Supplementary Information

Distribution-preserved sampling (DPS) for smarter machine learning assisted ultra-large-scale virtual screening

Alexander Trachtenberg, Alexander Spelkov, and Barak Akabayov*

Department of Chemistry and Data Science Research Center, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel.

*e-mail:akabayov@bgu.ac.il

Brief description of chemical descriptors in RDKit (https://github.com/milanimarcel/Descriptors-from-RDKit/blob/master/README.md):

- **MaxEStateIndex**: Returns a tuple of EState indices for the molecule, Reference: Hall, Mohney and Kier. JCICS *31* 76-81 (1991)
- **MinEStateIndex**: Returns a tuple of EState indices for the molecule, Reference: Hall, Mohney and Kier. JCICS *31* 76-81 (1991)
- **MaxAbsEStateIndex**: Returns a tuple of EState indices for the molecule, Reference: Hall, Mohney and Kier. JCICS *31* 76-81 (1991)
- **MinAbsEStateIndex**: Returns a tuple of EState indices for the molecule, Reference: Hall, Mohney and Kier. JCICS *31* 76-81 (1991)
- **QED**: QED stands for quantitative amount of similarity of drugs and the concept was first introduced by Richard Bickerton and his colleagues. The empirical logic of the QED measure reflects the underlying distribution of molecular properties, including molecular weight, logP, topological polar surface area, number of hydrogen scavenger donors and acceptors, number of aromatic rings and rotating screens and the presence of unwanted chemicals.
- **SPS**: An empirical measure of molecular three-dimensional complexity and spatial arrangement, useful for assessing scaffold diversity
- **MolWt**: The average molecular weight of the molecule
- **HeavyAtomMolWt**: The average molecular weight of the molecule without the hydrogens
- **ExactMolWt**: The exact molecular weight of the molecule
- **NumValenceElectrons**: The number of valence electrons of the molecule
- **NumRadicalElectrons**:The number of radical electrons of the molecule (says nothing about spin state)
- **MaxPartialCharge**: A partial charge is a non-integer charge value when measured in elementary charge units. Partial charge is more commonly called net atomic charge. It is represented by the Greek lowercase letter δ, namely, as δ− or δ+. Partial charges are created due to the asymmetric distribution of electrons in chemical bonds.
- **MinPartialCharge**: Min partial Charge
- **MaxAbsPartialCharge**:Returns molecular charge descriptors
- **MinAbsPartialCharge**:Returns molecular charge descriptors
- **FpDensityMorgan1**: Morgan fingerprint density
- **FpDensityMorgan2**: Morgan fingerprint density
- **FpDensityMorgan3**: Morgan fingerprint density
- **BCUT2D_MWHI** / **BCUT2D_MWLOW**: High and low eigenvalues of a weighted topological matrix derived using atomic molecular weights, capturing extremes in molecular-weight–based topology

