombDipolPolariz	Molecular linear free energy relation: Combined dipolarity/polarizability (S)	2D	PaDEL	http://dx.doi.org/10.1021/ci980339t
nBase	Number of basic groups	2D	PaDEL	
GATS1p	Geary autocorrelation: lag 1 weighted by polarizabilities	2D	PaDEL	ISBN: 978-3-527-31852-0 https://doi.org/10.2307/2986645
GATS1v	Geary autocorrelation: lag 1 weighted by van der Waals volumes	2D	PaDEL	ISBN: 978-3-527-31852-0 https://doi.org/10.2307/2986645
GATS2c	Geary autocorrelation: lag 2 weighted by charges	2D	PaDEL	ISBN: 978-3-527-31852-0 https://doi.org/10.2307/2986645
AATS0e	Average Broto-Moreau autocorrelation: lag 0 weighted by Sanderson electronegativities	2D	PaDEL	ISBN: 978-3-527-31852-0
AATS1e	Average Broto-Moreau autocorrelation: lag 1 weighted by Sanderson electronegativities	2D	PaDEL	ISBN: 978-3-527-31852-0
AATS2i	Average Broto-Moreau autocorrelation: lag 2 / weighted by first ionization potential	2D	PaDEL	ISBN: 978-3-527-31852-0

Table S3: Table (2 of 2) showing features used by the CAIRRE_v1 regression model, with the feature labels used by CAIRRE, the software used to generate the feature, and the DOI or ISBN of the reference for that feature.

Feature Label	Feature Description	Type	Software	DOI/ISBN
AATSC0v	Average centered Broto-Moreau autocorrelation: lag 0 weighted by van der Waals volumes	2D	PaDEL	ISBN: 978-3-527-31852-0
AATSC5m	Average centered Broto-Moreau autocorrelation: lag 5 weighted by mass	2D	PaDEL	ISBN: 978-3-527-31852-0
AATSC3s	Average centered Broto-Moreau autocorrelation: lag 3 weighted by I-state	2D	PaDEL	ISBN: 978-3-527-31852-0
AATSC4s	Average centered Broto-Moreau autocorrelation: lag 4 weighted by I-state	2D	PaDEL	ISBN: 978-3-527-31852-0
AATSC5s	Average centered Broto-Moreau autocorrelation: lag 5 weighted by I-state	2D	PaDEL	ISBN: 978-3-527-31852-0
AATSC8s	Average centered Broto-Moreau autocorrelation: lag 8 weighted by I-state	2D	PaDEL	ISBN: 978-3-527-31852-0
AATSC1m	Average centered Broto-Moreau autocorrelation: lag 1 weighted by mass	2D	PaDEL	ISBN: 978-3-527-31852-0
ATSC0v	Centered Broto-Moreau autocorrelation: lag 0 weighted by van der Waals volumes	2D	PaDEL	ISBN: 978-3-527-31852-0
ATSC2v	Centered Broto-Moreau autocorrelation: lag 2 weighted by van der Waals volumes	2D	PaDEL	ISBN: 978-3-527-31852-0
ATSC2e	Centered Broto-Moreau autocorrelation: lag 2 weighted by Sanderson electronegativities	2D	PaDEL	ISBN: 978-3-527-31852-0
ATSC1p	Centered Broto-Moreau autocorrelation: lag 1 weighted by polarizabilities	2D	PaDEL	ISBN: 978-3-527-31852-0
ATSC2p	Centered Broto-Moreau autocorrelation: lag 2 weighted by polarizabilities	2D	PaDEL	ISBN: 978-3-527-31852-0
ATSC5p	Centered Broto-Moreau autocorrelation: lag 5 weighted by polarizabilities	2D	PaDEL	ISBN: 978-3-527-31852-0
ETA_EtaP	Extended topochemical atom: Composite index η relative to molecular size	2D	PaDEL	http://dx.doi.org/10.1021/ci0342066
piPC1	Path counts: conventional bond order ID number of order 1 ($ln(1+x)$)	2D	PaDEL	https://doi.org/10.1002/9783527628766
IC1	Information content index: neighborhood symmetry of 1-order	2D	PaDEL	ISBN: 978-3-527-31852-0
CIC1	Complementary information content index: neighborhood symmetry of 1-order	2D	PaDEL	ISBN: 978-3-527-31852-0

S2 CAIRRE_v1 Feature Domain Ranges

S2.1 Classification Model

Table S4: Table showing the domain statistics of each feature used in the classification model. To provide additional information to the user, the Molecular Weight (*MolWeight*) is also included in this table despite not being used in the removal classification task.

