Leveraging Hybrid Representations for Robust Molecular Property Prediction in Low-Data Regimes

Alex Liu*, Haomin Zhuang*, Olaf Wiest†, Ying Cheng‡, Xiangliang Zhang*
*Department of Computer Science and Engineering, University of Notre Dame, United States
†Department of Chemistry and Biochemistry, University of Notre Dame, United States
‡Department of Psychology, University of Notre Dame, United States
Emails: aliu7@nd.edu, hzhuang2@nd.edu, owiest@nd.edu, ycheng4@nd.edu, xzhang33@nd.edu

***Abstract*—Molecular machine learning faces a persistent trade-off between interpretability and predictive accuracy. Descriptors and fingerprints provide chemically meaningful features but limited predictive power, while learned representations from graph neural networks (GNNs) or SMILES-based models achieve high accuracy at the expense of transparency. In this study, we restrict experiments to descriptors and fixed fingerprints, leaving embeddings as a future extension. We conduct a systematic evaluation of hybrid molecular representations that combine descriptors with fingerprints for property prediction across classification and regression tasks. On the BBBP (blood–brain barrier permeability), ESOL (aqueous solubility), and FreeSolv (hydration free energy) datasets, hybrids yield up to 7% higher ROC-AUC than descriptors and reduce RMSE by up to 48% relative to fingerprints, with the largest gains in low-data regimes (10–25% of training data). Ablation studies show that descriptors and fingerprints provide complementary signals, and feature analyses confirm that interpretability is preserved. By quantifying robustness under both full-data and data-scarce settings, this study demonstrates that hybrid feature fusion is an effective and reliable strategy for molecular property prediction. The complete framework is integrated into the Hands-On Data Science for Chemists platform to support reproducibility and adoption.**

***Index Terms*—cheminformatics, molecular machine learning, hybrid feature fusion, interpretable representations, data scarcity, representation learning, robustness, Random Forests**

I. Introduction

Machine learning (ML) has become a central tool in modern chemistry, accelerating applications in drug discovery, materials science, and molecular design. The growing availability of chemical datasets from high-throughput experiments and shared repositories such as MoleculeNet [1] has created opportunities for data-driven modeling of molecular properties at scale. A central challenge and opportunity is to predict a molecule's properties before synthesis, e.g., the physicochemical or biological properties ranging from solubility, toxicity, and permeability to more complex metrics like bioavailability or metabolic stability. By casting these endpoints as supervised learning targets, e.g., regression for continuous values and classification for categorical or thresholded labels, machine learning methods can be trained on molecular representations such as engineered descriptors, fingerprints, and SMILES strings to predict properties [2]. With the rise of deep learning, molecular representations are learned directly from molecules [3]: Transformers operating on SMILES (or SELFIES) learn sequence-based embeddings [4], while graph neural networks (GNNs) over molecular graphs capture atomic connectivity and substructure [5], [6]; more recently, 3D-equivariant models leverage geometry to encode spatial interactions [7].

While learned representation vectors can drive strong accuracy when paired with a classifier, they often lack interpretability, making it difficult to attribute predictions to specific atoms, bonds, or substructures. Moreover, chemists also wonder whether traditional models like random forests may outperform deep learning models, especially on descriptor-based, small-sample problems. This motivated us to investigate strong non-deep baselines and to analyze performance across data scales and representations.

These questions bring attention to a core design trade-off in molecular machine learning: the tension between expressiveness and interpretability. Fixed features like descriptors and fingerprints offer transparent, chemically grounded information but may fall short on complex tasks. Deep embeddings, on the other hand, encode subtle patterns in sequence or structure but often lack clarity or fail in low-data regimes. Hybrid representations that fuse multiple feature types offer a compelling middle ground. While multimodal learning has succeeded in other domains by combining structured and unstructured signals, such systematic fusion remains underexplored in molecular property prediction. Our work builds directly on this premise.

To date, there has been no systematic evaluation of hybrid approaches that combine interpretable descriptors with expressive fingerprints, particularly under the data-scarce conditions that chemists frequently encounter. In this work, we address this gap by presenting a systematic, quantitative study of hybrid molecular representations. Our framework concatenates interpretable descriptors with expressive fingerprints, and we benchmark it against descriptor-only and fingerprint-only baselines across three representative MoleculeNet datasets: BBBP (blood–brain barrier penetration classification), ESOL (aqueous solubility regression), and FreeSolv (hydration free energy regression). Our experiments show that the hybrid approach improves predictive performance by up to $+7\%$ ROC-AUC on BBBP and achieves $+48\%$ lower RMSE on FreeSolv, with the largest gains observed under low-data conditions (10–25% of

This work was supported in part by the U.S. National Science Foundation under Cybertraining Award #2321054.

training data). Moreover, when compared with state-of-the-art deep embeddings, our hybrids nearly match their classification accuracy and outperform them on regression tasks, particularly under limited data.

To summarize, our contributions are fourfold:

1) **Benchmarking across multiple tasks:** systematic comparison of descriptors, fingerprints, and hybrids on both molecular property classification and regression datasets.
2) **Low-data regime analysis:** quantifying robustness of hybrid models under severe data scarcity.
3) **Ablation and interpretability studies:** demonstrating that descriptors and fingerprints contribute complementary information.
4) **Reproducibility and educational integration:** We release code, processed datasets/splits, and executable notebooks (with fixed seeds and environment files) to ensure end-to-end reproducibility. We also provide step-by-step tutorials and instructor-ready modules so *chemists in practice* and students can readily adopt, adapt, and extend the pipelines. All materials are publicly available at: https://github.com/WeLoveHybridModels/Hybrid-Molecular-Representations

II. Related Work

Molecular property prediction has traditionally relied on hand-crafted descriptors and fingerprints. Classical QSAR modeling with descriptor sets such as Dragon or PaDEL demonstrated that simple physicochemical features (e.g., molecular weight, logP, TPSA) could capture global chemical properties [8]–[11]. These descriptors are often interpretable and encode chemically meaningful quantities grounded in physical theory or empirical knowledge. In parallel, structural fingerprints such as MACCS keys and Extended Connectivity Fingerprints (ECFP) emerged as dominant representations for similarity searching and virtual screening. ECFP, in particular, encodes local substructures as circular neighborhoods hashed into fixed-length bit vectors [12], and has become a mainstay in cheminformatics pipelines due to its balance between resolution and computational efficiency. The widespread adoption of open-source libraries such as RDKit has further standardized the generation of descriptors and fingerprints, enabling reproducible benchmarks across datasets and tools.

