Task 03: Representation Matters: Fingerprints, Descriptors and Beyond.

Submitted by: Group 02

Ahmed Hassaan
Maham Ayesha
Saleha Attar
Abdulrahman Walid
Fouzia Nasreen
Abdelsalam Helala

Morgan/ECFP:

Morgan or Extended Connectivity Fingerprints (ECFP) describe a molecule by encoding the local environment around each atom. Each atom is first assigned an identifier based on basic features such as atomic number, valence, hybridization, formal charge, and attached hydrogens. The algorithm then updates these identifiers over successive iterations by incorporating information from neighbouring atoms. As the radius expands, the method captures progressively larger circular substructures, producing a compact and chemically informative fingerprint [1].

The resulting environments are stored as bits in a fixed length vector, where each "1" marks the presence of a particular hashed substructure. These substructures frequently correspond to familiar medicinal chemistry features, including aromatic systems, heterocycles, hydrogen-bonding fragments and common functional groups. However, owing to this direct chemical interpretability, as well as their independence from molecular orientation, Morgan/ECFP fingerprints have become a standard tool for similarity searching, virtual screening, and other ligand-based modelling applications [2] [3]. Although ECFP provides a strong description of local molecular topology, it does not account for three dimensional conformations or long-range interactions within a molecule. For this reason, contemporary drug-discovery pipelines often pair ECFP with additional features such as physicochemical descriptors, graph-based representations or learned molecular embedding to achieve a more complete and informative characterization of molecular structure [4][5][6].

RDKit Classical Descriptors:

RDKit classical descriptors are atom type based numerical properties derived from established physicochemical models such as Crippen's logP fragments and Ertl's TPSA fragments [7], [8]. In cheminformatics, a representation refers to the transformation of a chemical structure which is usually a SMILES string or molecular graph, into a vector of interpretable features that describe size, polarity, lipophilicity, hydrogen-bonding capacity and structural motifs. RDKit's classical descriptors includes molecular weight, Crippen LogP, topological polar surface area (TPSA), hydrogen-bond donor/acceptor counts, rotatable bond count, aromatic ring count and molar refractivity. Many of these originate from established models in medicinal chemistry [7], [8]. These descriptors have formed the backbone of traditional QSAR modelling due to their interpretability and low computational cost.

These descriptors are relevant to binding affinity prediction because they capture key physicochemical determinants of ligand protein recognition. For example, LogP and TPSA

influence desolvation penalties and hydrophobic complementarity. Both of these are central energetic terms in the thermodynamic analysis of binding by Klebe and Böhm [9]. Also, hydrogen bond donor/acceptor counts are basic approximations of a molecule's capacity to form directional interactions that contribute to binding enthalpy [10]. Whereas, molecular weight and rotatable bond count relate to steric fit and conformational entropy loss upon binding. Although these descriptors do not capture full 3D structure or explicit energetic terms but they are highly relevant for binding affinity measurements.

Many ligands share characteristic physicochemical patterns in case of dopamine D2 receptor (DRD2) such as moderate lipophilicity, polar functional groups and basic amines important for receptor engagement [11]. RDKit classical descriptors do not directly encode receptor specific interactions but indirectly reflect these tendencies through features such as formal charge, hydrogen-bonding capacity, and polarity metrics. This makes them a useful baseline when modelling DRD2 affinity, even though more receptor aware or 3D features may be required for full accuracy.

RDKit classical descriptors remain limited by their purely 2D and additive nature. They cannot represent binding pose, induced fit, solvation structure or water displacement. These are the factors known to strongly influence affinity [9], [12]. As such, classical descriptors offer a valuable but incomplete picture. They are appropriate for initial ligand-based analysis but ideally complemented by more expressive structural or interaction driven representations in later modelling stages.

RDKit Graph Theory:

RDKit's shape and connectivity descriptors as for instance Chi, Kappa, Balaban J, and BertzCT are classic examples of graph-theoretic molecular indices [13]. They treat the molecule as a mathematical graph where atoms are vertices and bonds are edges, utilizing quantities like vertex degree and shortest-path distances to collapse the topology into single numeric values. The Chi connectivity indices and Kappa shape indices, introduced by Hall and Kier, operate on atom degrees and bond connectivity to quantify how linear or branched a molecular graph is, effectively encoding aspects of surface area and shape without requiring 3D coordinates [14]. Balaban's J index uses the distance matrix, for instance all shortest-path distances between atoms, to give a size-normalized measure of branching versus linearity, while Bertz's complexity index (BertzCT) combines edge counts and atom-type diversity into a single complexity score that increases with ring count, branching, and heteroatoms [15].

Physically, these abstract graph numbers correlate with coarse molecular properties like size, branching, rigidity, and complexity. These properties, in turn, influence boiling point, solubility, permeability, and often how a ligand can fit into a protein binding pocket. For example, Hall–Kier connectivity indices distinguish linear from branched isomers such as n-butane vs. isobutane in a way that mirrors differences in exposed surface area, while high BertzCT values typically signal molecules that are structurally intricate and potentially harder to synthesize [16]. Similarly, Balaban J and the Kappa indices separate extended, rod-like structures from compact, spherical, or highly branched ones, providing a 2D topological handle on "shape" that serves as a coarse proxy for how a ligand occupies space [14], [17].