- **BCUT2D_CHGHI** / **BCUT2D_CHGLO**: High and low eigenvalues based on partial atomic charges, representing charge distribution extremes
- **BCUT2D_LOGPHI** / **BCUT2D_LOGPLOW**: Eigenvalue descriptors using LogP contributions, capturing hydrophobicity extremes across the molecule
- **BCUT2D_MRHI** / **BCUT2D_MRLOW**: High and low eigenvalues derived from atomic molar refractivity values, reflecting polarizability extremes
- **AvgIpc**: Average Inner Polar Surface Area - An average per-atom inner polar surface measure, indicative of polarity distribution across the molecule
- **BalabanJ**: Balaban's J value for a molecule,Chem. Phys. Lett. 89:399-404 (1982).
- **BertzCT**:"A topological index meant to quantify ""complexity"" of molecules.J. Am. Chem. Soc. 103:3599-601 (1981)."
- **Chi0**: From equations (1),(9) and (10) of Rev. Comp. Chem. vol 2, 367-422, (1991)
- **Chi0n**: Similar to Hall Kier Chi0v, but uses nVal instead of valence This makes a big difference after we get out of the first row.Rev. Comput. Chem. 2:367-422 (1991).
- **Chi0v**: From equations (5),(9) and (10) of Rev. Comp. Chem. vol 2, 367-422, (1991)
- **Chi1**: From equations (1),(11) and (12) of Rev. Comp. Chem. vol 2, 367-422, (1991)
- **Chi1n**: Similar to Hall Kier Chi1v, but uses nVal instead of valence.Rev. Comput. Chem. 2:367-422 (1991).
- **Chi1v**: From equations (5),(11) and (12) of Rev. Comp. Chem. vol 2, 367-422, (1991)
- **Chi2n**: Similar to Hall Kier Chi2v, but uses nVal instead of valence This makes a big difference after we get out of the first row.Rev. Comput. Chem. 2:367-422 (1991).
- **Chi2v**: From equations (5),(15) and (16) of Rev. Comp. Chem. vol 2, 367-422, (1991)
- **Chi3n**: Similar to Hall Kier Chi3v, but uses nVal instead of valence This makes a big difference after we get out of the first row.Rev. Comput. Chem. 2:367-422 (1991).
- **Chi3v**: From equations (5),(15) and (16) of Rev. Comp. Chem. vol 2, 367-422, (1991)
- **Chi4n**: Similar to Hall Kier Chi4v, but uses nVal instead of valence.This makes a big difference after we get out of the first row.Rev. Comput. Chem. 2:367-422 (1991).
- **Chi4v**: From equations (5),(15) and (16) of Rev. Comp. Chem. vol 2, 367-422, (1991)
- **HallKierAlpha**: The Hall-Kier alpha value for a molecule. Rev. Comput. Chem. 2:367-422 (1991).
- **Ipc**: the information content of the coefficients of the characteristic polynomial of the adjacency matrix of a hydrogen-suppressed graph of a molecule.
- **Kappa1**: Kappa indices are calculated relative to the least branched (linear) and most branched (star) compounds with the same number of atoms as the molecule being investigated.
- **Kappa2**: Hall-Kier Kappa2 value
- **Kappa3**: Hall-Kier Kappa3 value
- **LabuteASA**: Labute's Approximate Surface Area (ASA from MOE)
- **PEOE_VSA1**: MOE(Molecular Operating Environment) Charge VSA Descriptor 1 (-inf < x < -0.30)
- **PEOE_VSA10**: MOE Charge VSA Descriptor
- **PEOE_VSA11**: MOE Charge VSA Descriptor 11 (0.15 <= x < 0.20)
- **PEOE_VSA12**: MOE Charge VSA Descriptor 12 (0.20 <= x < 0.25)
- **PEOE_VSA13**: MOE Charge VSA Descriptor 13 (0.25 <= x < 0.30)
- **PEOE_VSA14**: MOE Charge VSA Descriptor 14 (0.30 <= x < inf)
- **PEOE_VSA2**: MOE Charge VSA Descriptor 2 (-0.30 <= x < -0.25)
- **PEOE_VSA3**: MOE Charge VSA Descriptor 3 (-0.25 <= x < -0.20)
- **PEOE_VSA4**: MOE Charge VSA Descriptor 4 (-0.20 <= x < -0.15)