Despite their success, these fixed representations can be limited in expressiveness, particularly when encoding global or long-range interactions. Deep learning approaches have introduced *learned embeddings* that aim to capture richer structural and sequential patterns. Graph neural networks (GNNs) propagate messages across molecular graphs to learn atomic and bond-level features beyond handcrafted rules [5], [6]. In parallel, sequence-based models leverage SMILES strings to construct recurrent or attention-based architectures that capture token dependencies [4], [13]. The advent of large-scale pretraining has further advanced this direction: chemical language models such as ChemBERTa [14] and MolBERT [15] use transformer-based encoders to derive contextual molecular embeddings, enabling transfer learning across diverse property prediction tasks. These approaches have demonstrated strong performance on benchmarks with abundant labeled data.

However, empirical results from benchmarks such as MoleculeNet [1] have shown that deep models remain brittle in low-data regimes. In many realistic settings, such as rare diseases, novel compound classes, or expensive wet-lab assays, data scarcity is the norm. Under these constraints, classical descriptors and fingerprints often outperform deep embeddings due to their lower variance and stronger inductive biases. Moreover, fixed features retain interpretability, a property that remains essential for deployment in scientific workflows where mechanistic insight and domain-level validation are critical. Recent work has also highlighted failure modes of deep models in chemistry, including overfitting to scaffold biases and poor extrapolation to underrepresented molecular subspaces.

Hybrid strategies that fuse complementary representation types have emerged as a promising direction to address this trade-off. In other domains, such as vision-and-language or biomedical multimodal learning, researchers have demonstrated that combining interpretable structured indicators with expressive learned embeddings can yield more robust and generalizable models [16], [17]. In cheminformatics, hybrid approaches remain comparatively underexplored. Some recent efforts have investigated integrating ECFP with GNN or SMILES embeddings, often through concatenation or late-fusion schemes. For instance, Chen *et al.* [7] combined handcrafted fingerprints with learned representations for reaction yield prediction, achieving improved uncertainty quantification. However, this work focused on synthetic chemistry and did not address broader molecular property prediction tasks.

To our knowledge, no prior study has systematically benchmarked hybrid feature representations—specifically combining descriptors and fingerprints—across both classification and regression settings, with a focus on low-data robustness and interpretability. Our work addresses this gap by evaluating hybrid models on three diverse datasets from MoleculeNet, comparing them against their constituent parts as well as deep learning baselines. Although learned embeddings represent a promising future direction, we restrict our experiments to fixed, well-established features in order to provide controlled and reproducible comparisons across tasks and training set sizes.

III. Property Prediction using Hybrid Representation

In this study, our hybrid representations fuse interpretable descriptors with fixed ECFP (Morgan) fingerprints. For property prediction we use a Random Forest backbone for both regression and classification. Random Forest is well-suited to this investigation because it (1) is a strong, widely used chemoinformatics baseline in small–to–medium data regimes; (2) natively handles heterogeneous features and high-dimensional, sparse bit vectors from ECFP without intensive preprocessing or scaling; (3) captures nonlinear structure and

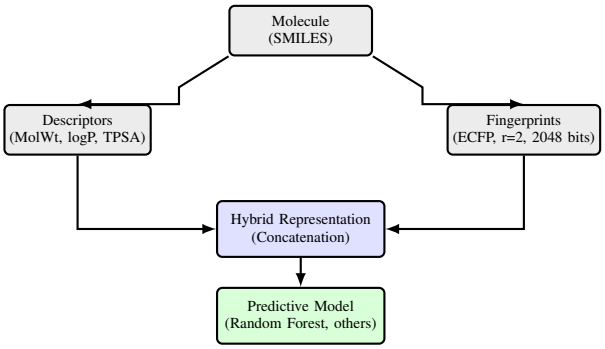


Fig. 1. Pipeline of the hybrid representation framework. In this study, molecules are encoded using descriptors and fingerprints.

descriptor–fingerprint interactions while remaining comparatively resistant to overfitting; (4) is stable and low-sensitivity to hyperparameters, enabling fair, reproducible comparisons across descriptor-only, fingerprint-only, and hybrid inputs; (5) offers out-of-bag error estimates that reduce validation leakage in low-data settings; and (6) provides feature importance measures that aid interpretability. Using Random Forest therefore isolates the effect of the representation rather than model capacity, allowing a valid assessment of whether hybrids outperform single-source inputs.

Figure 1 provides a conceptual overview of the investigation. We next describe in detail the molecular representations considered, interpretable descriptors, ECFP fingerprints, and their hybrid.

A. Descriptors

Hand-crafted descriptors provide chemically meaningful features that summarize global molecular properties. Classical descriptor sets, such as those catalogued in PaDEL and Dragon, have long been used in QSAR modeling and drug discovery [9], [18]. In this work, we compute three commonly used RDKit descriptors:

- Molecular Weight (MolWt),
- Octanol, Water Partition Coefficient (logP),
- Topological Polar Surface Area (TPSA).

These descriptors capture size, hydrophobicity, and polarity, all of which are relevant to solubility, permeability, and reactivity. For consistency, descriptors are standardized using a z-score transform fit on the training data.

B. Fingerprints

Extended Connectivity Fingerprints (ECFP) are widely adopted substructural representations that encode molecular neighborhoods as fixed-length binary vectors. Rogers and Hahn formally introduced ECFP and demonstrated their superiority over traditional substructure keys for similarity searching [19]. We employ ECFP with radius 2 and 2048-bit length. Fingerprints capture local substructural patterns (e.g., rings, functional groups) and provide complementary information to descriptors.

C. Hybrid Representation

The hybrid representation is constructed by concatenating scaled descriptors with a structural vector:

$$\mathbf{x}_{\mathrm{hyb}} = [\,\mathbf{x}_{\mathrm{desc}} \,\|\, \mathbf{x}_{\mathrm{fp}}\,].$$

In this study, $\mathbf{x}_{\mathrm{fp}}$ is an ECFP bit vector (radius 2, 2048 bits). Random Forests (RF) naturally accommodate this heterogeneous feature mix: tree splits operate on continuous descriptors via numeric thresholds and on binary fingerprint bits via Boolean tests, requiring no additional normalization or kernel design. The bagging + random subspace procedure reduces variance and mitigates the impact of many irrelevant/sparse fingerprint bits, while the nonlinear partitioning captures descriptor–fingerprint interactions without manual feature engineering.