Feature Label	Maximum	Minimum	Mean	Standard Deviation
MolWeight	*949.178286*	*27.010899*	*264.390095*	*146.989080*
BP_pred	536.000000	-24.000000	287.275449	122.626333
LogVP_pred	3.630000	-12.080000	-3.874311	3.717308
CATMoS_NT	1.000000	0.000000	0.317365	0.465451
CATMoS_LD50	15613.000000	0.680000	1350.912156	1882.727613
ETA_Beta	39.000000	0.500000	18.323353	9.185966
MLFER_L	22.529000	0.462000	7.800527	4.117409
piPC1	4.043051	0.693147	2.902996	0.666550
minwHBa	3.500000	-4.475013	0.483650	1.279784
GATS1e	2.000000	0.190584	0.789274	0.307147
GATS1m	2.000000	0.430018	0.732147	0.262866
SpMax1_Bhs	7.058291	2.821143	4.339267	0.849711
VR3_Dze	36.200057	0.069684	10.409691	7.826091
MAXDN2	10.305860	0.000000	2.048614	2.062916
ATSC2v	1609.936259	-790.566340	-28.541592	338.843228
AATSC1m	33.988177	-371.240410	-4.531341	35.447509
VPC_4	23.758912	0.000000	1.993444	2.965994

S2.2 Regression Model

Table S5: Table (1 of 2) showing the domain statistics of each feature used in the regression model. To provide additional information to the user, the Molecular Weight (MolWeight) is also included in this table despite not being used in the removal efficiency calculation.

Feature Label	Maximum	Minimum	Mean	Standard Deviation
MolWeight	*949.178286*	*27.010899*	*264.390095*	*146.989080*
LogHL_pred	-1.010000	-10.950000	-5.361856	2.476649
LogD55_pred	8.710000	-4.950000	3.244850	2.902082
LogKoc_pred	6.500000	0.200000	3.389042	1.528117
FUB_pred	0.980000	0.000000	0.180539	0.262747
CATMoS_LD50	15613.000000	0.680000	1350.912156	1882.727613
MAXDP	7.174016	0.000000	2.303102	1.922410
MAXDP2	7.174016	0.000000	2.513611	1.677571
SpMax1_Bhs	7.058291	2.821143	4.339267	0.849711
naasC	12.000000	0.000000	2.275449	2.673141
minsOH	10.918809	0.000000	2.066768	3.769267
minaaCH	2.498101	0.000000	1.242938	0.951757
XLogP	11.566000	-0.767000	4.313216	2.581333
MLFER_L	22.529000	0.462000	7.800527	4.117409
MLFER_E	4.382000	-1.270000	1.429275	1.049868
CombDipolPolariz	3.541000	-0.226000	1.333838	0.743734
nBase	1.000000	0.000000	0.065868	0.248052
GATS1p	2.139393	0.316176	0.891020	0.371077
GATS1v	2.000000	0.234153	0.863652	0.282618
GATS2c	1.769931	0.033399	0.996434	0.287729
AATS0e	12.928053	7.083820	8.405660	1.380960
AATS1e	10.087251	7.305580	8.013815	0.643176
AATS2i	207.939296	133.033475	152.821015	17.021435
AATSC0v	85.192721	0.655508	45.478622	15.766703

Table S6: Table (2 of 2) showing the domain statistics of each feature used in the regression model.

Feature Label	Maximum	Minimum	Mean	Standard Deviation
AATSC5m	263.668456	-251.895606	-4.387329	48.898160
AATSC3s	0.910627	-0.774054	0.003304	0.170766
AATSC4s	0.525109	-4.943878	-0.178620	0.559082
AATSC5s	1.530399	-3.547924	-0.132057	0.597756
AATSC8s	1.849486	-11.258316	-0.083199	0.969408
AATSC1m	33.988177	-371.240410	-4.531341	35.447509
ATSC0v	5702.802395	4.661393	1144.347493	852.283917
ATSC2v	1609.936259	-790.566340	-28.541592	338.843228
ATSC2e	3.903690	-4.784404	-0.044205	1.106547
ATSC1p	3.407941	-3.799628	-0.010113	0.808431
ATSC2p	7.904251	-4.666832	-0.200569	2.066034
ATSC5p	9.755275	-19.505172	-0.870983	3.153636
ETA_EtaP	2.120460	0.059630	0.764868	0.370599
piPC1	4.043051	0.693147	2.902996	0.666550
IC1	3.714512	0.811278	2.316459	0.637691
CIC1	4.179279	0.000000	2.058857	0.785556