- **PEOE_VSA5**: MOE Charge VSA Descriptor 5 (-0.15 <= x < -0.10)
- **PEOE_VSA6**: MOE Charge VSA Descriptor 6 (-0.10 <= x < -0.05)
- **PEOE_VSA7**: MOE Charge VSA Descriptor 7 (-0.05 <= x < 0.00)
- **PEOE_VSA8**: MOE Charge VSA Descriptor 8 (0.00 <= x < 0.05)
- **PEOE_VSA9**: MOE Charge VSA Descriptor 9 (0.05 <= x < 0.10)
- **SMR_VSA1**: MOE MR VSA Descriptor 1 (-inf < x < 1.29)
- **SMR_VSA10**: MOE MR VSA Descriptor 10
- **SMR_VSA2**: MOE MR VSA Descriptor 2 (1.29 <= x < 1.82)
- **SMR_VSA3**: MOE MR VSA Descriptor 3 (1.82 <= x < 2.24)
- **SMR_VSA4**: MOE MR VSA Descriptor 4 (2.24 <= x < 2.45)
- **SMR_VSA5**: MOE MR VSA Descriptor 5 (2.45 <= x < 2.75)
- **SMR_VSA6**: MOE MR VSA Descriptor 6 (2.75 <= x < 3.05)
- **SMR_VSA7**: MOE MR VSA Descriptor 7 (3.05 <= x < 3.63)
- **SMR_VSA8**: MOE MR VSA Descriptor 8 (3.63 <= x < 3.80)
- **SMR_VSA9**: MOE MR VSA Descriptor 9 (3.80 <= x < 4.00)
- **SlogP_VSA1**: MOE logP VSA Descriptor 1 (-inf < x < -0.40)
- **SlogP_VSA10**: MOE logP VSA Descriptor 10 (0.40 <= x < 0.50)
- **SlogP_VSA11**: MOE logP VSA Descriptor 11 (0.50 <= x < 0.60)
- **SlogP_VSA12**: MOE logP VSA Descriptor 12 (0.60 <= x < inf)
- **SlogP_VSA2**: MOE logP VSA Descriptor 2 (-0.40 <= x < -0.20)
- **SlogP_VSA3**: MOE logP VSA Descriptor 3 (-0.20 <= x < 0.00)
- **SlogP_VSA4**: MOE logP VSA Descriptor 4 (0.00 <= x < 0.10)
- **SlogP_VSA5**: MOE logP VSA Descriptor 5 (0.10 <= x < 0.15)
- **SlogP_VSA6**: MOE logP VSA Descriptor 6 (0.15 <= x < 0.20)
- **SlogP_VSA7**: MOE logP VSA Descriptor 7 (0.20 <= x < 0.25)
- **SlogP_VSA8**: MOE logP VSA Descriptor 8 (0.25 <= x < 0.30)
- **SlogP_VSA9**: MOE logP VSA Descriptor 9 (0.30 <= x < 0.40)
- **TPSA**: The polar surface area (PSA) or topological polar surface area (TPSA) of a molecule is defined as the surface sum over all polar atoms or molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms. PSA is a commonly used medicinal chemistry metric for the optimization of a drug's ability to permeate cells.
- **EState_VSA1**: EState VSA Descriptor 1 (-inf < x < -0.39)
- **EState_VSA10**: EState VSA Descriptor 10 (9.17 <= x < 15.00)
- **EState_VSA11**: EState VSA Descriptor 11 (15.00 <= x < inf)
- **EState_VSA2**: EState VSA Descriptor 2 (-0.39 <= x < 0.29)
- **EState_VSA3**: EState VSA Descriptor 3 (0.29 <= x < 0.72)
- **EState_VSA4**: EState VSA Descriptor 4 (0.72 <= x < 1.17)
- **EState_VSA5**: EState VSA Descriptor 5 (1.17 <= x < 1.54)
- **EState_VSA6**: EState VSA Descriptor 6 (1.54 <= x < 1.81)
- **EState_VSA7**: EState VSA Descriptor 7 (1.81 <= x < 2.05)
- **EState_VSA8**: EState VSA Descriptor 8 (2.05 <= x < 4.69)
- **EState_VSA9**: EState VSA Descriptor 9 (4.69 <= x < 9.17)
- **VSA_EState1**: VSA EState Descriptor 1 (-inf < x < 4.78)
- **VSA_EState10**: VSA EState Descriptor 10 (11.00 <= x < inf)
- **VSA_EState2**: VSA EState Descriptor 2 (4.78 <= x < 5.00)
- **VSA_EState3**: VSA EState Descriptor 3 (5.00 <= x < 5.41)
- **VSA_EState4**: VSA EState Descriptor 4 (5.41 <= x < 5.74)
- **VSA_EState5**: VSA EState Descriptor 5 (5.74 <= x < 6.00)

