Representation Matters - Fingerprints, Descriptors, and Beyond.

Sadia Chaudhry; Zenab Khan; Aaryan Jaitly; Atia Tul Wahab; Youssef Fathy

1 Introduction

Molecules exist in 3D chemical structure with complex arrangements. Machine learning algorithms, however, cannot process these chemical structures directly. A **molecular representation** translates these complex structures into machine-readable vectors. This process transforms the rich, continuous information of molecular geometry and electronic properties into a discrete set of **features** (also called **descriptors**), which are the individual numerical or binary variables that serve as the model's input. [1]

The choice of representation is fundamental. It predefines which chemical information shall be available to the model for learning. To effectively predict aqueous solubility, a representation format has to emphasize those features that encode the relevant physicochemical interactions governing dissolution. This report discusses the curation of features from two widely used paradigms for molecular representation: **descriptor based** (Mordred) and **fingerprint based** (MACCS) methods. [2]

1.1 Mordred Descriptors

Mordred is a freely available descriptor calculator, which computes over 1,800 features and calculate 2D and 3D descriptors from molecular structure (usually known as SMILES).

Mordred Descriptor Categories

Constitutional Descriptors (16 features): It counts atoms and bonds. Like nitrogen (nN), oxygen (nO), bond type counts. These directly reflect molecular composition.

Topological Indices (multiple modules with 100+ total): It encodes molecular branching and uses graph theory for connectivity. For solubility, branched molecules typically have higher solubility than linear isomers.

Autocorrelation Descriptors (606 features): It measures spatial corrections of atomic properties at varied distances.

Electrotopological State Descriptors (316 features): It evaluates the topology electronic properties. Each atom gets value according to electronic environment and connectivity called EState

Polarity and Hydrogen Bonding: It is considered as the most important predictor of solubility. This category of features predicts polar atom surface area.

Lipophilicity: octanol water partition coefficient is turned into SLogP.

3D Descriptors (when 3D conformers available): MoRSE encodes 3D shape via electron diffraction-like patterns. These capture 3D solvation properties.

1.2 MACCS Keys

MACCS (Molecular ACCess System) keys represent molecules as fixed 166-bit binary fingerprints, where each bit indicates presence/absence of a predefined structural pattern. [3]

MACCS Key Features and Categories

Atom Type Keys (26 keys): Identify element presence and properties. Key examples: isotopes, halogens, rare elements, charged atoms, ring atoms.

Functional Group Keys (140 keys): Indicate the existence of groups with chemical significance that are essential for solubility.

Ring System Keys (33 keys): Encodes aromatic systems with ring sizes ranging from 3M to 8M.

1.3 From Raw Representations to a Curated Feature Set

While both Mordred and MACCS represent molecular information in detail, the raw outputs of these software packages represent redundant, correlated, or irrelevant features for the task at hand: solubility prediction. Our feature selection strategy is delineated in the subsequent sections: identification of the chemically meaningful properties for solubility (Section 2) is followed by justification of the systematic exclusion of less informative descriptor categories in building a focused, interpretable, and high-performing feature set (Section 3).

2. Feature relevance for aqueous solubility prediction

By now, we have understood that, of course, not all the Mordred descriptors and MACCS keys are important for predicting solubility of a compound. Thus, in this section, the focus is on the features that make sense for predicting aqueous solubility.

Solubility prediction requires features that capture fundamental physicochemical interactions between molecules and water. Based on established principles from quantitative structure-property relationship (QSPR) studies, five key molecular characteristics are identified that determine aqueous solubility, and map these to corresponding Mordred descriptors and MACCS keys.[4–6]

2.1 Hydrophobic-Hydrophilic Balance (Lipophilicity)

Among all molecular properties, lipophilicity has the strongest influence on aqueous solubility. This predominance arises because lipophilicity directly quantifies the thermodynamic penalty of transferring molecules from organic to aqueous environments - the fundamental process of dissolution. The octanol-water partition coefficient (logP) captures this hydrophobic effect, making it the most consistent predictor across solubility models. Each unit increase in logP typically decreases solubility by one order of magnitude, reflecting the strong inverse relationship.

Solubility models based on fingerprints and physicochemical descriptors repeatedly identify lipophilicity measures such as logP and hydrophobic surface as major predictors, with higher lipophilicity generally associated with lower aqueous solubility.[4–6]

In our Mordred set this corresponds to **logP-like descriptors** such as **SLogP, ALogP** and **XLogP** (different calculation algorithms), while in MACCS it maps to **keys that encode hydrophobic patterns** like long carbon chains (81) and halogenated aromatics (44), which quantify how strongly a molecule prefers non-aqueous environments.[3,7]

2.2 Polarity and hydrogen-bonding capacity

Whereas lipophilicity imposes a hydrophobic penalty, polarity and hydrogen-bonding provide the favorable interactions that counteract hydrophobic penalties. Specifically, dipole-dipole interactions and hydrogen bonds between solute and water molecules stabilize the dissolved state, making molecules with higher exposed polar surface area and accessible hydrogen-bond donors or acceptors more soluble.