S3 Feature Rankings

S3.1 Classification Model

Table S7: Table displaying the ranking of the 16 features used by the CAIRRE_v1 classification model in order of importance based on total usage by decision nodes within the GBDT model. Also included are the rankings and ranking statistics of the use of the 16 features in the first decision nodes (Node0) of the GBDT estimators.

Feature Label	Total Rank	Total Freq	Total Percent [%]	Node0 Rank	Node0 Freq	Node0 Percent [%]
minwHBa	1	82	10.93	6	13	10.40
ATSC2v	2	77	10.27	7	8	6.40
LogVP_pred	3	61	8.13	10	2	1.60
MLFER_L	4	59	7.87	1	28	22.40
MAXDN2	5	54	7.20	4	15	12.00
SpMax1_Bhs	6	53	7.07	2	18	14.40
BP_pred	7	53	7.07	3	15	12.00
VR3_Dze	8	51	6.80	8	6	4.80
ETA_Beta	9	47	6.27	5	13	10.40
AATSC1m	10	47	6.27	11	2	1.60
CATMoS_LD50	11	42	5.60	9	4	3.20
piPC1	12	32	4.27	12	1	0.80
GATS1m	13	30	4.00	16	0	0.00
GATS1e	14	27	3.60	15	0	0.00
VPC_4	15	26	3.47	13	0	0.00
CATMoS_NT	16	9	1.20	14	0	0.00

S3.2 Regression Model

Table S8: Table displaying the ranking of the 39 features used by the CAIRRE_v1 regression model in order of importance based on total usage by decision nodes within the GBDT model. Also included are the rankings and ranking statistics of the use of the 40 features in the first decision nodes (Node0) of the GBDT estimators.

Feature	Total Rank	Total Freq	Total Percent [%]	Node0 Rank	Node0 Freq	Node0 Percent [%]
SpMax1_Bhs	1	67	9.24	1	22	14.67
AATSC1m	2	35	4.83	2	17	11.33
CIC1	3	33	4.55	29	0	0.00
GATS1v	4	32	4.41	3	14	9.33
ATSC1p	5	31	4.28	8	7	4.67
AATSC4s	6	28	3.86	6	8	5.33
ATSC5p	7	28	3.86	21	2	1.33
AATSC5s	8	26	3.59	5	9	6.00
CombDipolPolariz	9	25	3.45	4	11	7.33
GATS2c	10	24	3.31	14	4	2.67
AATSC5m	11	24	3.31	13	4	2.67
LogHL_pred	12	24	3.31	19	2	1.33
MLFER_L	13	22	3.03	12	5	3.33
AATSC3s	14	22	3.03	28	1	0.67
XLogP	15	22	3.03	32	0	0.00
CATMoS_LD50	16	20	2.76	30	0	0.00
GATS1p	17	20	2.76	7	8	5.33
IC1	18	20	2.76	15	3	2.00
AATS2i	19	18	2.48	9	7	4.67
ATSC0v	20	17	2.34	25	1	0.67
ETA_EtaP	21	15	2.07	20	2	1.33
piPC1	22	14	1.93	39	0	0.00
AATSC8s	23	14	1.93	11	5	3.33
naasC	24	14	1.93	10	6	4.00
ATSC2p	25	14	1.93	23	1	0.67
LogD55_pred	26	13	1.79	16	3	2.00
AATSC0v	27	12	1.66	31	0	0.00
ATSC2v	28	12	1.66	18	2	1.33
ATSC2e	29	12	1.66	24	1	0.67
MAXDP2	30	12	1.66	38	0	0.00
MLFER_E	31	10	1.38	27	1	0.67
AATS1e	32	8	1.10	17	2	1.33
AATS0e	33	7	0.97	33	0	0.00
MAXDP	34	7	0.97	36	0	0.00
minaaCH	35	6	0.83	35	0	0.00
FUB_pred	36	6	0.83	26	1	0.67
LogKoc_pred	37	5	0.69	34	0	0.00
nBase	38	4	0.55	22	1	0.67
minsOH	39	2	0.28	37	0	0.00