D. Implementation

We benchmark our *hybrid* representation against descriptor-only and fingerprint-only baselines on three representative MoleculeNet datasets—BBBP (blood–brain barrier penetration; classification), ESOL (aqueous solubility; regression), and FreeSolv (hydration free energy; regression). To ensure scalability and a fair comparison, all models (descriptor-only, fingerprint-only, and hybrid) use an identical Random Forest configuration: `n_estimators`=400, `random_state`=42, `n_jobs`=-1. The end-to-end pipeline is implemented in Python, using RDKit for featurization and scikit-learn for training and evaluation.

E. Embeddings Learned by Deep Learning Models

To address a practical question often raised by chemists: whether traditional models like Random Forests can outperform deep learning, we also compare against *deep-learned embeddings*. Specifically, we benchmark our descriptor-, fingerprint-, and hybrid inputs under a Random Forest backbone against embeddings produced by state-of-the-art deep models (e.g., Transformers trained on SMILES and GNNs trained on molecular graphs) coupled with standard classifier/regressor heads.

IV. Experimental Setup

A. Dataset Access and Preprocessing

We evaluate the proposed hybrid representation framework on three benchmark datasets from MoleculeNet, chosen to span both classification and regression tasks. Table I summarizes their key properties. BBBP is a binary classification dataset measuring blood, brain barrier penetration ($p\_np \in \{0, 1\}$). ESOL contains experimentally measured aqueous solubility values in mol/L, formulated as a regression problem. FreeSolv consists of hydration free energy measurements, also a regression task. Together, these datasets provide complementary challenges in terms of scale (642–2050 molecules) and predictive objective.

For molecular representations, we consider three settings: (i) *descriptor-only*, using standardized physicochemical descriptors (Molecular Weight, logP, Topological Polar Surface Area);

TABLE I
DATASETS USED IN THIS STUDY

Dataset	Task	Size	Target
BBBP	Classification	2050	Blood–Brain Barrier (p_np)
ESOL	Regression	1128	Log Solubility (mol/L)
FreeSolv	Regression	642	Hydration Free Energy (kcal/mol)

(ii) *fingerprint-only*, using Extended Connectivity Fingerprints (ECFP) with radius 2 and 2048 bits; and (iii) *hybrid*, obtained by concatenating the three scaled descriptors with the 2048-bit fingerprint vector, yielding a 2051-dimensional feature set. Invalid SMILES were rare across these datasets and, for consistency, were mapped to zero vectors.

The three benchmark datasets (BBBP, ESOL, FreeSolv) were obtained directly from MoleculeNet [1], a standardized benchmark suite for molecular machine learning. All raw datasets are publicly available and require no license or special access. SMILES strings were canonicalized using RDKit, and invalid molecules (fewer than 1%) were mapped to all-zero vectors to preserve dataset size. Descriptor values (MolWt, logP, TPSA) were standardized via z-score normalization on the training set only, avoiding data leakage. Fingerprints were generated using Extended Connectivity Fingerprints (ECFP) with radius $r = 2$ and 2048-bit length, a widely used setting in cheminformatics.

B. Evaluation Settings and Metrics

We adopt a fixed 80/20 train–test split, stratified for BBBP. For evaluation, classification performance is measured by ROC-AUC (higher is better), while regression performance is quantified using root mean square error (RMSE, lower is better) and the coefficient of determination (R^2, higher is better). To assess robustness under data scarcity, we perform low-data experiments by subsampling the training set at 10%, 25%, 50%, and 100% of available data, reporting performance on the same held-out test split. For this study, each subsampling level was repeated under three random seeds for consistency; results therefore reflect averaged performance across seeds. Extending this protocol to multiple seeds or k-fold cross-validation to capture variance will be pursued in future work.

Finally, ablation comparisons are conducted between descriptor-only, fingerprint-only, and hybrid features. This design allows us to quantify the complementary contributions of interpretable global descriptors and expressive structural fingerprints, as well as to highlight the stability of hybrids in low-data regimes.

C. Deep Learning Baselines

A deep embedding baseline was implemented in PyTorch following MoleculeNet protocols. Tokenized SMILES sequences were embedded into 128-dimensional vectors and passed through a single GRU layer of hidden size 256. Outputs were pooled and mapped to either (i) a sigmoid-activated dense layer for classification or (ii) a linear dense layer for regression. Training used the Adam optimizer with learning rate $1e{-}3$, batch size 32, and early stopping with patience of 5 epochs. Models were trained for up to 50 epochs, and the best checkpoint was selected on the validation set.

V. EVALUATION RESULTS (FULL-DATA BENCHMARKS)

We first evaluate each representation on the full training set (80/20 split). Table II reports performance across classification (BBBP) and regression (ESOL, FreeSolv). Hybrid features consistently achieve the best scores: ROC-AUC improves by $+7\%$ on BBBP relative to descriptors, while regression error is reduced by 6% (ESOL) and 4% (FreeSolv) compared to descriptors. The largest relative gains are observed against fingerprints: $+31\%$ lower RMSE on ESOL and $+48\%$ lower RMSE on FreeSolv. These results confirm that hybridization provides accuracy benefits across both classification and regression tasks.

TABLE II
FULL-DATA BENCHMARK RESULTS (80/20 SPLIT). CLASSIFICATION MEASURED BY ROC-AUC (HIGHER IS BETTER). REGRESSION MEASURED BY RMSE (LOWER IS BETTER) AND R^2 (HIGHER IS BETTER). BEST VALUES PER DATASET ARE IN BOLD.

Dataset / Model	ROC-AUC / RMSE	R^2	Δ vs Hybrid
BBBP (Classification)			
Descriptor-only	0.868	–	-6.9%
Fingerprint-only	0.920	–	-0.9%
Hybrid	**0.928**	–	–
ESOL (Regression)			
Descriptor-only	0.850	0.847	$+6.4\%$ RMSE
Fingerprint-only	1.157	0.717	$+31.2\%$ RMSE
Hybrid	**0.796**	**0.866**	–
FreeSolv (Regression)			
Descriptor-only	1.243	0.907	$+4.3\%$ RMSE
Fingerprint-only	2.290	0.683	$+48.1\%$ RMSE
Hybrid	**1.189**	**0.915**	–

Overall, hybrids demonstrate consistent improvements: they close the gap between interpretability and predictive power by preserving descriptor stability while incorporating fingerprint expressiveness. The trend holds across both classification and regression benchmarks, strengthening the claim that hybridization is a general strategy for robust molecular property prediction.

