ECFP Transformations

Albert Dinstl

November 2025

Abstract

We present a comprehensive ablation study investigating the effects of various transformations on Extended-Connectivity Fingerprints (ECFPs) for molecular property prediction. Through 84 systematic experiments, we evaluate how affine transformations, noise injection, normalization techniques, and novel block radius linear mixing affect both distance preservation and downstream task performance. Our results demonstrate that count-based ECFPs exhibit greater robustness to noise than binary ECFPs, while most affine transformations (rotation, permutation, reflection, shear) preserve distances perfectly with minimal performance impact. We identify block radius linear mixing with ReLU activation as achieving the best classification performance (ROC-AUC = 0.8521 on BACE), while rotation of count ECFPs yields the lowest regression error (RMSE = 0.6775 on ESOL).

1 Introduction

Extended-Connectivity Fingerprints (ECFPs) have become a standard representation for molecular structures in cheminformatics and drug discovery [1]. These fixed-length binary or count vectors encode local chemical environments and have proven effective for molecular similarity assessment and property prediction tasks. However, the effect of various mathematical transformations on ECFP representations remains underexplored.

The ultimate goal of this research is to design a Graph Neural Network (GNN) architecture that produces structurally equivalent molecular embeddings to ECFPs when frozen, essentially establishing ECFP's downstream task performance as a lower bound. Understanding how different transformations affect ECFP representations is a crucial step toward this goal, as it reveals which mathematical operations preserve the essential structural information encoded in these fingerprints.

In this work, we systematically investigate how different classes of transformations affect both (1) the preservation of pairwise distance relationships between molecular representations, and (2) performance on downstream molecular property prediction tasks. We evaluate 84 different experimental configurations across three diverse molecular datasets, encompassing regression and classification tasks.

2 Methodology

2.1 Datasets

We evaluate our transformations on three benchmark molecular property prediction datasets:

- **ESOL (Aqueous Solubility):** Regression task predicting water solubility of compounds
- **LIPO (Lipophilicity):** Regression task predicting octanol-water partition coefficients
- **BACE:** Binary classification task predicting Beta-site APP Cleaving Enzyme 1 inhibition

2.2 Fingerprint Representations

We test two variants of ECFPs:

- **Binary ECFPs:** Standard bit-vector representations where each bit indicates the presence/absence of a structural feature
- **Count ECFPs:** Integer-valued vectors encoding the frequency of each structural feature

For block radius mixing experiments, we use multi-radius representations that concatenate ECFP features computed at radii 0–3, capturing both local and extended chemical environments.

2.3 Transformations Evaluated

We organize our experiments into five categories:

Baselines: Unmodified binary and count ECFPs establish baseline performance.

Affine Transformations: Six geometric transformations applied to fingerprint vectors:

- Rotation (orthogonal transformation)
- Permutation (coordinate reordering)
- Translation (additive shift)
- Shear (non-uniform scaling along axes)
- Reflection (coordinate sign flipping)
- Scaling (multiplicative factor)

Noise Injection: Gaussian noise ($\sigma = 0.2$) added to fingerprint components to evaluate robustness.

Normalization: Two standardization techniques:

- L2 Normalization (unit-length vectors)
- Standardization (zero mean, unit variance)

Block Radius Linear Mixing: Novel approach applying learnable linear transformations to multi-radius fingerprints with three activation functions:

- Identity (linear combination)
- ReLU (rectified linear activation)
- Tanh (hyperbolic tangent activation)

2.4 Evaluation Metrics

Downstream Task Performance:

- Root Mean Squared Error (RMSE) for regression tasks (ESOL, LIPO)
- ROC Area Under Curve (ROC-AUC) for classification (BACE)

Distance Preservation: We compute Spearman rank correlations between pairwise distances in the original and transformed spaces for 300 randomly sampled molecules per dataset. Three distance metrics are evaluated:

- Tanimoto distance (standard for binary fingerprints)
- Euclidean distance
- Cosine distance

Perfect preservation yields correlation = 1.0, while complete destruction yields correlation ≈ 0.

