

ArcMol Enables Task-Adaptive Spherical Representation Learning for Molecular Property Prediction

Lijuan Chen

chenlijuan125@163.com

Department of Hematology, State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University <https://orcid.org/0000-0002-8076-163X>

Yurong Zou

West China Hospital, Sichuan University <https://orcid.org/0009-0000-2677-511X>

Zhongning Guo

West China Hospital, Sichuan University

Zihan Zou

West China Hospital, Sichuan University

Duanyang Qin

West China Hospital, Sichuan University

Dingguo Xu

Sichuan University <https://orcid.org/0000-0002-9834-8296>

Taijin Wang

Zenitar Biomedical Technology Co., Ltd.

Article

Keywords:

Posted Date: April 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9018430/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

**ArcMol Enables Task-Adaptive Spherical Representation Learning for
Molecular Property Prediction**

Yurong Zou¹, Zhongning Guo¹, Zihan Zou¹, Duanyang Qin¹, Dingguo Xu³, Taijin
Wang^{2,*} and Lijuan Chen^{1,2,*}

¹State Key Laboratory of Biotherapy and Collaborative Innovation Center of
Biotherapy, West China Hospital, Sichuan University, Chengdu, 610041, China

²Chengdu Zenitar Biomedical Technology Co., Ltd., Chengdu, 610045, China

³MOE Key Laboratory of Green Chemistry and Technology, College of Chemistry,
Sichuan University, Chengdu, 610064, China

*Corresponding author: State Key Laboratory of Biotherapy and Collaborative
Innovation Center of Biotherapy, West China Hospital, Sichuan University, Chengdu,
610041, China and Chengdu Zenitar Biomedical Technology Co., Ltd., Chengdu,
610045, China

E-mail addresses: chenlijuan125@163.com.

Abstract

Molecular properties are measured under different experimental and representational settings, each capturing a distinct aspect of molecular behavior. Integrating such task-dependent molecular information remains challenging, especially when labeled data are scarce and structurally similar molecules exhibit markedly different properties. Here we introduce ArcMol, a task-adaptive framework for molecular property prediction that combines explicit physicochemical descriptors with learned molecular representations. ArcMol organizes these features in a hyperspherical latent space, where molecular similarity is defined by angular relationships. Across more than 150 molecular property prediction benchmarks, ArcMol achieves performance comparable to state-of-the-art methods, while showing improved robustness in low-data settings and for structurally similar molecules. Analyses of the learned representations reveal task-relevant separation, ordered property gradients, and coherent local neighborhoods. We demonstrate ArcMol in a real-world application targeting the kinase IRAK4, where it prioritizes candidate molecules for drug repurposing and leads to the identification of nanomolar type II inhibitors. These results suggest ArcMol as a general framework for molecular property prediction in discovery settings.

Introduction

Molecular property prediction is a central component of modern drug discovery, supporting applications ranging from early virtual screening to lead optimization and the assessment of safety and drug-likeness.^{1,2} Recent advances in high-throughput experimentation and computational chemistry have led to the accumulation of large molecular datasets, providing new opportunities for data-driven modeling.³ However, practical molecular property prediction remains challenging. Prediction tasks often differ substantially in their underlying physical meaning, data scale, and statistical distribution, and are frequently limited by data scarcity and property cliffs, where small structural variations give rise to large changes in molecular properties.⁴⁻⁶ These challenges hinder the development of stable and reliable models within a unified framework.