However, because these descriptors are derived purely from the 2D molecular graph, they do not explicitly encode conformational flexibility, stereochemistry, or the precise 3D arrangement of pharmacophoric features. Consequently, they are informative but incomplete representations of binding and should ideally be complemented by 3D or interaction-focused features in future models [15].

Redundancy and Correlation:

RDKit descriptors often include multiple features that end up describing nearly the same thing. Many size dependent values like molecular weight, number of heavy atoms, or surface area related measures tend to rise and fall together. Because they behave so similarly, they don't each provide new chemical information and instead introduce unnecessary duplication [18]. Correlation appears when several descriptors follow the same trend across different molecules. This is especially common with properties linked to size, branching, or aromatic character. When such highly correlated features are used in models such as linear regression or logistic classifiers, they can create multicollinearity, which makes the model less stable and harder to interpret [19].

Several RDKit's topological and graph-theory descriptors stem from very similar mathematical ideas. Indices derived from adjacency information, distance relationships, or connectivity patterns often end up grouping together because they reflect comparable aspects of molecular structure. Removing some of these overlapping descriptors generally leads to cleaner datasets and more reliable models [20]. Morgan fingerprints can also carry repeated information. A single structural motif like an aromatic ring or a common heterocycle may trigger several fingerprint bits at once. When models such as Random Forests, XGBoost, or neural networks are trained on these fingerprints, the extra correlated bits increase dimensionality without adding real value and may even hide important patterns [21].

To deal with redundant information, techniques such as PCA, feature clustering, variance filtering, and mutual-information methods are commonly applied. These approaches help narrow the representation down to the most meaningful features, improving the performance and interpretability of QSAR and affinity prediction models [22].

Conformers and Pharmacophores:

In all representations we used so far, the molecule is treated as a 2D graph or a set of counted features. Morgan fingerprints focus on local connectivity, and RDKit descriptors summarize properties like size, polarity and hydrogen-bonding capacity. These are useful, but they completely ignore how the molecule actually looks in three dimensions. In reality, a ligand is not flat. It can adopt many 3D shapes, called conformers, because single bonds can rotate. Only some of these conformers will fit well into the dopamine D_2 receptor pocket and form the right interactions. Binding affinity therefore depends on the 3D shape of the ligand and how well this shape complements the protein surface, not only on the 2D graph description. Models that only see 2D features can partially learn these effects, but they miss important geometric details that contribute to ΔG, such as exact distances, angles and packing in the pocket [23].

Pharmacophores were introduced exactly to capture this "important 3D information" in a simpler way. A pharmacophore describes a small set of key interaction features (for example:

hydrogen-bond donors or acceptors, positive or negative charges, aromatic rings, hydrophobic groups) and the distances and angles between them in space [24], [25]. For DRD2 antagonists, typical pharmacophore patterns include a basic (protonated) amine that can interact with a conserved Asp residue, at least one aromatic or heteroaromatic ring for π–π or hydrophobic contacts, and additional hydrogen-bonding features in defined relative positions [26]. Our current descriptors do count donors, acceptors and aromatic rings, but they do not know *where* these features are in 3D or how far they are from each other. This means two ligands with very different pharmacophore geometry can still look quite similar in our 2D fingerprints.

Because of this, a natural next step for future work is to complement our current 2D-based models with simple 3D and pharmacophore-based features. One option is to generate a small conformer ensemble for each ligand (for example with RDKit) and then compute 3D pharmacophore fingerprints or shape descriptors on these conformers [24]. Another option, now that high-resolution DRD2 structures are available [26], is to derive basic protein–ligand interaction fingerprints (hydrogen bonds, salt bridges, aromatic contacts) from docked poses for a subset of compounds. These features could then be combined with Morgan fingerprints and RDKit descriptors in a hybrid model and compared against our current baseline.

Conclusion and Synthesis:

1. The Strengths of 2D Representations: Our investigation into molecular representations reveals that standard 2D approaches provide a robust baseline for binding affinity prediction.

a) Morgan Fingerprints (ECFP): Effectively encode the local chemical environment and functional groups, capturing specific substructures like aromatic systems and hydrogen-bonding fragments that are essential for medicinal chemistry.
b) RDKit's classical descriptors: Translate the chemical structure into interpretable physicochemical properties. Metrics such as LogP and TPSA act as proxies for thermodynamic driving forces like desolvation penalties and hydrophobic complementarity.
c) RDKit Graph Theory: Simultaneously, graph-theoretic indices (e.g., Kappa, Balaban J) offer a computational method to quantify molecular shape and branching complexity without requiring coordinate generation.