3 Results

3.1 Baseline Performance

Binary and count ECFP baselines establish the following performance levels:

3.2 Affine Transformations

Rotation: Achieves perfect distance preservation across all metrics (Tanimoto, Euclidean, Cosine ≥ 0.999) with the best overall regression performance. Count ECFP with rotation yields ESOL RMSE = 0.6775, the lowest error observed across all single-transformation experiments.

Dataset	Binary ECFP	Count ECFP
ESOL (RMSE ↓)	1.0992	**0.7606**
LIPO (RMSE ↓)	0.8260	**0.7302**
BACE (ROC-AUC ↑)	**0.8672**	0.8545

Table 1: Baseline performance for binary and count ECFPs. Count ECFPs outperform binary ECFPs on regression tasks, while binary ECFPs show slight advantage on classification.

Permutation: Exhibits perfect distance preservation (all correlations = 1.000) across all metrics and fingerprint types. Downstream performance remains comparable to baselines, with BACE Binary ROC-AUC = 0.8694. Recognized as the best transformation for distance preservation.

Translation: Shows selective distance disruption—Euclidean distance is perfectly preserved (correlation = 1.000), while Tanimoto (0.24–0.99) and Cosine (0.23–0.99) distances are partially degraded. Despite distance disruption, downstream performance remains reasonable (ESOL Count RMSE = 0.8161).

Shear: Maintains excellent distance preservation (Tanimoto = 0.9996–1.000) with strong downstream performance. BACE Binary achieves ROC-AUC = 0.8729, the best result among affine transformations for classification.

Reflection: Achieves near-perfect or perfect distance preservation across all metrics. Performance closely matches rotation and permutation, confirming that coordinate sign flips do not meaningfully alter the fingerprint's predictive information.

Scaling: Moderately degrades distance preservation (Tanimoto = 0.89–0.98 for binary, 0.95–0.98 for count) and downstream performance (ESOL Binary RMSE = 1.1338). Count fingerprints show better resilience to scaling than binary fingerprints.

3.3 Noise Injection

Gaussian noise ($\sigma = 0.2$) reveals differential robustness between fingerprint types:

- **Binary ECFPs:** Severely impacted with Tanimoto correlation ≈ 0.90 and ESOL RMSE degrading to 1.7148 (55% increase over baseline)
- **Count ECFPs:** Demonstrate robustness with Tanimoto ≈ 0.98–0.99 and ESOL RMSE = 0.9437 (24% increase). BACE classification remains nearly unchanged (ROC-AUC = 0.8580)

3.4 Normalization

L2 Normalization: Exhibits unique metric-specific behavior. Euclidean distance is severely disrupted (correlation $\approx$ 0.01–0.56) due to projection onto the unit hypersphere, while Cosine distance is perfectly preserved (correlation $\approx$ 0.999). Downstream performance remains comparable to baselines (ESOL Count RMSE = 0.9614).

Standardization: Causes severe distance destruction across all metrics (Tanimoto = 0.26–0.71) and yields the worst downstream performance among normalization techniques (ESOL Binary RMSE = 1.3319, ESOL Count RMSE = 1.3906). Surprisingly, BACE classification remains stable (Count ROC-AUC = 0.8561), suggesting classification tasks are more robust to this transformation.

3.5 Block Radius Linear Mixing

Multi-radius baselines concatenating ECFP features at radii 0–3 already improve performance:

- LIPO Count: RMSE = 0.6752 (best count-based result overall)
- BACE Binary: ROC-AUC = 0.8831 (best binary classification baseline)

Applying learnable linear transformations with different activations yields:

Identity Activation: Linear mixing without nonlinearity achieves LIPO Count RMSE = 0.6661 and ESOL Binary RMSE = 1.0465.

