1. Rationale for Excluding Other Descriptors

1.1 Overview of Exclusion Strategy

To prevent the "Curse of Dimensionality" and ensure interpretability, only a limited number of non-redundant descriptors were retained. Although Mordred may provide over 1,600 descriptors, many of them are highly correlated or contain duplicate chemical information. The following categories were removed to reduce noise and multicollinearity while retaining important physicochemical signals.

1.2 Redundant Lipophilicity Measurements

- **Excluded:** MLogP, XLogP, nOctanol, FilterItLogP, WLogP, LogP_Consensus.
- **Reason:** All LogP variations have a high correlation and represent the same attribute. Including numerous variants pushes the model exclusively towards lipophilicity, overshadowing other features. **SLogP** is retained as the most robust and thoroughly proven method.

1.3 Explicit Atom Counts

- **Excluded:** nN, nO, nS, nF, nCl, nBr.
- **Reason:** Raw atom counts include information that is already captured by better aggregate descriptors:
    - **TopoPSA** identifies solvation-relevant heteroatoms more accurately than raw counts.
    - **MACCS keys** capture functional groups (such as halogens) more explicitly.
    - Removing atom counts minimizes feature redundancy while preserving chemical meaning.

1.4 Abstract Graph-Theoretical Indices

- **Excluded:** Wiener, Balaban, Zagreb, BertzCT, WalkCounts.
- **Reason:** These topological indices duplicate the information found in **Molecular Weight** and **VdwVolume** because they primarily correspond with molecule size. Additionally, their contribution to solubility lacks a clear physical explanation compared to volume or surface area.

1.5 3D Geometric Descriptors

- **Excluded:** WHIM, GETAWAY, RDF, 3D-MoRSE, Moments of Inertia, Radius of Gyration.
- **Reason:** Because 3D descriptors rely on a single calculated conformer, significant noise is introduced, and the dynamic solvated state is rarely

represented accurately. A more consistent and conformer-independent signal is produced by **2D topological descriptors**, such as KappaShapeIndex.

1.6 Autocorrelation Descriptors

- **Excluded:** Moreau-Broto, Moran, Geary, MATS/GATS.
- **Reason:** These descriptors are chemically ambiguous; different structures might provide the same value, and they encode mathematical patterns instead of comprehensible chemical properties. Clearer functional information related to solubility is provided by **MACCS keys**.

1.7 Molar Refractivity (MR)

- **Excluded:** MolarRefractivity, SMR.
- **Reason:** MR has a significant correlation with both **SLogP** (polarizability) and **Molecular Weight** (volume). The principle of descriptor orthogonality is violated by MR since both qualities are already present in the selected feature set.

References

[1] Mannhold, R., et al. (2009). "Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log P Methods on More Than 96,000 Compounds." *Journal of Pharmaceutical Sciences*, 98(3), 861-893.

[2] Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*. Wiley-VCH.

[3] Delaney, J. S. (2004). "ESOL: Estimating Aqueous Solubility Directly from Molecular Structure." *Journal of Chemical Information and Computer Sciences*, 44(3), 1000-1005.