- **VSA_EState6**: VSA EState Descriptor 6 (6.00 <= x < 6.07)
- **VSA_EState7**: VSA EState Descriptor 7 (6.07 <= x < 6.45)
- **VSA_EState8**: VSA EState Descriptor 8 (6.45 <= x < 7.00)
- **VSA_EState9**: VSA EState Descriptor 9 (7.00 <= x < 11.00)
- **FractionCSP3**: The fraction of C atoms that are SP3 hybridized.
- **HeavyAtomCount**: Number of heavy atoms of a molecule.
- **NHOHCount**: Number of -NH and -OH groups
- **NOCount**: Number of nitrogen and oxygen atoms
- **NumAliphaticCarbocycles**: returns the number of aliphatic (containing at least one non-aromatic bond) carbocycles for a molecule
- **NumAliphaticHeterocycles**: returns the number of aliphatic (containing at least one non-aromatic bond) heterocycles for a molecule
- **NumAliphaticRings**: returns the number of aliphatic (containing at least one non-aromatic bond) rings for a molecule
- **NumAmideBonds**: The total count of amide bonds in the molecule, suggesting potential hydrogen-bonding capacity
- **NumAromaticCarbocycles**: The number of aromatic carbocycles for a molecule
- **NumAromaticHeterocycles**: The number of aromatic rings containing heteroatoms (e.g., N, O, S), indicative of key pharmacophores
- **NumAromaticRings**: The number of aromatic rings for a molecule
- **NumAtomStereoCenters**: Count of stereogenic atoms with defined chirality centers
- **NumBridgeheadAtoms**: Number of atoms at bridgehead positions in polycyclic structures, reflecting topological complexity
- **NumHAcceptors**: Number of Hydrogen Bond Acceptors
- **NumHDonors**: Number of Hydrogen Bond Donors
- **NumHeteroatoms**: Number of Heteroatoms
- **NumHeterocycles**: Total number of ring systems containing at least one heteroatom
- **NumRotatableBonds**: Number of Rotatable Bonds
- **NumSaturatedCarbocycles**: returns the number of saturated carbocycles for a molecule
- **NumSaturatedHeterocycles**: returns the number of saturated heterocycles for a molecule
- **NumSaturatedRings**: Number of Saturated Rings
- **NumSpiroAtoms**: Count of atoms that serve as shared nodes in spiro ring systems
- **NumUnspecifiedAtomStereoCenters**: Count of stereocenters present in the molecule whose chiral configuration is unspecified
- **RingCount**: Number of All Rings
- **Phi**: Flexibility Index / Kier Flexibility Index – A metric of overall molecular flexibility computed according to topological and valence properties
- **MolLogP**: octanol/water partition coefficient (Wildman-Crippen LogP value)
- **MolMR**: Molar Refractivity(Wildman-Crippen MR value)
- **fr_Al_COO**: Number of aliphatic carboxylic acids
- **fr_Al_OH**: Number of aliphatic hydroxyl groups
- **fr_Al_OH_noTert**: Number of aliphatic hydroxyl groups excluding tert-OH
- **fr_ArN**:Number of N functional groups attached to aromatics
- **fr_Ar_COO**: Number of Aromatic carboxylic acide
- **fr_Ar_N**: Number of aromatic nitrogens
- **fr_Ar_NH**: Number of aromatic amines
- **fr_Ar_OH**: Number of aromatic hydroxyl groups
- **fr_COO**: Number of carboxylic acids

- **fr_COO2**: Number of carboxylic acids
- **fr_C_O**: Number of carbonyl O
- **fr_C_O_noCOO**: Number of carbonyl O, excluding COOH
- **fr_C_S**: Number of thiocarbonyl
- **fr_HOCCN**: Number of C(OH)CCN-Ctert-alkyl or C(OH)CCNcyclic
- **fr_Imine**: Number of Imines
- **fr_NH0**: Number of Tertiary amines
- **fr_NH1**: Number of Secondary amines
- **fr_NH2**: Number of Primary amines
- **fr_N_O**: Number of hydroxylamine groups
- **fr_Ndealkylation1**: Number of XCCNR groups
- **fr_Ndealkylation2**: Number of tert-alicyclic amines (no heteroatoms, not quinine-like bridged N)
- **fr_Nhpyrrole**: Number of H-pyrrole nitrogens
- **fr_SH**: Number of thiol groups
- **fr_aldehyde**: Number of aldehydes
- **fr_alkyl_carbamate**: Number of alkyl carbamates (subject to hydrolysis)
- **fr_alkyl_halide**: Number of alkyl halides
- **fr_allylic_oxid**: Number of allylic oxidation sites excluding steroid dienone
- **fr_amide**: Number of amides
- **fr_amidine**: Number of amidine groups
- **fr_aniline**: Number of anilines
- **fr_aryl_methyl**: Number of aryl methyl sites for hydroxylation
- **fr_azide**: Number of azide groups
- **fr_azo**: Number of azo groups
- **fr_barbitur**: Number of barbiturate groups
- **fr_benzene**: Number of benzene rings
- **fr_benzodiazepine**: Number of benzodiazepines with no additional fused rings
- **fr_bicyclic**: Number of Bicyclic groups
- **fr_diazo**: Number of diazo groups
- **fr_dihydropyridine**: Number of dihydropyridines
- **fr_epoxide**: Number of epoxide rings
- **fr_ester**: Number of esters
- **fr_ether**: Number of ether oxygens (including phenoxy)
- **fr_furan**: Number of furan rings
- **fr_guanido**: Number of guanidine groups
- **fr_halogen**: Number of halogens
- **fr_hdrzine**: Number of hydrazine groups
- **fr_hdrzone**: Number of hydrazone groups
- **fr_imidazole**: Number of imidazole rings
- **fr_imide**: Number of imide groups
- **fr_isocyan**: Number of isocyanates
- **fr_isothiocyan**: Number of isothiocyanates
- **fr_ketone**: Number of ketones
- **fr_ketone_Topliss**: Number of ketones excluding diaryl, a,b-unsat. dienones, heteroatom on Calpha
- **fr_lactam**: Number of beta lactams
- **fr_lactone**: Number of cyclic esters (lactones)