S4 Descriptions of Features

S4.1 Introduction to Chemical Graphs

Many of the descriptors used by the CAIRRE_v1 model suite rely on a class of chemical descriptor known as *Chemical Graph Descriptors*. Therefore, understanding chemical graph descriptors requires an understanding of chemical graphs. Simply put, a chemical graph is a mathematical representation of a chemical's structure based on graph theory.

A typical chemical graph is composed of a series of vertices (or nodes) and edges, with the vertices representing the atoms and the edges representing bonds. This concept is demonstrated in figure S1 for a methane molecule. In this example, the chemical structure is converted to a graph consisting of 5 vertices and 4 edges representing the 5 atoms and 4 covalent bonds in a typical methane molecule. Further depth and complexity can be added to the chemical graphs by incorporating labels to the vertices and edges. This is demonstrated in figure S1 with the use of colours to represent the atom types in the chemical graph of methane. In this example, the carbon atom is represented in green while the hydrogen atoms are shown in blue.

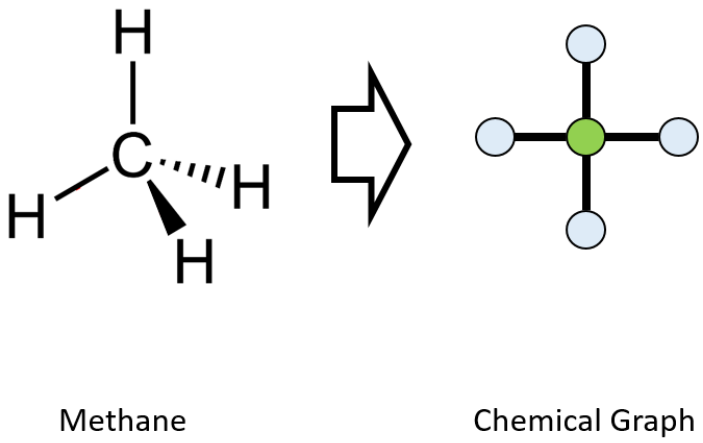

Figure S1: Example of a chemical graph based on the methane molecule. In the chemical graph, the vertices representing hydrogen atoms are blue and the vertex representing the carbon atom is green.

The labels assigned to the nodes and edges of a chemical graph can vary depending on the need, ranging from basic information such as the atomic number on the nodes (demonstrated in figure S1 through the use of colour) and the bond type on the edges, to more complex information like electronegativity, polarizability, etc. For use in machine learning, these chemical graphs are often represented by AxA matrices, where A is the number of atoms (or nodes) in the chemical graph. The matrix representing a simple chemical graph for methane is presented in equation S1, in which the values of the diagonals - representing the vertices (V) of the graph (G) - are the atomic numbers of the atoms. The off-diagonal values in the graph's matrix represent the edges (E) in the graph connecting the vertices. In this example, the off-diagonals will either contain the value of 1 for bonded pairs of atoms, or 0 for non-bonded pairs of atoms shown in equation S1. The value of these edges can vary depending on the bond type to incorporate more information about the chemical structure.

$$G(V,E)_{Methane} = \begin{bmatrix} 6 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{S1}$$

S4.2 Burden Modified Matrix

A practical example of a chemical graph matrix used in the CAIRRE_v1 model suite is the Burden matrix. The Burden matrix is used to describe the connectivity and bonding found within a chemical graph. This takes the form of an AxA square matrix, where A is the total number of atoms in the chemical graph. Since atoms cannot be bonded to itself, additional data is added to the matrix by inserting the atomic numbers along the matrix diagonal. The off diagonal values between bonded atoms (B_{ij}) are calculated using equation S2, where π^* is the conventional bond order of the chemical bond between atoms i and j (See table S9).

$$B_{ij} = \pi^* 10^{-1} \tag{S2}$$

For terminally bonded atoms, the value in the matrix is augmented by 0.01. For all non-bonded atom pairs, a value of 0.001 is assigned to the corresponding matrix element. An example of the Burden matrix calculated for a methane molecule is shown in equation S3. In this example, all four of the edges are considered to be *terminal* and therefore the value of B_{ij} is augmented by 0.01.