Recent progress in molecular pretraining has shifted the prevailing paradigm toward pretraining followed by task-specific fine-tuning to improve task relevance.^{7,8} This strategy has delivered substantial performance gains and partially alleviated the limitations of universal molecular representations.^{9,10} In most existing approaches,¹¹ task adaptation is achieved through modifications to model parameters or prediction heads.¹² Task-specific differences in the molecular information involved, discrimination patterns, and decision structures are instead absorbed implicitly by shared representations, rather than being modeled explicitly.^{6,13} Such approaches can be effective when sufficient training data are available. Their behavior under highly heterogeneous or data-limited conditions remains less well understood.¹⁴⁻¹⁶

In realistic molecular property prediction settings, different tasks emphasize distinct aspects of molecular structure.¹⁷⁻²⁰ Properties such as lipophilicity or solubility are largely governed by global physicochemical characteristics,¹⁹ whereas endpoints related to bioactivity or selectivity can be highly sensitive to subtle modifications of local functional groups or three-dimensional conformations.²⁰ Regression tasks often require preserving smooth and ordered property gradients, while classification tasks demand clear decision boundaries between closely related molecules.¹² These considerations suggest that effective task adaptation may require more than parameter updates alone. Task-aware selection and fusion of molecular information at the representation level, together with latent space organizations aligned with task-specific objectives, may therefore be beneficial.²¹⁻²⁴ The structural organization of the representation space may play an important role in supporting stable and discriminative predictions in high-similarity and low-data regimes.

Motivated by these considerations, we introduce ArcMol, a task-adaptive spherical representation learning framework for molecular property prediction. ArcMol integrates multiple molecular representations and adaptively selects and fuses them in a task-driven manner, organizing molecular embeddings within a hyperspherical latent space to support angular-based task discrimination. The framework incorporates heterogeneous molecular information, including explicit physicochemical descriptors and learned representations, into a unified latent embedding. ArcMol is evaluated on more than 150 fine-grained molecular property prediction tasks spanning both classification and regression settings, where it matches

the performance of current state-of-the-art methods and shows improved robustness under low-data and high-similarity conditions. Analyses of the learned representations provide insight into how task-relevant structure emerges in the latent space. ArcMol is further applied to real-world molecular discovery tasks targeting the therapeutically important kinase IRAK4, enabling drug repurposing and binding-mode-aware candidate prioritization, and leading to the identification of nanomolar type II IRAK4 inhibitors. These results suggest that ArcMol can serve as a general framework for molecular property prediction in discovery settings.

Results

Complementarity in Molecular Representation Space

To investigate how explicit chemical knowledge and implicit topological semantics interact within molecular representations, we conducted a systematic evaluation across 47 benchmark tasks from MoleculeNet.¹¹ By comparing the performance of 14 mainstream molecular embedding models²⁵⁻³³ across three evaluation settings (details in Supplementary Information Notes), including embeddings alone and embeddings augmented with physicochemical descriptors derived from RDKit³⁴ or symbolic regression,³⁵ we observed consistent and statistically significant performance gains from descriptor incorporation (Fig. 1a). These results reveal a clear complementarity between explicit and implicit molecular representations. While embedding models effectively capture complex topological patterns, learned representations alone are insufficient to achieve optimal performance across tasks. Incorporation of physicochemical descriptors introduces explicit feature references that systematically enhance predictive performance.

To further characterize representational differences between symbolic regression derived Formula features and general RDKit descriptors, we performed a metric based analysis (Fig. 1b). This analysis evaluates feature properties along four dimensions reflecting task relevance and information structure, including task alignment, information abundance, internal redundancy, and knowledge coverage. Across tasks, Formula features exhibit markedly stronger task alignment, as indicated by consistently a higher mean absolute correlations than RDKit descriptors (Fig. 1b).

This increased task specificity is accompanied by reduced information abundance, suggesting that Formula features encode highly concentrated and partially overlapping information. In contrast, RDKit descriptors exhibit lower redundancy and broader information coverage, highlighting a trade off between task sensitivity and representational completeness. Guided by this trade off, we compared feature selection strategies based on mutual information and linear correlation at $K = 110$ (Fig. 1c). Across both classification and regression benchmarks, mutual information based selection yields more robust performance than correlation based selection, and performance remains stable under permutations of descriptor input order. This robustness is further supported by permutation and feature ordering analyses (Supplementary Fig. S1).