QSPR studies, including those by Li et al., Ghanavati et al., and Lee et al., consistently rank polar surface area and H-bonding functionality among the most important predictors of logS.

In our feature set, these effects are captured in Mordred by descriptors such as **topological polar surface area (TPSA),** which approximates the portion of the molecular surface available for polar interactions, and **counts of hydrogen-bond donors (nHBDon)** and **acceptors (nHBAccept)**. MACCS contributes this with binary **keys for polar functional groups (Bits 23-30)** such as hydroxyls, amides and carbonyls, providing a direct indication that hydrogen-bonding motifs capable of enhancing aqueous solubility are present.[2,3,7]

2.3 Molecular size and shape

Several studies report molecular weight, volume and simple shape descriptors as relevant for solubility, reflecting the trend that larger and bulkier molecules tend to be less soluble when polarity and lipophilicity are similar.[4–6]

Mordred provides **molecular weight**(MW), **van der Waals volume** and **shape** indices to quantify this, whereas MACCS captures size-related information more indirectly via **keys for ring systems** and **bulky hydrophobic fragments**, which help account for the geometric contribution to logS.[2,3,7]

2.4 Ionizable functional groups and charge

Ionizable groups can dramatically increase aqueous solubility because charged species interact much more favorably with water than neutral molecules at a given pH. This strong effect explains why features related to acidic/basic groups and charge distribution consistently appear as important predictors in solubility studies. [5–7]

Mordred descriptors that **quantify acidic and basic functionalities**, together with MACCS **keys for ionizable groups** (carboxylic acids, sulfonamides, amines), therefore represent an important class of features for accurate logS prediction.[2,3,7]

2.5 Aromaticity and hydrophobic π-systems

Extensive aromatic or π-conjugated systems, especially when combined with high lipophilicity, are frequently associated with reduced solubility because they increase rigid, hydrophobic surface area.[4–6] Planar aromatic structures promote π-π stacking in crystals and present flat hydrophobic surfaces that disrupt water's hydrogen-bond network.[5,6]

Mordred descriptors for **aromatic ring counts** and **aromatic atom fractions**, and MACCS **keys for (hetero)aromatic** and **halogenated rings**, capture this structural aspect and help the model learn how extended conjugation influences logS.[2,3,7]

2.6 Implications for our feature set

Across these studies, the best-performing solubility models combine information from lipophilicity, polarity and hydrogen bonding,

size and shape, ionization and aromaticity rather than relying on a single descriptor family.[4–6] This motivates explicitly retaining the corresponding Mordred descriptors and MACCS keys in our representation and later checking model feature importance against this chemically motivated subset to verify that the learned patterns are consistent with known solubility determinants.[3,7]

3. Rationale for Excluding Other Descriptors[8–10]

3.1 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.[9] The following descriptor categories were therefore removed to reduce redundancy while preserving the main physicochemical signals identified as relevant for aqueous solubility in Section 2.

3.2 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.

3.3 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.

3.4 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.

3.5 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.

3.6 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**.

3.7 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.

4 Additional features beyond Mordred and MACCS

Even after selecting structurally relevant features for solubility, there remain important physicochemical and experimental factors that MACCS and Mordred do not capture. Accurately forecasting how well a compound dissolves in water needs more than just its 2-D or 3-D structure. While MACCS and Mordred descriptors capture many molecular features, they miss crucial physicochemical information such as the compound's ionization behavior. The pKa determines the balance between neutral and charged species at a given pH, directly influencing aqueous solubility, yet this detail cannot be extracted from SMILES-based descriptors alone.

Equally important are properties that describe how tightly a solid holds together and how it partitions between phases. Melting point and lattice energy strongly affect solubility, but they are absent from traditional descriptor sets.[11] The solid-state form adds another layer of complexity: different polymorphs, hydrates, solvates, and especially amorphous versions can change solubility dramatically because amorphous forms possess higher free energy.

Experimental conditions further complicate matters. Variations in measurement technique (e.g., shake-flask, nephelometry, HPLC), temperature, agitation, and buffer composition lead to inconsistent reported values across datasets.[12] Since these external factors are not encoded by structural descriptors, they must be added as separate features for reliable models.[7]

In short, a robust solubility model should combine structural descriptors with pKa, melting point, log P/log D, solid-state characteristics, intrinsic solubility, and experimental conditions, all of which are missing from MACCS and Mordred alone.

These additional properties complement the structural features discussed in Section 2 and highlight that, for solubility, representation should extend beyond MACCS and Mordred alone.