2. The Challenge of Redundancy: While these descriptors are informative, they are not statistically independent. Our analysis highlights significant redundancy and correlation among features, particularly those related to molecular size and complexity. The presence of multicollinearity in these high-dimensional vectors can destabilize linear models and obscure the true drivers of affinity. Therefore, dimensionality reduction techniques (such as PCA or variance filtering) are a strict requirement before training machine learning models on this dataset.
3. The Critical Gap: Geometry and Interaction The most significant limitation identified across all current representations is the lack of 3D spatial awareness. Morgan

fingerprints and RDKit descriptors treat molecules as static 2D graphs, ignoring conformational flexibility and the specific spatial arrangement of pharmacophoric features. Binding affinity is fundamentally a 3D geometric, can capture the surround environment of the interactions, such as distances and angles between hydrogen bond donors/acceptors. and energetic fit between the ligand and the protein pocket (e.g., DRD2).

4. Future Directions To transcend the limitations of the current baseline, future modeling efforts must move beyond 2D topology. The integration of 3D pharmacophore fingerprints and conformer ensembles is necessary to represent the spatial reality of ligand-receptor interactions. Combining these explicit geometric features with the established physicochemical baselines (Hybrid Models) offers the most promising path toward accurate affinity prediction.

References:

1. D. Rogers and M. Hahn, "Extended-Connectivity Fingerprints," pp. 742–754, 2010, doi: 10.1021/ci100050t.
2. A. Cereto-massagué, M. José, C. Valls, M. Mulero, S. Garcia-vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening," vol. 71, no. 2015, pp. 58–63, 2023.
3. C. Geniesse, A. S. Pappu, and V. Pande, "Chemical Science MoleculeNet : a benchmark for molecular machine learning †," pp. 513–530, 2018, doi: 10.1039/c7sc02664a.
4. S. Riniker and G. A. Landrum, "Open-source platform to benchmark fingerprints for ligand-based virtual screening," 2013.
5. J. Jiménez-luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence".
6. K. Yang *et al.*, "Analyzing Learned Molecular Representations for PropertyPrediction," 2019, doi: 10.1021/acs.jcim.9b00237.
7. P. Ertl, B. Rohde, and P. Selzer, "Fast calculation of molecular polar surface area as a sum of fragment-based contributions," J. Med. Chem., vol. 43, no. 20, pp. 3714–3717, 2000.
8. S. A. Wildman and G. M. Crippen, "Prediction of physicochemical parameters by atomic contributions," J. Chem. Inf. Comput. Sci., vol. 39, no. 5, pp. 868–873, 1999.
9. G. Klebe and H.-J. Böhm, "Energetic and entropic factors determining binding affinity in protein–ligand interactions," J. Recept. Signal Transduct., vol. 17, pp. 459–473, 1997.
10. P. R. Connelly et al., "Enthalpy of hydrogen bond formation in a protein–ligand binding reaction," Proc. Natl. Acad. Sci. USA, vol. 91, no. 5, pp. 1964–1968, 1994.
11. S. Wang et al., "Structure of the D2 dopamine receptor bound to the atypical antipsychotic drug risperidone," Nature, vol. 555, pp. 269–273, 2018.

12. A. Tropsha, “Best practices for QSAR model development, validation, and exploitation,” Mol. Inform., vol. 29, pp. 476–488, 2010.
13. RDKit graph descriptors documentation: https://www.rdkit.org/docs/source/rdkit.Chem.GraphDescriptors.html
14. Hall, L. H., & Kier, L. B. (1991). The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. *Reviews in Computational Chemistry*, 2, 367-422.
15. Guha, R. (2007). A survey of quantitative descriptions of molecular structure. *Current Computer-Aided Drug Design*.
16. Bertz, S. H. (1981). The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12), 3599-3601.
17. Balaban, A. T. (1982). Highly discriminating distance-based topological index. *Chemical Physics Letters*, 89(5), 399-404.
18. Yap, C. W. (2011). PaDEL-descriptor: An open-source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7), 1466–1474.
19. Hawkins, D. M. (2003). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1), 1–12.
20. Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*. Wiley-VCH.
21. Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints (ECFP). *Journal of Chemical Information and Modeling*, 50(5), 742–754.
22. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
23. Gohlke H, Klebe G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed Engl. 2002 Aug 2;41(15):2644-76. doi: 10.1002/1521-3773(20020802)41:15<2644::AID-ANIE2644>3.0.CO;2-O. PMID: 12203463.
24. Lin S-K. Pharmacophore Perception, Development and Use in Drug Design. Edited by Osman F. Güner. *Molecules*. 2000; 5(7):987-989. https://doi.org/10.3390/50700987
25. Langer T, Krovat EM. Chemical feature-based pharmacophores and virtual library screening for discovery of new leads. Curr Opin Drug Discov Devel. 2003 May;6(3):370-6. PMID: 12833670.
26. Wang, S., Che, T., Levit, A. *et al.* Structure of the D2 dopamine receptor bound to the atypical antipsychotic drug risperidone. *Nature* 555, 269–273 (2018). https://doi.org/10.1038/nature25758