ReLU Activation: Produces the best overall results:

- ESOL Binary: RMSE = 1.0163 (best binary regression performance)
- BACE Binary: ROC-AUC = 0.8521 (**best classification performance overall**)

Tanh Activation: Comparable to identity activation with BACE Binary ROC-AUC = 0.8724, demonstrating that different nonlinearities offer complementary performance trade-offs.

3.6 Best Performing Configurations

Across all 84 experiments, the top-performing configurations are:

4 Discussion

4.1 Transformation Effects on Distance Preservation

Our systematic evaluation reveals distinct patterns in how transformations affect distance preservation:

Dataset	Best Configuration	Score
ESOL (RMSE ↓)	Count + Rotation	0.6775
LIPO (RMSE ↓)	Count + Block Mixing (Identity)	0.6661
BACE (ROC-AUC ↑)	Binary + Block Mixing (ReLU)	0.8521

Table 2: Best performing experimental configurations per dataset.

Perfect Preservation: Rotation, permutation, reflection, and shear transformations preserve all three distance metrics nearly perfectly (correlations $\geq$ 0.999), confirming that these isometric transformations maintain the geometric structure of the fingerprint space.

Metric-Specific Effects: L2 normalization demonstrates that transformations can selectively preserve certain metrics while destroying others. The perfect preservation of cosine distance but destruction of Euclidean distance reflects the mathematical properties of unit-sphere projection.

Degradation Patterns: Scaling shows moderate degradation, standardization shows severe degradation, and noise injection shows differential effects based on fingerprint type. These patterns suggest a hierarchy of transformation "severity" in terms of structural information loss.

4.2 Binary vs. Count ECFPs

Count-based ECFPs consistently demonstrate superior robustness to noise (Tanimoto correlation 0.98–0.99 vs. 0.90 for binary) and better baseline regression performance. However, binary ECFPs show slight advantages in classification tasks and are less affected by standardization's pathological behavior on count distributions.

4.3 Block Radius Linear Mixing

The multi-radius approach with learnable mixing successfully improves upon single-radius baselines, particularly with ReLU activation for classification. This suggests that learnable combinations of features at different radii capture complementary chemical information, and that nonlinear activation functions can enhance representation quality for specific tasks.

4.4 Implications for GNN Design

These findings provide crucial guidance for designing GNN architectures that mimic ECFP behavior:

- Permutation invariance (like graph neural networks naturally possess) aligns perfectly with ECFP properties, as permutation preserves all distances

- Rotation invariance should be incorporated where possible, as it maintains performance while enabling geometric flexibility
- Multi-scale aggregation (analogous to multi-radius ECFPs) with learnable mixing should be prioritized
- Count-based aggregation mechanisms may offer better robustness than binary indicators

5 Conclusion

Through this comprehensive ablation study of 84 experimental configurations across three molecular property prediction datasets, we have systematically characterized how various transformations affect ECFP representations. Our key findings include:

1. Most affine transformations (rotation, permutation, reflection, shear) preserve distance relationships perfectly with minimal impact on downstream performance
2. Count ECFPs exhibit significantly greater robustness to noise than binary ECFPs
3. L2 normalization creates metric-specific preservation patterns, while standardization severely degrades both distances and performance
4. Block radius linear mixing with learned transformations improves performance, with ReLU activation achieving the best classification results
5. The best regression performance (ESOL RMSE = 0.6775) comes from count ECFP with rotation, while the best classification (BACE ROC-AUC = 0.8521) uses binary block mixing with ReLU

These results establish a foundation for understanding ECFP transformation properties and provide empirical guidance for designing GNN architectures that leverage ECFP-like representations. Future work will focus on incorporating these insights into frozen GNN embeddings that match or exceed ECFP performance while maintaining interpretability and geometric properties.

References

[1] David Rogers and Mathew Hahn. "Extended-connectivity fingerprints". In: *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 742–754. DOI: 10.1021/ci100050t.