Annex - RDKit Full Descriptor List

Descriptors used in Machine Learning (ML) model development (Table S1 and S2) retrieved from RDKit software. The respective bibliographical foundation of each descripted is described in detail in the RDKit WebBook Documentation[1]. The descriptor calculations were made by converting each SMILES string (representing each molecular entry), running on top of Python version 3.9.8[2] and using RDKit package, version 2025_03_4 (Q1 2025)[3]. The Chemical information Type is presented, and the chemical species are sorted in the database file in the repository. The code used to generate descriptor values are in the online repository. The database uses as set of 106 for encoding values for each compound. The descriptors encode electronic, structural and van der Waals surface area (VSA) features.

Table S1

List of RDKit Descriptors used in ML model development (*n* = 106 descriptors).

Descriptor	Chemical information Type	Meaning
BalabanJ (1)	Structural	Distance sum of the two end-vertex for each edge. Connectivity of a molecule based on its graph structure (BalabanJ index has been proven to be relevant to network branching).
BertzCT (1)	Structural	Complexity index, considering both the variety of kinds of bond connectivity's and atom types; information contents related to bond connectivity and atom type diversity.
0χ , 1χ (2)	Structural	This descriptor signifies a retention index derived directly from gradient retention times. Considers individual atoms and their valence (0) and pairs of directly connected atoms (1).
0χn – 4χn (5)	Structural	This descriptor signifies a retention index derived directly from gradient retention times (normalized indexes).
0χv – 4χv (5)	Structural	This descriptor signifies atomic valence connectivity index. Indices weighted by atomic valence

EState_VSA1 – EState_VSA11 (11)	VSA	MOE-type (QSAR model) descriptors using EState indices and surface area contributions. RDKit automatically calculates the EState index for each atom in the molecule, assigning each atom to a predefined EState range (bin). Sums the VSA of atoms within each bin (surface area grouped by EState range).
Hall Kier α (1)	Structural	Descriptor fingerprint that displays the difference between active and inactive molecules.
HeavyAtomCount (1)	Structural	The number of heavy atoms in the molecule.
HeavyAtomMolWt (1)	Structural	The average molecular weight of the molecule ignoring hydrogens.
κ_1 , κ_2 , κ_3 **(3)**	Structural	This descriptor signifies # κ shape index: $(n-1) \times 2 / m^2$
Max and Min AbsEStateIndex **(2)**	Electronic	Maximum and minimum absolute E-State
Max and Min AbsPartialCharge **(2)**	Electronic	Maximum and minimum absolute partial charge
Max and Min EStateIndex **(2)**	Electronic	Maximum and minimum E-State
Max and Min PartialCharge **(2)**	Electronic	Maximum and minimum partial charge
Mol logP (1)	Structural	Wildman-Crippen logP value.
MolMR (1)	Electronic	Wildman-Crippen molar refractivity.
MolWt (1)	Structural	The average molecular weight of the molecule.
NHOH Count (1)	Structural	The number of NHs or OHs.
NO Count (1)	Structural	The number of Nitrogens and Oxygens.
$n_{alicarb}$ **(1)**	Structural	The number of aliphatic carbocycles.
n_{alihet} **(1)**	Structural	The number of aliphatic heterocycles.
n_{alirig} **(1)**	Structural	The number of aliphatic rings.
$n_{arocarb}$ **(1)**	Structural	The number of aromatic carbocycles.

n_{arohet} (1)	Structural	The number of aromatic heterocycles.
n_{arorig} (1)	Structural	The number of aromatic rings.
n_{Ha} (1)	Structural	The number of Hydrogen Bond Acceptors.
n_{Hd} (1)	Structural	The number of Hydrogen Bond Donors.
n_{het} (1)	Structural	The number of Heteroatoms.
n_{radele} (1)	Structural	The number of radical electrons.
n_{rot} (1)	Structural	The number of rotatable bonds.
$n_{satcarb}$ (1)	Structural	The number of saturated carbocycles.
n_{sathet} (1)	Structural	The number of saturated heterocycles.
n_{satrig} (1)	Structural	The number of saturated rings.
n_{ele} (1)	Structural	The number of valence electrons.
PEOE_VSA1 – PEOE_VSA14 (14)	VSA	MOE-type (QSAR model) descriptors using partial charges estimated using the Gasteiger PEOE and surface area contributions (sum the VSA of atoms within a given partial charge range).
SMR_VSA1 – SMR_VSA10 (10)	VSA	MOE-type (QSAR model) descriptors using atomic contributions to molar refractivity and surface area contributions (sums the VSA of atoms in each bin).
SlogP_VSA1 – SlogP_VSA12 (12)	VSA	MOE-type (QSAR model) descriptors using logP (calculated using the Crippen method) contributions and surface area sum of surface areas of atoms whose logP contribution falls within a specific range).
TPSA (1)	VSA	The total polar surface area of a molecule based upon fragment calculations.
VSA_EState1 – VSA_EState10 (10)	VSA	MOE-type (QSAR model) descriptors using EState indices and surface area contributions. The sum of VSA contributions of atoms whose EState values fall into specific predefined ranges (surface area grouped by predefined VSA ranges).

Bibliography

(1) *RDKit: Open-source cheminformatics - Descritptor Guide Online Webbook.* https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors (accessed July 1st, 2025).

(2) *Python Software Foundation - Python Language Reference, version 3.9.8*. http://www.python.org (accessed *July 1st, 2025).*

(3) Landrum, G. *RDKit: Open-source cheminformatics* 2025_03_4 (Q1 2025) Release - June 2025. http://www.rdkit.org/ (accessed July 1st, 2025).