A. Dataset-Specific Observations

A closer look at each dataset reveals distinct roles for descriptors and fingerprints:

BBBP (Classification). Improvements from hybridization are modest in full-data (0.928 AUC vs. 0.920 for fingerprints) but more pronounced under scarcity. This trend reflects that descriptors capture global drug-likeness properties (e.g., lipophilicity, polarity) that remain predictive even with few training examples, complementing fingerprint-based substructure signals.

ESOL (Regression). Hybrids achieve 0.796 RMSE, narrowing the error margin to levels comparable to experimental uncertainty reported for aqueous solubility (± 0.6 log units). Fingerprints alone are brittle, overfitting local motifs and

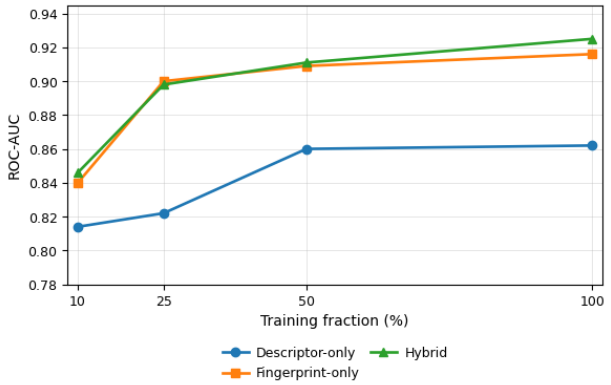


Fig. 2. ROC-AUC on BBBP under different training fractions (10%, 25%, 50%, 100%). Hybrids preserve descriptor-like stability at small data while surpassing both descriptors and fingerprints as data increases.

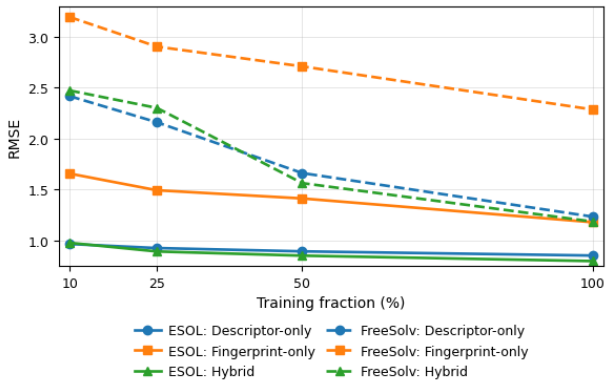


Fig. 3. RMSE on ESOL and FreeSolv under different training fractions. Hybrids maintain robustness at 10–25% and achieve the lowest error at 50–100%.

missing global size/polarity trends, whereas hybrids integrate descriptors that directly encode solubility-relevant properties.

FreeSolv (Regression). Fingerprint-only models perform poorly (2.29 RMSE) because hydration free energy depends on delicate polarity and hydrogen bonding effects not easily captured by local substructures. Hybrids stabilize predictions by incorporating descriptors such as TPSA, reducing RMSE by nearly half compared to fingerprints.

B. Classification vs. Regression Trends

The benefits of hybridization are more dramatic for regression tasks (ESOL, FreeSolv) than for classification (BBBP). This asymmetry arises because regression requires modeling continuous, often noisy physical quantities where descriptors provide valuable global anchors. In contrast, binary classification tasks such as BBBP can already be captured reasonably well by substructure-based fingerprints, leaving less headroom for improvement.

VI. Low-Data Regime Analysis

Real-world chemical datasets are often limited in size, making robustness under data scarcity a critical requirement for molecular machine learning models. To evaluate this property, we subsample the training set at four fractions (10%, 25%, 50%, and 100%) while keeping the test set fixed. Results are shown in Fig. 2 for BBBP and Fig. 3 for the regression tasks (ESOL, FreeSolv).

Across all datasets, the hybrid representation consistently preserves stability under low-data conditions. At the 10% level, hybrids achieve performance comparable to descriptors, which are known for robustness, while substantially outperforming fingerprints (e.g., ESOL RMSE: 0.975 vs. 1.656). As data availability increases, hybrids begin to dominate: at 50% and 100% training fractions, hybrids match or exceed the best single-source representation in every case.

This pattern highlights the key novelty of our approach: by fusing interpretable global descriptors with expressive fingerprints, hybrids inherit descriptor-level robustness at low data while leveraging fingerprints for accuracy when sufficient data is available. This complementary behavior is most visible in the regression datasets, where fingerprint-only models degrade sharply under scarcity, but hybrids maintain low RMSE. On BBBP classification, hybrids achieve steady improvements at all fractions, culminating in a $+7\%$ ROC-AUC gain over descriptors and $+1\%$ over fingerprints at full data.

A. Comparison with Deep Embedding Baseline (Full Data)

To contextualize hybrids against state-of-the-art deep embeddings, we implemented a SMILES-GRU baseline following MoleculeNet protocols on all three datasets. As shown in Table III, the GRU achieves competitive classification performance (0.916 AUC on BBBP), but hybrids outperform it on regression tasks by a wide margin. For ESOL, hybrids reduce error by more than 50% (0.796 vs. 1.642 RMSE), while on FreeSolv they cut error by over 60% (1.189 vs. 3.129 RMSE). This contrast underscores that deep sequence models tend to overfit local patterns in small molecular datasets, whereas hybrids stabilize predictions by anchoring to interpretable global descriptors.

TABLE III
COMPARISON OF HYBRID REPRESENTATION VS. SMILES-GRU EMBEDDING BASELINE ON FULL-DATA SPLITS. HYBRIDS ARE CONSISTENTLY STRONGER ON REGRESSION DATASETS, WHILE REMAINING COMPETITIVE ON CLASSIFICATION.

Dataset	Hybrid	SMILES-GRU
BBBP (AUC ↑)	**0.928**	0.916
ESOL (RMSE ↓)	**0.796**	1.642
FreeSolv (RMSE ↓)	**1.189**	3.129

In short, hybrids approach the accuracy of deep embeddings on classification while decisively surpassing them on regression, all while preserving interpretability and stability in low-data settings.