The incorporation of physicochemical descriptors is associated with systematic changes in latent space organization and representational similarity (Fig. 1d to g). Using t-SNE visualization³⁶ and cluster quality metrics such as the Silhouette coefficient,³⁷ which reflects intra class compactness and inter class separation, and the Davies Bouldin index,³⁸ which measures cluster dispersion, we observed improved intra class compactness and inter class separation following descriptor integration (Fig. 1d, e). These changes are consistent with the introduction of explicit physicochemical constraints that guide data driven representations toward greater physicochemical coherence. To further examine relationships between physicochemical descriptors and learned molecular feature spaces, we applied nonlinear centered kernel alignment (CKA)³⁹ to quantify representational similarity across model families (Fig. 1f, g).

Rule based fingerprints such as MACCSKeys show high similarity to physicochemical descriptors, reflecting their shared reliance on predefined molecular features. In contrast, large scale pretrained models such as BioT5²⁶ and UniMol³³ exhibit substantially lower similarity and higher complementarity with descriptor based feature spaces, with complementarity values exceeding 0.75. Consistent with this complementarity, combining such models yields net performance gains in more than 40 percent of the benchmarks (Fig. 1h). Collectively, these results indicate that improved predictive performance is associated with the integration of complementary physicochemical and learned molecular information within a unified latent space.

Architecture of the ArcMol

Motivated by the observed complementarity between physicochemical descriptors and learned molecular representations, we developed ArcMol, a hierarchical predictive framework for molecular property prediction, whose architecture is illustrated in Fig. 2. ArcMol integrates diverse molecular inputs within a shared latent space, enabling explicit chemical descriptors and learned representations to be modeled jointly. To account for task dependent heterogeneity, the framework adaptively emphasizes informative chemical features and routes different representations to components specialized for capturing molecular topology or sequence patterns. The fused representations are constrained to a unit hypersphere, such that molecular similarity is governed primarily by angular relationships rather than feature magnitude.

Task specific supervision is incorporated through a dual head design tailored to different prediction settings. In classification tasks, latent representations are shaped by joint multilayer perceptron heads and angle based discrimination to promote clear separation between classes. In regression tasks, continuous properties are predicted using paired multilayer perceptron heads, with an auxiliary proxy based angular signal supporting local consistency within the spherical latent space. Detailed architectural components and mathematical formulations are provided in the Methods section.

Benchmark Performance on MoleculeNet

We benchmarked ArcMol on nine representative MoleculeNet datasets, comprising a total of 47 subtasks, and compared it with 17 state of the art baseline methods. These baselines span a broad range of modeling paradigms, including graph message passing networks such as D-MPNN,⁴⁰ graph contrastive learning approaches such as MolCLR²⁸ and MolMCL,⁴¹ three dimensional geometric pretraining models such as UniMol,³³ and sequence based large language models including MolT5,²⁵ BioT5,²⁶ and LLM4SD.⁴² Table 1 summarizes performance on both classification and regression benchmarks, evaluated using AUC and RMSE, respectively. Across these tasks, ArcMol achieves the best performance on six of the nine benchmarks and ranks close to the top performing method on the remaining tasks, indicating competitive and consistent performance across diverse molecular prediction settings.

In classification benchmarks, ArcMol performs competitively on datasets characterized by complex toxicity and bioactivity endpoints. On the SIDER dataset,

which focuses on multi organ toxicity prediction, ArcMol achieves an AUC of 73.0 percent, exceeding the previous best result reported for MolMCL by 4.3 percent. On the BACE benchmark for BACE one inhibition, ArcMol attains an AUC of 87.9 percent, compared with 85.7 percent obtained by the three dimensional pretraining model Uni Mol. These results indicate improved overall classification performance across diverse benchmarks.

For regression benchmarks, ArcMol shows strong performance on tasks dominated by physicochemical properties. On ESOL, which measures aqueous solubility, and Lipo, which evaluates lipophilicity, ArcMol achieves RMSE values of 0.517 and 0.578, respectively. Across both classification and regression benchmarks, ArcMol demonstrates competitive performance under a unified evaluation setting.