Table S9: Traditional bond order values used in the Burden Matrix for conventional bond types.

Bond Type	π^*
Single	1.0
Double	2.0
Triple	3.0
Aromatic	1.5

$$B(V,E)_{Methane} = \begin{bmatrix} 6 & B_{ij}+0.01 & B_{ij}+0.01 & B_{ij}+0.01 & B_{ij}+0.01 \\ B_{ij}+0.01 & 1 & 0.001 & 0.001 & 0.001 \\ B_{ij}+0.01 & 0.001 & 1 & 0.001 & 0.001 \\ B_{ij}+0.01 & 0.001 & 0.001 & 1 & 0.001 \\ B_{ij}+0.01 & 0.001 & 0.001 & 0.001 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 0.110 & 0.110 & 0.110 & 0.110 \\ 0.110 & 1 & 0.001 & 0.001 & 0.001 \\ 0.110 & 0.001 & 1 & 0.001 & 0.001 \\ 0.110 & 0.001 & 0.001 & 1 & 0.001 \\ 0.110 & 0.001 & 0.001 & 0.001 & 1 \end{bmatrix} \tag{S3}$$

S4.3 Autocorrelations: General

Autocorrelation descriptors are calculated by transforming a chemical structure into a chemical graph. A chemical graph is an un-directed graph composed of nodes and edges, with the nodes representing the atoms and their atomic properties, and the edges representing the bonds between those atoms. When no knowledge of the 3-dimensional geometry of the molecule is not known, simple topological information can still be extracted from a chemical graph. Autocorrelation descriptors leverage this information to link the chemical structure to some observable (boiling point, melting point, etc.) based on of a general autocorrelation function shown in equation S4.

$$AC_l = \int_a^b f(x)f(x+l)\,dx \tag{S4}$$

When dealing with the information contained in the chemical graph, the distances between atoms (or nodes) are discrete and form an ordered sequence of values for $f(x_i)$. Equation S4 can therefore be rewritten as:

$$AC_l = \sum_{i=1}^{n-l} f(x_i)f(x_{i+l}) \tag{S5}$$

Both equations S4 and S5 include a property known as *Lag*, delimited by the variable l. Autocorrelation calculations performed on a chemical graphs are often weighted by some chemical property, w, to better capture the chemistry of the molecule. The variable of x_i denotes the node in the graph representing atom i. Since the chemical graph is composed of discrete points, the atomic property is therefore represented as $f(x)$ and we can define the *lag* as the topological distance d_{ij}, which is simply the number of edges (or bonds) between two atoms (i and j) in the chemical graph.

S4.4 Autocorrelations: Broto-Moreau

The Broto-Moreau autocorrelation function is often referred to as the *Autocorrelation of a Topological Surface (ATS)*. This feature describes the distribution of an atomic property along the topological surface of a molecular graph. In equation S6, w is some atomic property, A is the total number of atoms in the chemical graph, d is the topological distance, and δ_{ij} is the Kronecker delta.

$$ATS_d = \sum_{i=1}^{A}\sum_{j=1}^{A} \delta_{ij}(w_i w_j)_d \tag{S6}$$

Since the graph distance (d) is a discrete variable, equation S6 can be solved for every value of d from 0 to the maximum topological distance (D) in the graph. This value of D is commonly known as the topological diameter of the graph. Solving this equation therefore generates a series for any atomic property, w, shown in equation S7.

$$\langle ATS_0, ATS_1, ..., ATS_D \rangle_w \tag{S7}$$

The *Broto-Moreau* autocorrelation function is therefore finite when calculated for discrete chemicals. As such, an *Average Autocorrelation of a Topological Surface* ($AATS$ or $\overline{ATS}$) can be calculated for an atomic property using equation S8, where Δ is the sum over all values of δ_{ij} for the distance d.

$$\overline{ATS}_d = \frac{1}{\Delta} \sum_{i=1}^{A} \sum_{j=1}^{A} \delta_{ij} (w_i w_j)_d \tag{S8}$$