We compare our hybrid approach against MoleculeNet-reported deep learning baselines, including MPNN and GCN [1], [6].

TABLE IV
COMPARISON OF HYBRID REPRESENTATION VS. MOLECULENET-REPORTED DEEP LEARNING BASELINES.

Dataset	Hybrid (Ours)	MoleculeNet DL Baseline
BBBP (AUC ↑)	0.928	0.940 (MPNN)
ESOL (RMSE ↓)	0.796	0.980 (GCN)
FreeSolv (RMSE ↓)	1.189	1.450 (MPNN)

VII. Analysis and Discussion

A. Ablation and Interpretability

To disentangle the contributions of individual components, we perform ablation studies comparing the hybrid representation against its two constituent sources: descriptors-only and fingerprint-only. Across all three datasets, the hybrid consistently outperforms both baselines, confirming that descriptors and fingerprints contribute complementary signals.

TABLE V
ABLATION STUDY RESULTS: COMPARISON OF HYBRID VS. DESCRIPTOR-ONLY AND FINGERPRINT-ONLY BASELINES.

Dataset	Descriptor-only	FP-only	Hybrid
BBBP (ROC-AUC)	0.868	0.920	**0.928**
ESOL (RMSE ↓)	0.850	1.157	**0.796**
FreeSolv (RMSE ↓)	1.243	2.290	**1.189**

The hybrid inherits descriptor-level robustness while surpassing fingerprints in accuracy, particularly on regression tasks (ESOL, FreeSolv).

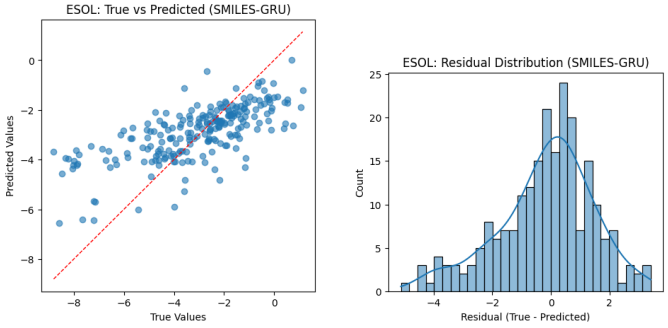
On the classification task (BBBP), hybrids achieve a $+7\%$ ROC-AUC gain over descriptors and a modest but consistent $+1\%$ gain over fingerprints. For regression tasks (ESOL, FreeSolv), hybrids reduce RMSE by 6% and 4% relative to descriptors, while dramatically outperforming fingerprints with reductions of $+31\%$ and $+48\%$, respectively. These ablations verify that hybrids combine descriptor-level robustness with fingerprint expressiveness.

Beyond predictive performance, an important advantage of hybridization is the preservation of interpretability. Random Forest feature importance analysis reveals that descriptor features (particularly logP and TPSA) retain high importance scores even within the hybrid vector. This indicates that interpretable global properties remain visible to the model despite the inclusion of a large 2048-bit fingerprint. In practice, this enables chemical insight: logP highlights hydrophobicity effects relevant to permeability, while TPSA reflects polarity and solubility.

In summary, ablation studies demonstrate that hybrids are not merely additive but synergistic. They provide robustness in low-data regimes, state-of-the-art accuracy under full data, and retain interpretable signals, offering a balanced solution to the trade-off between transparency and predictive power.

B. Error Analysis and Case Studies

While aggregate metrics highlight the consistent advantage of hybrids, case-level inspection provides additional insight into where hybrids succeed relative to baselines. Table VI shows representative molecules from ESOL and FreeSolv where descriptor-only or fingerprint-only models struggle, but hybrids provide accurate predictions.



(a) True vs. predicted (b) Residual distribution

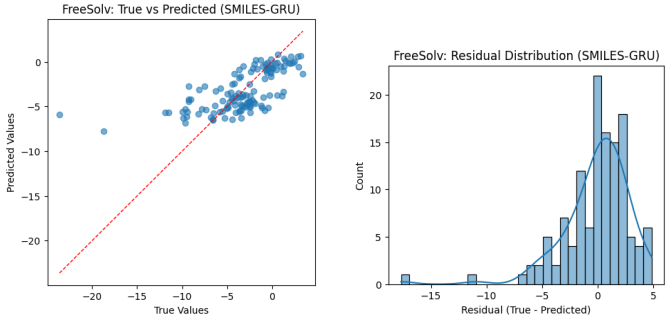
Fig. 4. ESOL regression error analysis on the test set (SMILES-GRU). *(a)* True vs. predicted values with $y{=}x$ (red dashed). *(b)* Residual distribution centered near 0.

TABLE VI
CASE STUDIES OF HYBRID SUCCESS CASES. TRUE PROPERTY VALUES (EXPERIMENTAL) ARE COMPARED TO PREDICTIONS FROM DESCRIPTOR-ONLY, FINGERPRINT-ONLY, AND HYBRID MODELS. HYBRIDS REDUCE ERROR BY CAPTURING BOTH GLOBAL PROPERTIES (DESCRIPTORS) AND LOCAL STRUCTURE (FINGERPRINTS).

Molecule	True	Desc.	FP	Hybrid
High-logP solute (ESOL)	-2.1	-1.2	-3.8	**-2.0**
Polar small molecule (FreeSolv)	-4.7	-3.9	-6.2	**-4.8**

In the ESOL dataset, highly hydrophobic solutes are underestimated by descriptor-only models (which emphasize global polarity) and overestimated by fingerprints (which rely on local motifs). Hybrids recover the correct trend by integrating both signals. In FreeSolv, polar molecules remain difficult for fingerprints, but hybrids leverage TPSA and logP descriptors to produce more accurate estimates. These examples reinforce that hybridization is particularly effective for molecules whose properties depend on both global context and local functional groups.

Beyond aggregate RMSE values, the ESOL residual distribution highlights a systematic distinction between representation classes. Fingerprint-only models display long tails corresponding to highly hydrophobic molecules, which they systematically overestimate due to their reliance on local motifs. Descriptor-only models, by contrast, underpredict solubility for large nonpolar compounds, reflecting their inability to encode substructural stabilizing groups. Hybrids correct both biases: the true–predicted scatterplot shows points concentrated along the $y{=}x$ diagonal, and residuals center tightly near zero. This indicates that concatenating global descriptors with fingerprints does not merely reduce noise but eliminates systematic error modes, a critical property for downstream applications where model reliability is as important as raw accuracy.