Generalization to Extended Tasks and Data-Limited Settings

To assess the generalization of ArcMol under conditions relevant to real-world drug discovery, we expanded the evaluation to 73 independent molecular property prediction tasks spanning six categories,⁴³ including general physicochemical properties, absorption, distribution, metabolism, excretion, and toxicity (Fig. 3a). Compared with commonly used academic benchmarks, these tasks cover a broader range of property types, data distributions, and biological contexts. In this setting, ArcMol matches or exceeds the performance of state-of-the-art pretrained models, including UniMol_V2_1.1B and MolCLR, on approximately 89 percent of the tasks. This trend is observed consistently across multiple ADMET categories, indicating

competitive performance across diverse molecular property prediction settings. Representative results are shown in Fig. 3a. ArcMol achieves an RMSE of 0.516 on the general physicochemical regression task logD, outperforming UniMol_V2_1.1B at 0.608, and attains an AUC of 0.949 on the blood brain barrier permeability benchmark bbb_logbb, exceeding MolCLR at 0.937. Across multiple endpoints related to metabolism, excretion, and toxicity, including CYP inhibition and carcinogenicity, ArcMol consistently matches or exceeds the performance of strong baseline methods.

Model performance was further evaluated under activity-cliff conditions using the MoleculeACE benchmark⁴⁴ (Fig. 3b), which comprises 30 challenging datasets. ArcMol shows strong performance across these benchmarks. For example, on CHEMBL4203_Ki, ArcMol achieves an RMSE of 0.676, substantially lower than that of Uni Mol V2 1.1B at 0.903. Activity-cliff datasets are characterized by high structural similarity coupled with large variations in bioactivity, posing a well known challenge for molecular property prediction. Under these conditions, ArcMol shows enhanced sensitivity to activity variations in datasets characterized by high structural similarity.

Model behavior was also examined under data-limited settings by varying the fraction of available training data (Fig. 3c). On both the classification task bbb_logbb and the regression task CHEMBL2147_Ki, predictive performance improves steadily as the training set size increases from 20 percent to 100 percent. Models trained with approximately 20 percent of the data already achieve performance close to that

obtained with the full dataset, with near saturation observed at around 60 percent. These results indicate that ArcMol can achieve competitive predictive performance with limited training data.

Structure of the latent representation space

To compare how different molecular representation models organize their latent spaces under supervised learning, we analyzed the hidden representations of ArcMol alongside several widely used baseline models, focusing on global distributions and local neighborhood structure (Fig. 4).

In the bbb_logbb classification task, the latent spaces learned by different models exhibit distinct distribution patterns (Fig. 4a). Models based on RDKit fingerprints exhibit relatively dispersed, point-like representations after supervised training, where samples from different classes occupy overlapping but still distinguishable regions. MolCLR enhances local similarity among samples of the same class after fine-tuning, although overlap remains near the class boundary. In contrast, ArcMol displays higher intra-class compactness in the t-SNE projection, and when the latent vectors are further projected onto a three-dimensional hypersphere, molecules from the same class concentrate within stable angular regions with clear separation between classes (Fig. 4c). This observation is further supported by quantitative clustering metrics, with higher Silhouette scores and lower Davies–Bouldin indices indicating a more compact and separable latent partitioning for ArcMol in this task (Fig. S2). Consistent with this organization, the geometric discriminant score derived from the ArcMol latent space shows a one-to-one correspondence with the predicted class probability (Fig. S3).