S4.5 Autocorrelations: Geary

The Geary autocorrelation function, or the *Geary Coefficient* (GC), operates using a similar premise to that of the Broto-Moreau, taking the form of a distance type function ranging from 0 to ∞. In equation S9, $\overline{w}$ is the average of the atomic property, w, over the whole molecule.

$$GC(d) = \frac{\frac{1}{2\Delta} \sum_{i=1}^{A} \sum_{j=1}^{A} (w_i - w_j)^2}{\frac{1}{A-1} \sum_{i=1}^{A} (w_i - \overline{w})^2} \tag{S9}$$

S4.6 Extended Topological Atom (ETA): General

The extended topological atom (ETA) descriptors are expansions on the TAU descriptors.[1] The TAU descriptors are calculated using the molecular graph of a chemical developed in the valence electron mobile (VEM) environement, separating the nodes of the graph by their core and valence electronic environments. In the ETA formalism, the electronic structure of a node within the chemical graph can be described by a series of equations, each relating the parameter to a different observable property of a chemical. The first of these functions, the α value of a vertex, defines the core electron count of a non-hydrogen atom. This function is shown in equation S10, where z is the atomic number, z^v is the number of valence electrons, and PN is the elemental period number for atom i in the chemical graph.

$$\alpha_i = \frac{z_i - z_i^v}{z_i^v} \frac{1}{PN_i - 1} \tag{S10}$$

S4.7 Extended Topological Atom (ETA): β

The first of these descriptors used in the CAIRRE_v1 models is the β parameter, described by equations S11 and S12. For atom i, this equation considers the contribution of all sigma bonds (x_σ) and the contributions of all π bonds (y_π) centered on the atom, adjusted by a correction factor (δ). The values of x and y are selected based on the difference in electronegativity of the two bonded atoms (table S10) and a value of 0.5 is chosen

for δ when a lone pair of electrons capable of resonance is present. In the absence of this lone pair, a value of 0 is assigned to δ. Once β_i is computed for each atom in the molecular graph, the final β descriptor is calculated using equation S12.

$$\beta_i = \sum x_\sigma + \sum y_\pi + \delta \tag{S11}$$

$$\beta = \sum_{i=1}^{A} \beta_i \tag{S12}$$

Table S10: Contributions of the sigma bonds (x) and π bonds (y) to β_i based on the difference in electronegativity (Δe) between the two bonded atoms.

Δe	x	y
≤ 0.3	0.50	1.0
> 0.3	0.75	1.5

S4.8 Extended Topological Atom (ETA): η

Another of the ETA descriptors used by the CAIRRE_v1 model suite is the η parameter, described by equation S13. This parameter is the composite index considering both the bonded and non-bonded interactions of the atoms in the chemical graph. In this equation, r_{ij} is the topological distance between atoms i and j, and γ_i is defined in equation S14 as the ratio between the α_i and β_i parameters for atom i.

$$\eta = \sum_{i<j} \left[\frac{\gamma_i \gamma_j}{r_{ij}^2} \right]^{0.5} \tag{S13}$$

$$\gamma_i = \frac{\alpha_i}{\beta_i} \tag{S14}$$

S4.9 MaxDP and MaxDP2

Maximum positive intrinsic state difference (MaxDP) in the molecule considers the intrinsic states of the atoms in the molecular graphs of discrete chemicals. The first of the MaxDP values can be calculated using equation S15. In this equation, ΔI_i is defined as the Field effect on the i^{th} atom due to the perturbation of all other atoms[2,3] and is calculated using equation S16.

$$MaxDP = max_i \left| \Delta I_i(\delta_1^v) \right| \tag{S15}$$

$$\Delta I_i(\delta^v) = \sum_j \frac{I_i - I_j}{(d_{ij}+1)^2} \tag{S16}$$

To calculate ΔI_i, one needs to calculate the intrinsic state, I, for every atom, i, in the molecular graph using equation S17. In this equation, δ^v is the number of valence electrons and δ is the number of sigma electrons on the selected atom.

$$I(\delta^v) = \frac{\delta^v + 1}{\delta} \tag{S17}$$

$$\delta_1^v = \frac{Z^v - nB_H}{Z - Z^v - 1} \tag{S18}$$

The models found within the CAIRRE_v1 model suite use two versions of the MaxDP descriptors: MaxDP and MaxDP2 calculated using equations S15 and S19, respectively. The key difference between these descriptors can be found in the calculation of the δ^v parameter, where MaxDP relies on the equation S18 and MaxDP2 on equation S20. In both equations for δ^v, Z is the atomic number, Z^v is the number of valence electrons on the atom, and nB_H is the number of hydrogen atoms bonded to the atom.