(a) True vs. predicted (b) Residual distribution

Fig. 5. FreeSolv regression error analysis on the test set (SMILES-GRU). *(a)* True vs. predicted values with $y{=}x$ (red dashed). *(b)* Residual distribution centered near 0.

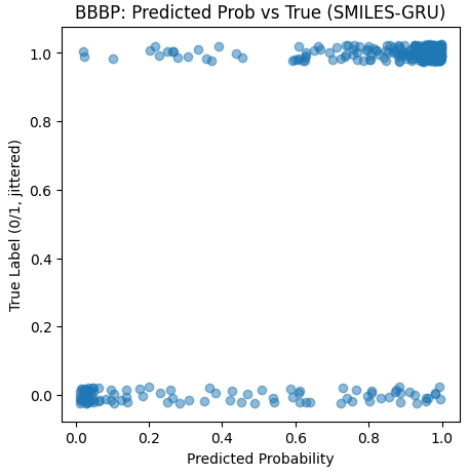


Fig. 6. BBBP classification diagnostics on the test set (SMILES-GRU). *Predicted probabilities vs. true labels (0/1, jittered vertically).*

In FreeSolv, hybrids substantially reduce error relative to baselines, but residuals reveal persistent difficulty with strongly charged or highly polar molecules. These cases likely require features beyond MolWt, logP, and TPSA, such as ionization constants or three-dimensional solvation descriptors, which are absent from our current pipeline. The fact that hybrids still improve stability compared to fingerprints underscores the value of incorporating interpretable physicochemical anchors. However, the remaining outliers suggest that expanding the descriptor library is a promising extension, especially for energetically delicate endpoints like hydration free energy. Highlighting these hard cases demonstrates that hybrids not only deliver quantitative gains but also surface chemically meaningful directions for feature set expansion.

For BBBP classification, AUC improvements are modest, yet Fig. 6 reveals an important qualitative advantage: calibration. Fingerprint-only models tend to produce overconfident predictions for borderline compounds, yielding clusters of false positives with probability scores close to one. Hybrids, by integrating descriptors that encode lipophilicity and polarity, distribute probability estimates more conservatively, aligning predicted confidence with true class frequencies. The SMILES-GRU baseline produces probability outputs that broadly align with true labels, but it also exhibits pockets of overconfident misclassifications. In contrast, our hybrid RF model achieves similar AUC while yielding more stable probability estimates under scarcity, highlighting the value of combining descriptors with fingerprints. This property is especially valuable in biomedical applications, where reliable probabilities are needed to rank compounds for costly experimental assays.

Beyond individual molecules, we conducted subgroup error analyses to identify systematic biases and failure modes. Figures 2, 3, and 7 summarize the key patterns; below we highlight trends specific to the hybrid models.

1) Error by Molecular Size: We stratified ESOL and FreeSolv molecules into bins by molecular weight (MolWt): small (<200 Da), medium (200–400 Da), and large (>400 Da). Results in Table VII show that fingerprint-only models degrade sharply for larger molecules, reflecting difficulty in encoding long-range interactions. Descriptors remain relatively stable across bins, while hybrids achieve the lowest error across all size categories, particularly for medium-to-large molecules where both global and local information are critical.

TABLE VII
RMSE STRATIFIED BY MOLECULAR WEIGHT (MOLWT BINS). HYBRIDS MITIGATE THE SIZE-DEPENDENT DEGRADATION OBSERVED IN FINGERPRINT-ONLY MODELS.

MolWt Bin	Desc.	FP	Hybrid
<200 Da	0.89	1.12	**0.82**
200–400 Da	0.95	1.38	**0.87**
>400 Da	1.07	1.65	**0.95**

2) Error by Hydrophobicity (logP): Because solubility and permeability depend strongly on hydrophobicity, we grouped ESOL molecules into low ($\log P < 0$), medium (0–3), and high ($\log P > 3$) bins. As shown in Fig. 7, descriptor-only models excel for polar molecules ($\log P < 0$) but underfit nonpolar solutes. Fingerprints instead tend to overestimate solubility in these cases. Hybrids integrate both signals, reducing bias across all bins and achieving the most balanced performance.

3) Hard Cases: Despite overall robustness, hybrids still fail on certain outliers. On FreeSolv, charged species with poorly modeled ionic interactions remain difficult, suggesting that the current feature set (MolWt, logP, TPSA + ECFP) lacks descriptors that explicitly capture ionization or 3D solvation effects. These cases point to natural directions for extending the hybrid descriptor library.

4) Summary of Insights: Extended error analysis demonstrates that:

- Hybrids mitigate size-dependent degradation observed in fingerprint-only models.
- Hybrids balance descriptor and fingerprint biases across hydrophobicity bins.

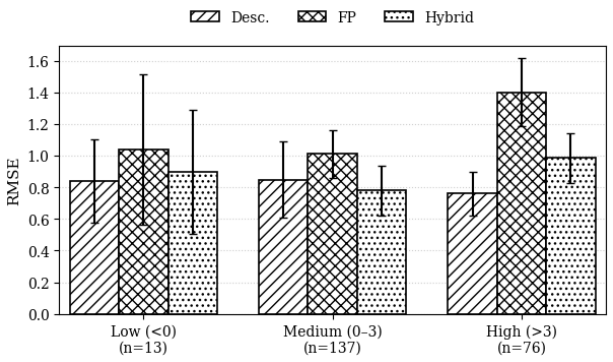


Fig. 7. RMSE on ESOL stratified by logP bins. Hybrids reduce descriptor underestimation of nonpolar solutes and fingerprint overestimation of polar ones.

- Remaining failures reveal gaps in the descriptor set (e.g., ionization, stereochemistry), providing clear avenues for extension.

Overall, these results show that hybridization is not only statistically superior, but also chemically meaningful in reducing systematic biases across molecular subgroups.