- **fr_methoxy**: Number of methoxy groups (-OCH_3)
- **fr_morpholine**: Number of morpholine rings
- **fr_nitrile**: Number of nitriles
- **fr_nitro**: Number of nitro groups
- **fr_nitro_arom**: Number of nitro benzene ring substituents
- **fr_nitro_arom_nonortho**: Number of non-ortho nitro benzene ring substituents
- **fr_nitroso**: Number of nitroso groups, excluding NO_2
- **fr_oxazole**: Number of oxazole rings
- **fr_oxime**: Number of oxime groups
- **fr_para_hydroxylation**: Number of *para*-hydroxylation sites
- **fr_phenol**: Number of phenols
- **fr_phenol_noOrthoHbond**: Number of phenolic -OH groups, excluding *ortho* intramolecular H bond substituents
- **fr_phos_acid**: Number of phosphoric acid groups
- **fr_phos_ester**: Number of phosphoric ester groups
- **fr_piperdine**: Number of piperdine rings
- **fr_piperzine**: Number of piperzine rings
- **fr_priamide**: Number of primary amides
- **fr_prisulfonamd**: Number of primary sulfonamides
- **fr_pyridine**: Number of pyridine rings
- **fr_quatN**: Number of quarternary nitrogens
- **fr_sulfide**: Number of thioether
- **fr_sulfonamd**: Number of sulfonamides
- **fr_sulfone**: Number of sulfone groups
- **fr_term_acetylene**: Number of terminal acetylenes
- **fr_tetrazole**: Number of tetrazole rings
- **fr_thiazole**: Number of thiazole rings
- **fr_thiocyan**: Number of thiocyanates
- **fr_thiophene**: Number of thiophene rings
- **fr__unbrch_alkane**: Number of unbranched alkanes of at least 4 members (excludes halogenated alkanes)
- **fr_urea**: Number of urea groups