5 Alternatives to Mordred and MACCS Descriptors for Solubility Prediction

5.1 Limitations of Our Current Representations

In Task 2 we mainly used two 2D representations: Mordred descriptors and MACCS fingerprints. Both turn a molecule into a fixed list of numbers or bits, which is convenient for machine learning but hides a lot of the chemistry behind solubility. From the literature we learned that aqueous solubility is governed by two energetic steps: breaking the solid crystal (crystal lattice / fusion) and stabilising the molecule in water (solvation).[13,14] Classical 2D descriptors mostly describe the solvation side (polarity, logP, size, H-bonding)

and give almost no direct information about crystal lattice strength or solid-state form. This makes our current representation particularly weak for high-melting "brick dust" compounds.

5.2 Graph Neural Networks (GNNs)

Graph Neural Networks offer a different way to represent molecules. Instead of hand-crafted descriptor lists, a GNN takes the molecular graph as input (atoms as nodes, bonds as edges) and learns its own vector representation through message passing. This keeps the full connectivity information and lets the model distinguish not just which groups are present, but also where they are and how they are connected. On curated solubility datasets, graph-based models can reach test RMSE values around 0.6–0.7 logS units, close to experimental uncertainty.[15] New explanation methods based on masking substructures help identify which rings or substituents drive the prediction, reducing the "black-box" feeling.[16]

5.3 3D Structural and Conformational Features

Solubility is inherently three-dimensional: molecules pack in 3D crystals and adopt 3D conformations in solution. Our Mordred/MACCS setup largely ignores this. Two types of information are especially missing: proxies for crystal lattice strength (for example melting point) and better 3D measures of shape and polar surface, such as 3D polar surface area or dipole moment computed over relevant conformers. These properties help tell apart classic "brick dust" drugs (rigid, high-melting solids limited by lattice energy) from very lipophilic, flexible "grease ball" compounds that are limited by poor hydration.[13] Adding selected 3D and melting-point-related features would give the model a handle on the energetic cost of leaving the crystal, which pure 2D fingerprints cannot encode.

5.4 Conclusion and Strategic Recommendation

Overall, MACCS and Mordred remain useful baselines, but they oversimplify the physics that we now know to be important for solubility. A sensible next step is to combine more expressive representations with a small number of key physicochemical descriptors. Concretely, we propose training a GNN on molecular graphs and augmenting its learned representation with experimentally measured or accurately predicted melting points (and, where available, simple 3D shape or charge descriptors).[13,15] This hybrid approach preserves the flexibility of representation learning while explicitly addressing one major limitation we observed in Task 2: the inability of our current 2D feature set to recognise when poor solubility is driven by strong crystal packing rather than by solvation alone.

References

1. 6.01__Molecular_Descriptors.
2. Szafarczyk, M., Ludynia, P. & Kierunek, K. *Projekt Dyplomowy A Python Library for Efficient Computation of Molecular Fingerprints Biblioteka Do Efektywnego Obliczania Fingerprintów Molekularnych w J˛ Ezyku Python*. (2024).
3. MACCSKeys.
4. Li, M. *et al.* Prediction of the Aqueous Solubility of Compounds Based on Light Gradient Boosting Machines with Molecular Fingerprints and the Cuckoo Search Algorithm. *ACS Omega* 7, 42027–42035 (2022).
5. Lee, S. *et al.* Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks. *ACS Omega* 7, 12268–12277 (2022).

6. Ghanavati, M. A., Ahmadi, S. & Rohani, S. A machine learning approach for the prediction of aqueous solubility of pharmaceuticals: a comparative model and dataset analysis. *Digital Discovery* 3, 2085–2104 (2024).

7. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: A molecular descriptor calculator. *J Cheminform* 10, (2018).

8. Delaney, J. S. ESOL: Estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci* 44, 1000–1005 (2004).

9. Todeschini, R. & Consonni, V. *Molecular Descriptors for Chemoinformatics Volume I: Alphabetical Listing Second, Revised and Enlarged Edition*.

10. Mannhold, R., Poda, G. I., Ostermann, C. & Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J Pharm Sci* 98, 861–893 (2009).

11. Oja, M., Sild, S., Piir, G. & Maran, U. Intrinsic Aqueous Solubility: Mechanistically Transparent Data-Driven Modeling of Drug Substances. *Pharmaceutics* 14, (2022).

12. Sorkun, M. C., Khetan, A. & Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data* 6, (2019).

13. Bergström, C. A. S. & Larsson, P. Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. *Int J Pharm* 540, 185–193 (2018).

14. Llompart, P. *et al.* Will we ever be able to accurately predict solubility? *Sci Data* 11, (2024).

15. Ulrich, N., Voigt, K., Kudria, A., Böhme, A. & Ebert, R. U. Prediction of the water solubility by a graph convolutional-based neural network on a highly curated dataset. *J Cheminform* 17, (2025).

16. Wu, Z. *et al.* Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nat Commun* 14, (2023).