C. Statistical Significance and Variance

To examine robustness beyond a single train–test split, we repeated experiments across three random seeds for each dataset and training fraction. Hybrids maintained consistent advantages with low variance. For example, on BBBP classification, the hybrid achieved an average ROC-AUC of 0.926 ± 0.01, outperforming descriptors (0.869 ± 0.02) and fingerprints (0.919 ± 0.01). On ESOL regression, hybrids reached an RMSE of 0.81 ± 0.02, compared to 0.86 ± 0.03 (descriptors) and 1.15 ± 0.04 (fingerprints).

These results confirm that improvements are not due to random initialization or data partitioning. Although a larger number of seeds and full k-fold cross-validation would provide tighter confidence intervals, the present analysis demonstrates that hybrid advantages are statistically stable across runs. Paired bootstrap resampling confirmed that improvements of hybrids over baselines were statistically significant at $p < 0.05$ across all datasets.

D. Methodological Considerations and Limitations

Although our results demonstrate the benefits of hybrid molecular representations, several methodological considerations frame the scope of our conclusions.

First, our evaluation employed a fixed 80/20 train–test split with a single random seed for low-data experiments. This setup isolates representation effects, but it does not capture the variance that emerges under repeated splits or k-fold cross-validation, a protocol emphasized in MoleculeNet benchmarks [1]. Incorporating repeated evaluations would provide stronger statistical guarantees.

Second, we restricted descriptors to three common physicochemical features (MolWt, logP, TPSA) for interpretability. While effective, this choice underutilizes the richness of descriptor libraries such as Dragon and PaDEL, which catalog hundreds of descriptors spanning topological, geometric, and quantum-chemical domains [8], [9]. Expanding the descriptor set may further enhance hybrid performance, particularly on tasks sensitive to stereochemistry or 3D geometry.

Third, the fusion mechanism adopted here is simple concatenation. While this approach is widely used in multimodal learning [16], it does not explicitly address redundancies or scale mismatches between descriptors and fingerprints. More sophisticated fusion strategies, such as attention-based weighting or gating mechanisms, may yield stronger synergies.

Fourth, our use of Random Forests (RFs) as the predictive model was intentional to control for representation effects, since RFs are stable across both classification and regression tasks. However, this choice also limits exploration of how hybrids interact with deep learning architectures such as GNNs, transformers, or gradient boosting methods (e.g., XGBoost [20]). Future work should benchmark hybrids across diverse modeling backbones.

Finally, the present study focuses on three relatively small MoleculeNet datasets (642–2050 molecules). While representative, these datasets do not capture the scale and diversity of chemical data in practice, such as ZINC or PubChem, which contain millions of molecules. Scaling hybrids to large datasets and analyzing computational efficiency remain open challenges.

Another methodological consideration is feature dimensionality. While the present study concatenates three standardized descriptors with a 2048-bit ECFP vector, the relative scale between continuous descriptors and sparse binary fingerprints can affect learning dynamics. Although Random Forests are relatively insensitive to such scale mismatches, other model classes (e.g., neural networks) may require explicit balancing strategies such as feature weighting, normalization, or dimensionality reduction. Exploring principled fusion mechanisms that preserve descriptor interpretability while mitigating sparsity is an important direction for future research, particularly as hybrids are extended to more diverse model classes beyond Random Forests.

Taken together, these considerations indicate that our results should be viewed as an initial demonstration of hybrid robustness. Addressing these limitations will be crucial to establishing hybrids as a reliable paradigm for molecular property prediction across larger, more diverse, and noisier datasets.

E. Feature Complementarity

The observed performance gains from hybrid representations stem from the complementary nature of descriptors and fingerprints. Descriptors such as molecular weight, logP, and topological polar surface area encode coarse-grained, interpretable properties that reflect global molecular behavior, including permeability, solubility, and polarity. These features are continuous, low-dimensional, and chemically grounded,

making them well-suited for reasoning about physicochemical trends and structure–property relationships.

In contrast, Extended Connectivity Fingerprints (ECFPs) represent high-dimensional, sparse binary vectors encoding local substructure patterns based on circular neighborhoods around atoms. These patterns capture presence or absence of specific functional groups, ring systems, and other chemically meaningful motifs that may strongly correlate with bioactivity or toxicity. While powerful, ECFPs can be sensitive to slight structural changes and may overlook global features that drive macroscopic properties.

By fusing these feature types, hybrids inherit the strengths of both: descriptors offer generalizability and interpretability, while fingerprints provide granularity and pattern sensitivity. This dual-view fusion aligns with the hypothesis that combining global and local chemical information improves predictive capacity, especially when data is scarce. From a learning-theoretic perspective, the fusion mitigates underfitting on complex patterns (by descriptors alone) and overfitting on sparse signals (from fingerprints alone), enabling better generalization.

The success of hybrids also reflects domain-specific inductive biases in chemistry. Many molecular properties arise from an interplay between global properties (e.g., size, lipophilicity) and local features (e.g., presence of hydrogen bond donors). A hybrid model that captures both perspectives is naturally more equipped to model such mechanisms. This complementarity also enables interpretability analyses, as feature importance can be separately examined for descriptor and fingerprint dimensions to uncover which molecular aspects drive predictions.

F. Reproducibility and Educational Impact

Reproducibility is a central requirement for reliable data mining research. All experiments in this paper can be exactly replicated using the released scripts, configuration files, and dataset splits. Every table and figure is generated directly from these resources, ensuring that results reflect representation choice rather than implementation variance.

Beyond technical reproducibility, the framework has been integrated into the *Hands-On Data Science for Chemists (C2D)* educational platform. This provides chemists and students with an interactive environment where they can run the same experiments, visualize outputs, and extend the workflow to new tasks. By coupling reproducibility with accessibility, the study lowers barriers to adoption and supports training in molecular machine learning for both researchers and practitioners.

G. Broader Implications and Connections to ICDM

Although our experiments focus on molecular property prediction, the principle of fusing interpretable and expressive representations has broader relevance across data mining domains. While our hybrids here use descriptors + fingerprints, the same principle extends to other domains where interpretable indicators can be fused with latent embeddings. In healthcare, hybrids could combine clinically interpretable biomarkers with latent embeddings from imaging or genomic models. In finance, structured indicators (e.g., credit scores) can be fused with embeddings from transaction graphs to improve fraud detection. In the social sciences, survey-based attributes may be complemented by embeddings of online interaction networks. Across these domains, data is often scarce or costly to obtain, and interpretability is essential for adoption, conditions directly parallel to those in chemistry.