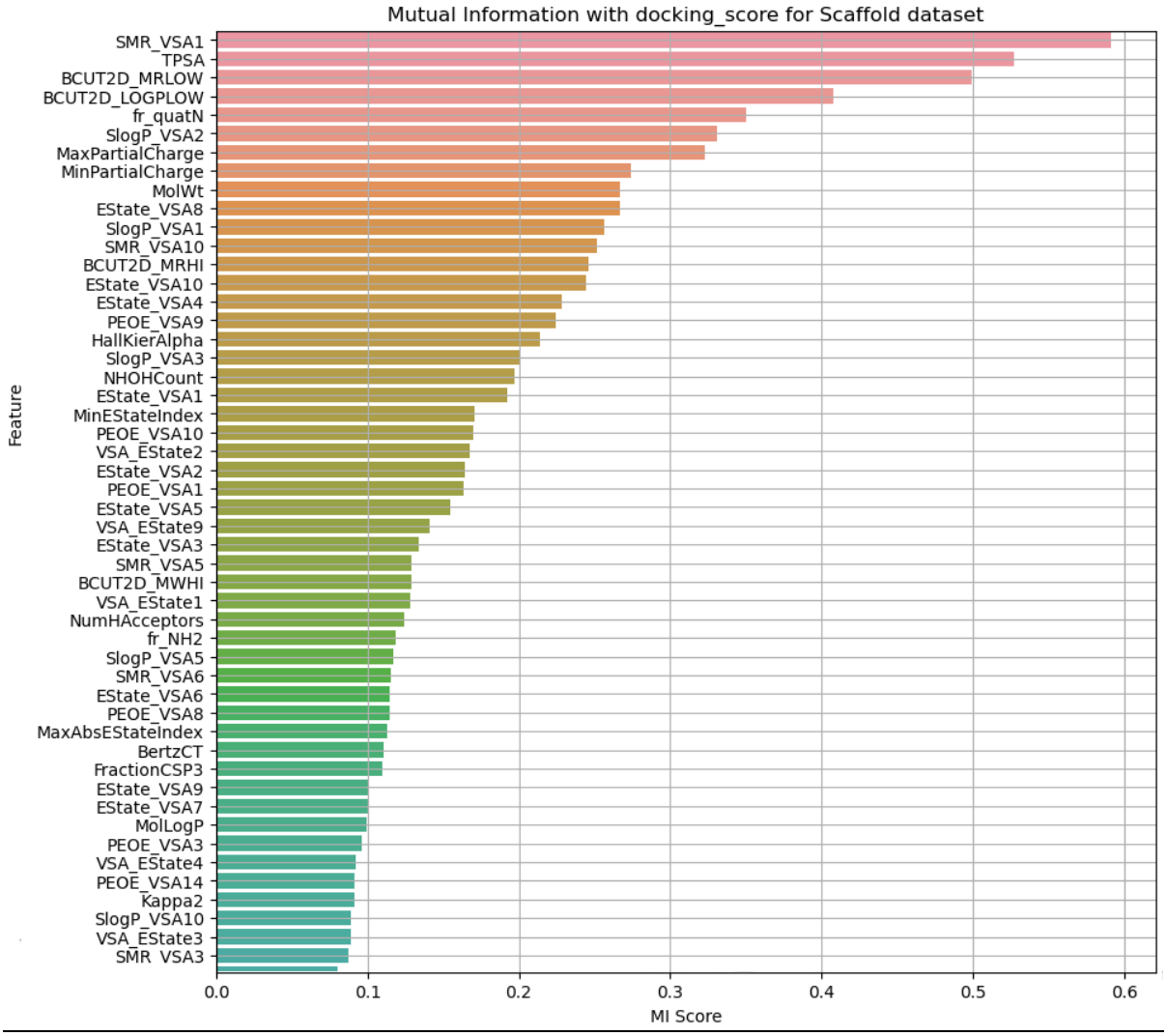

Figure S1: Top 50 RDKit chemical descriptors ranked by mutual information (MI) with docking scores for the **Scaffold**-based dataset.