$$MaxDP2 = max_i \left|\Delta I_i(\delta_2^v)\right| \tag{S19}$$

$$\delta_2^v = Z^v - nB_H \tag{S20}$$

S4.10 Atom Type Electrotopological States

The atom type electropological state descriptors[4] are a group of topological indices based on E-state indices calculated using equation S21. In this equation, the value of I_i and ΔI_i are calculated using equations S16 and S17, respectively.

$$S_i = I_i + \Delta I_i \tag{S21}$$

Once the states have been calculated for each atom in the molecule, those atoms are then classified into atom types according to a variety of criteria including but not limited to: atom identity, valence state, and the number of bonded hydrogen atoms. The final indices for a selected atom type then becomes the some of the S_i values for each atom of that type.

The CAIRRE_v1 model suite relies on a variety of these indices for both the classification and regression tasks, however the feature with label *minwHBA* was identified as the most important feature in the classification,

whereas the other 3 indices: *minaaCH*, *naasC*, and *minsOH* were among the least important features. Since the methodology for calculating these descriptors is identical, save for the atom type classification, only the *minwHBA* index will be discussed.

In the case of the *minwHBA* feature which is described as the "minimum e-states for weak hydrogen bond acceptors", the code identifies atoms in the molecular graph which would be considered weak hydrogen bond acceptors using a series of heuristic rules. Once the set of atoms has been determine, the *minwHBA* feature is calculated by simply taking the smallest value of S_i among those atom types. It is therefore possible to generate atom type electropological state indices using the minimum, maximum, or sum of the atom type subset.

S4.11 Complimentary Information Content Index

Expanding on the idea of indices based on molecular graphs, a series indices based on the field of information theory are also calculated by PaDEL[5] and used throughout the OPERA software.[6] Among the descriptors used by the CAIRRE_v1 model suite is the Information Content Index (IC) described by the *Shannon Formula*[7] shown in equation S22. In this equation, n is the total number of vertices (or atoms) in the molecular graph, n_i is the number of equivalent atoms of type i, and k is the order of neighbourhood symmetry ($k = 0, 1, 2..., K$). In this context, the *Shannon formula* describes the entropy (or complexity) of the probability distribution of the molecular graph, by taking the sum of all probabilities, $p_i = n_i/n$, of randomly selecting a vertex belonging to the i^{th} subset. The IC index is therefore a measure of the overall complexity of the molecular graph.

$$^{k}IC = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n} \tag{S22}$$

An important consideration when computing these indices is the definition of *equivalent* atoms, computed by considering the k surrounding layers to the atom's vertex. Both the IC and CIC descriptors used by CAIRRE are first-order ($k = 1$), meaning they only consider adjacent vertices when checking for equivalency. The equivalency definition used in this work was formulated by Sarkar et al.[8] and states that two vertices of a chemical graph are equivalent if they have the same number of edges and the same number of first-order neighbours that have identical degrees. When $k > 1$, this is expanded to second-, third-, etc. order neighbours.

The amount of information extracted by this index can be improved through the use of a complimentary information content index (CIC) defined by equation S23.[9]

$$^{k}CIC = \log_2 n -^{k} IC \tag{S23}$$

S4.12 Molecular Linear Free Energy Relation (MLFER)

The descriptors in the Molecular Linear Free Energy (MLFER) descriptor set are based on the relationship proposed by Abraham[10] relating some solvation property (SP) to a linear combination of interaction terms,

shown in equation S24. In this equation, R_2 is the excess molar refraction defined as the molar refraction of the solute minus the molar refraction of an alkane minus the equivalent volume. The variable π_2^H is a combined dipolarity and polarizability descriptors, $\sum \alpha_2^H$ is the overall solute hydrogen bonds acidity, $\sum \beta_2^H$ is the overall solute hydrogen bond basicity, and $\log L^{16}$ is the solute gas-hexadecane partition coefficient.[11]

$$\log SP = c + rR_2 + s\pi_2^H + a\sum \alpha_2^H + b\sum \beta_2^H + l\log L^{16} \tag{S24}$$