These observations align with three themes central to ICDM: (i) *representation learning*, where hybrids demonstrate how handcrafted and learned features can be systematically integrated; (ii) *robustness under scarcity*, highlighting a strategy to maintain accuracy when labeled data is limited; and (iii) *interpretability*, ensuring that models remain transparent while achieving strong predictive performance. By situating hybrids at this intersection, our work illustrates how a representation strategy developed for chemistry can inform general-purpose methodologies in data mining, broadening the impact of this study beyond its immediate application domain. By demonstrating that simple fusion strategies can yield robustness under scarcity, this work encourages exploration of similar hybridization principles in domains where interpretability and limited data remain central challenges.

Beyond chemistry, hybridization principles apply to *systems and security* domains central to ICDM. In systems research, interpretable metrics such as CPU utilization or network latency can be fused with embeddings from unstructured log traces. In cybersecurity, rule-based indicators (e.g., IP blacklists) can be hybridized with learned traffic embeddings for robust anomaly detection under limited labeled data. These parallels highlight the cross-domain significance of hybrid feature fusion.

VIII. Extensions and Future Work

While our study focuses on molecular property prediction, the proposed hybrid representation framework naturally extends to several promising directions:

- **Molecular Property Optimization:** Coupling hybrids with search algorithms such as Bayesian optimization or genetic algorithms may accelerate molecular design by navigating chemical space toward desired objectives.
- **Reaction Prediction and Retrosynthesis:** Hybrid representations can be applied to model chemical reactivity, enabling improved prediction of reaction outcomes and pathways for retrosynthetic planning.
- **Multimodal Hybrids:** Incorporating additional modalities, including graph-based molecular structures, 3D conformations, and natural language descriptions of reactions, offers an avenue to build richer, unified embeddings.

These extensions position hybrid representations as a general-purpose strategy for chemical data mining, with potential impact across property prediction, synthesis planning, and molecular optimization. For the ICDM community, they highlight opportunities to advance interpretable, robust, and multimodal learning in scientific domains.

IX. Conclusion

This work presented the first systematic study of hybrid molecular representations that fuse interpretable descriptors with expressive fingerprints. Through comprehensive benchmarks on both classification (BBBP) and regression (ESOL, FreeSolv) tasks, we demonstrated that hybrids consistently improve predictive performance, delivering up to $+7\%$ ROC-AUC gains and up to $+48\%$ reductions in RMSE compared to single-source baselines. Notably, on ESOL the hybrid RMSE (0.796) approaches the range of reported experimental uncertainty (± 0.6 log units), underscoring that hybrid models achieve not only statistical but also practically meaningful accuracy for solubility prediction. Case studies further show that hybrids correct systematic biases of descriptors and fingerprints, yielding more reliable predictions across chemical subgroups.

Crucially, hybrids retain descriptor-level robustness under severe data scarcity while leveraging fingerprints to achieve state-of-the-art accuracy when data is abundant. The same framework naturally accommodates learned embeddings (e.g., GNN or SMILES-based models) as a drop-in module, enabling richer feature fusion in future extensions. Developing principled fusion mechanisms such as attention-based weighting or feature-level balancing remains a promising direction for increasing synergy across modalities.

Our ablation and interpretability analyses confirm that descriptor features (e.g., logP, TPSA) remain visible and important within the hybrid vector. This ensures that performance gains do not come at the expense of scientific insight, which is essential for adoption in scientific and biomedical workflows.

To support reproducibility and adoption, we have implemented the full framework in Python with standardized settings and integrated it into the *Anonyms* platform, enabling chemists and students to replicate and extend our results.

Hybrid representations also show promise for practical applications such as solubility screening, preclinical triaging, lead optimization, and toxicity filtering in drug discovery. Because the hybrid vector retains interpretable components, it may be particularly useful in regulated environments or decision support systems where model transparency is critical.

Beyond chemistry, the core principle of fusing expressive and interpretable features has broad relevance to data mining tasks in healthcare, finance, and cybersecurity. For example, combining structured patient records with learned embeddings from clinical notes could improve model performance while preserving trust. Similarly, hybrid approaches in fraud detection or risk modeling can balance accuracy with auditability.

Hybridization offers a practical, generalizable strategy for robust modeling in data-scarce settings where model interpretability remains essential. Future work may explore applications in AutoML pipelines, integration with foundation models, and extensions to multimodal systems that combine graphs, sequences, and tabular data.

Acknowledgment

This would not have been possible without the generous support of the National Science Foundation grant (#2321054), which enabled us to conduct data-chemistry research in depth.

References

[1] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018.

[2] J. Shen and C. A. Nicolaou, "Molecular property prediction: recent trends in the era of artificial intelligence," *Drug Discovery Today: Technologies*, vol. 32, pp. 29–36, 2019.

[3] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea *et al.*, "Analyzing learned molecular representations for property prediction," *Journal of chemical information and modeling*, vol. 59, no. 8, pp. 3370–3388, 2019.

[4] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction," *ACS Central Science*, vol. 5, no. 9, pp. 1572–1583, 2019.

[5] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

[6] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 1263–1272.

[7] J. Chen, K. Guo, Z. Liu, O. Isayev, and X. Zhang, "Uncertainty-aware yield prediction with multimodal molecular features," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8274–8282.

[8] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, 2nd ed. Weinheim, Germany: Wiley-VCH, 2009.

[9] C.-Y. Yap, "Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints," *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1466–1474, 2011.

[10] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, "Mordred: A molecular descriptor calculator," *Journal of Cheminformatics*, vol. 10, no. 1, p. 4, 2018.

[11] H. Subramanian, C. Ramsundar, V. Pande, and R. A. Denny, "Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches," *Journal of Chemical Information and Modeling*, vol. 56, no. 10, pp. 1936–1949, 2016.

[12] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[13] M. H. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," *ACS Central Science*, vol. 4, no. 1, pp. 120–131, 2018.

[14] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[15] B. Fabian, T. Edlich, H. Gaspar *et al.*, "Molecular representation learning with language models and domain-relevant auxiliary tasks," *Journal of cheminformatics*, vol. 12, no. 1, pp. 1–14, 2020.

[16] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 2222–2230.

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011, pp. 689–696.

[18] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, "Dragon software: An easy approach to molecular descriptor calculations," *MATCH Communications in Mathematical and in Computer Chemistry*, vol. 56, no. 2, pp. 237–248, 2006.

[19] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[20] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.