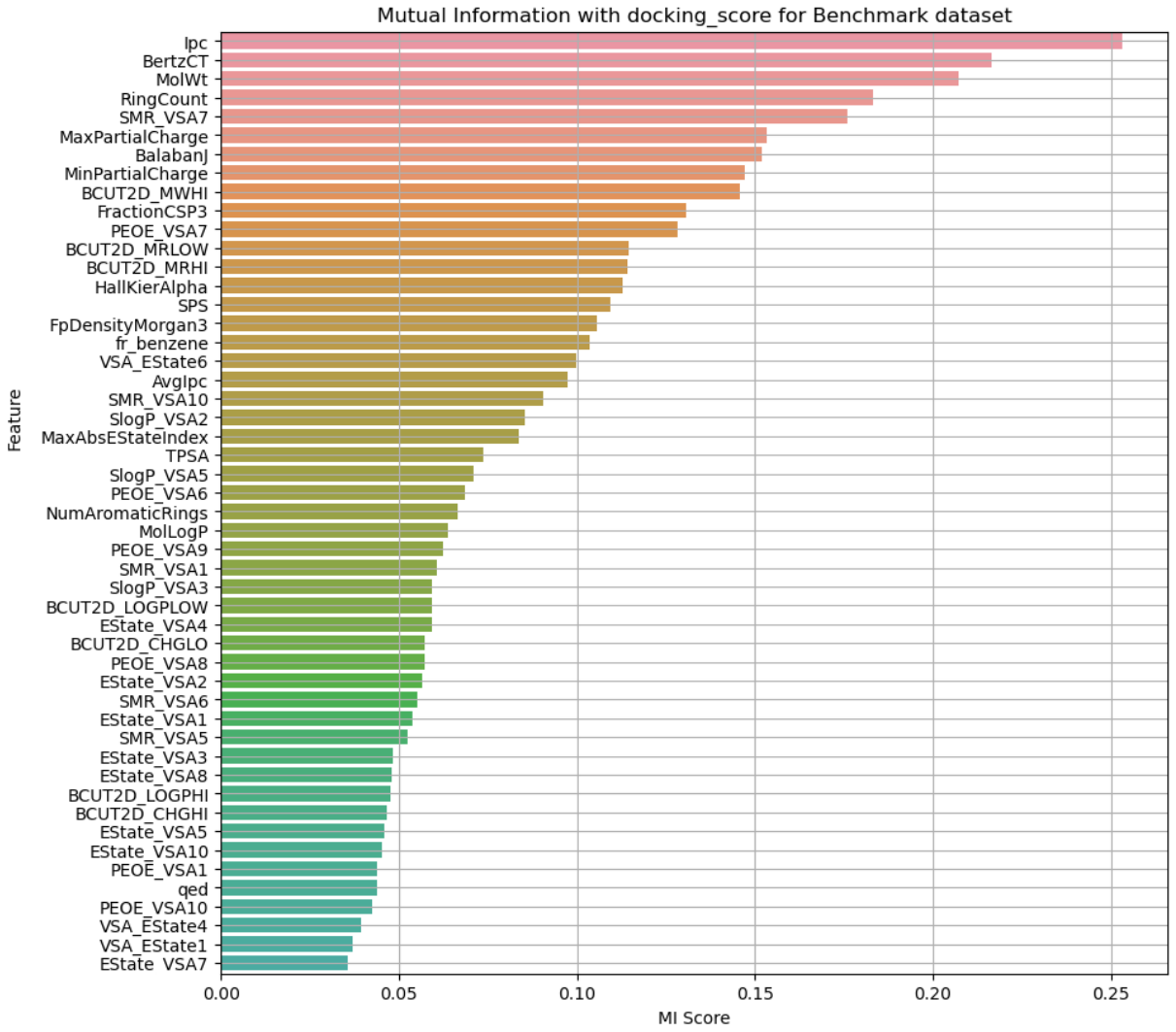

Figure S2: Top 50 RDKit chemical descriptors ranked by mutual information (MI) with docking scores for **Benchmark** dataset.

Dataset	Splitting Method	Subset Fraction	RMSE	R^2	Training Time [min.]
Scaffold	Random	1%	0.4736	0.8164	1.5220
Scaffold	Preserved	1%	0.4716	0.8179	1.3196
Benchmark	Random	1%	0.7343	0.5055	1.2515
Benchmark	Preserved	1%	0.7292	0.5124	1.0856
Scaffold	Random	5%	0.4421	0.8400	8.1097
Scaffold	Preserved	5%	0.4435	0.8389	7.0252
Benchmark	Random	5%	0.6839	0.5712	7.1124
Benchmark	Preserved	5%	0.6832	0.5720	6.2098
Scaffold	Random	10%	0.4309	0.8480	14.9277
Scaffold	Preserved	10%	0.4316	0.8474	14.1543
Benchmark	Random	10%	0.6626	0.5974	14.8106
Benchmark	Preserved	10%	0.6603	0.6002	13.2891
Scaffold	Random	25%	0.4209	0.8549	38.7931
Scaffold	Preserved	25%	0.4203	0.8553	34.9784
Benchmark	Random	25%	0.6328	0.6326	37.0249
Benchmark	Preserved	25%	0.6331	0.6325	33.9888
Scaffold	Random	50%	0.4173	0.8574	73.8511
Scaffold	Preserved	50%	0.4176	0.8572	71.2890
Benchmark	Random	50%	0.6104	0.6582	76.7545
Benchmark	Preserved	50%	0.6117	0.6571	70.2797
Scaffold	Random	75%	0.4170	0.8575	105.2672
Scaffold	Preserved	75%	0.4179	0.8569	102.3032
Benchmark	Random	75%	0.5974	0.6722	116.5927
Benchmark	Preserved	75%	0.5978	0.6725	114.8751

Table S1: RMSE, $R^{2,}$ and training time in minutes for training a Random Forest model on different subsets of Scaffold and Benchmark datasets using different splitting methods.