Using OPERA,[6] all five descriptors in equation S24 can be estimated for use in other machine learning tasks. The two descriptors of interest used by the CAIRRE_v1 model suite are the $\log L^{16}$ (labeled as MLFER_L in tables S1 and S2) and R_2 (labeled as MLFER_E in table S2) components. To estimate $\log L^{16}$, PaDEL uses the additive scheme proposed by Svecik et al.[12] shown in S25. In this equation, FG_i is the number of functional group i, SC_j are the structural contributions of functional group j, and IC_k are the interaction contributions of functional group k. In equation S25, the terms l_i, m_j, and n_k are fitted regression terms.

$$\log L^{16}(X) = \sum_i l_i \times FG_i + \sum_j m_j \times SC_j + \sum_k n_k \times IC_k \tag{S25}$$

The other term in equation S24 used by the CAIRRE_v1 model suite is the R_2 parameter, or the excess molar refraction of the molecule. To estimate this property, PaDEL uses the method proposed by Abraham and Whiting[13] described by equation S26. In this equation, $\mathrm{MR_x}$(observed) is the molar refraction of the chemical being modeled and $\mathrm{MR_x}$(alkane of same V_x) is the molar refraction of a reference alkane.

$$R_2 = \mathrm{MR_x}(\text{observed}) - \mathrm{MR_x}(\text{alkane of same } V_x) \tag{S26}$$

The value of $\mathrm{MR_x}$ is calculated using equation S27, where η is the refractive index and V_x is the volume of the solute being tested. The value of equation S26 becomes zero when the molar refraction of the chemical being modeled matches a reference alkane with the same value of V_x.

$$\mathrm{MR_x} = \frac{10\left(\eta^2 - 1\right)V_x}{\eta^2 + 2} \tag{S27}$$

References

[1] Kunal Roy and Gopinath Ghosh. Qstr with extended topochemical atom indices. 2. fish toxicity of substituted benzenes. *Journal of chemical information and computer sciences*, 44(2):559–567, 2004.

[2] P Gramatica, M Corradi, and V Consonni. Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. *Chemosphere*, 41(5):763–777, 2000.

[3] Lemont B Kier, Lowell H Hall, and Jack W Frazer. An index of electrotopological state for atoms in molecules. *Journal of Mathematical Chemistry*, 7(1):229–241, 1991.

[4] Lowell H Hall and Lemont B Kier. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *Journal of Chemical Information and Computer Sciences*, 35(6):1039–1045, 1995.

[5] Chun Wei Yap. Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry*, 32(7):1466–1474, 2011.

[6] Kamel Mansouri, Christopher Grulke, Richard Judson, and Antony Williams. Opera: A free and open source qsar tool for predicting physicochemical properties and environmental fate endpoints. In *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*, volume 255. AMER CHEMICAL SOC 1155 16TH ST, NW, WASHINGTON, DC 20036 USA, 2018.

[7] CE Shannon. A mathematical theory of communication (vol. 27, pp. 379–423). *Urbana: University of Illinois Press*, 1984.

[8] Rina Sarkar, AB Roy, and PK Sarkar. Topological information content of genetic molecules—i. *Mathematical Biosciences*, 39(3-4):299–312, 1978.

[9] Subhash C Basak, DK Harriss, and VR Magnuson. Comparative study of lipophilicity versus topological molecular descriptors in biological correlations. *Journal of pharmaceutical sciences*, 73(4):429–437, 1984.

[10] Michael H Abraham. Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. *Chemical Society Reviews*, 22(2):73–83, 1993.

[11] James A Platts, Darko Butina, Michael H Abraham, and Anne Hersey. Estimation of molecular linear free energy relation descriptors using a group contribution approach. *Journal of Chemical Information and Computer Sciences*, 39(5):835–845, 1999.

[12] Pavel Havelec and Jiří GK Ševčík. Extended additivity model of parameter log (l 16). *Journal of Physical and Chemical Reference Data*, 25(6):1483–1493, 1996.

[13] Michael H Abraham, Garry S Whiting, Ruth M Doherty, and Wendel J Shuely. Hydrogen bonding. part 13. a new method for the characterisation of glc stationary phases—the laffort data set. *Journal of the Chemical Society, Perkin Transactions 2*, (8):1451–1460, 1990.