In the ChEMBL2147_Ki regression task, the organization of the latent space differs from that observed in classification settings (Fig. 4b). Models based on RDKit fingerprints produce relatively dispersed latent distributions, in which activity values show limited alignment with the representation space. MolCLR exhibits a well-structured latent distribution after fine-tuning; however, activity values are distributed non-monotonically across this structure, without a clear correspondence between representation organization and activity magnitude. Uni-Mol captures certain high-activity structural motifs, but highly active molecules are distributed across multiple separated regions, leading to discontinuities in activity-related organization. By contrast, ArcMol exhibits a more continuous organization of activity values within its latent space. In the t-SNE projection, molecules arrange approximately along the activity scale, while in the hyperspherical projection, latent vectors shift progressively in angular direction from low to high activity, with intermediate-activity molecules naturally occupying intermediate positions (Fig. 4d). In regression, experimental activity values show a clear linear correlation with cosine similarity to the highest-activity proxy anchor, indicating a stable angular ordering of activity in the latent space (Fig. S4). Consistently, pairwise cosine similarities among proxy anchors form an ordered structure along the activity scale (Fig. S5).

At the local scale, neighborhood-based retrieval experiments reveal differences in latent-space consistency across models (Fig. 4e,f). Using a k-nearest-neighbor (k-NN) retrieval scheme, where predictions are inferred from the k closest neighbors in the latent space, ArcMol exhibits relatively stable predictive performance across a

wide range of neighborhood sizes in both the bbb_logbb classification and CHEMBL2147_Ki regression tasks. By contrast, the performance of MolCLR and Uni-Mol varies more noticeably as k increases. In particular, the performance of ArcMol at $k = 1$ suggests that nearest neighbors in its latent space tend to be consistent in terms of molecular properties or activities. At the molecule level, we further examined how ArcMol represents property cliffs and continuous structural changes. For pairs of structurally similar molecules with divergent properties, ArcMol maps them to separated regions on the hypersphere, facilitating discrimination of property cliffs (Fig. 4g and Fig. S6). In regression settings, analysis of a series of R-group modifications indicates that ArcMol produces smooth and monotonic angular trajectories on the hypersphere that track the experimental activity trend and align with the proxy anchors (Fig. 4h and Fig. S7). Interpolation experiments show that spherical linear interpolation in the ArcMol latent space produces smoother predicted activity trajectories than linear interpolation in the original feature space (Fig. S8).

Drug repurposing for IRAK4 inhibition.

As an application of ArcMol in a realistic molecular discovery context, we examined its performance on a drug repurposing task targeting interleukin-1 receptor-associated kinase 4 (IRAK4), a serine/threonine kinase involved in Toll-like receptor and interleukin-1 receptor signaling.⁴⁵ Although IRAK4 has been extensively investigated as a therapeutic target, no selective IRAK4 inhibitors have yet been approved. A dataset comprising approximately 6,449 compounds with experimentally

reported IRAK4 activities was assembled from ChEMBL⁴⁶ and the literature,^{47,48} and model performance was evaluated using a scaffold-based data split.²⁸ On the held-out test set, ArcMol achieved higher predictive accuracy than baseline molecular representation models, as reflected by increased R^2 values and reduced RMSE (Fig. 5a,b). For the repurposing analysis, ArcMol was applied to rank FDA-approved compounds, with standard filtering applied to avoid training overlap and molecules falling outside the kinase-relevant molecular size range (Supplementary Table 1). This ranking recovered several FDA-approved drugs previously reported to interact with IRAK4, including sunitinib⁴⁹ and pacritinib,⁵⁰ as well as midostaurin, a clinical derivative of the staurosporine scaffold with reported IRAK4 activity (Fig. 5c).⁵¹ The top-ranked compounds identified in this analysis are summarized in Supplementary Table 2.

To further assess structural compatibility, molecular docking was performed for the top-ranked candidates, yielding docking scores with a mean of approximately -9.1 kcal/mol. A representative binding pose of pacritinib illustrates a predicted type I IRAK4 binding mode (Fig. 5d), consistent with previously reported IRAK4 inhibitors. In contrast, pexidartinib, an FDA-approved CSF1R inhibitor used for the treatment of symptomatic tenosynovial giant cell tumor,⁵² was predicted to adopt a type II binding conformation⁵³ within the IRAK4 pocket (Fig. 5e). Reported IRAK4 inhibitors are predominantly associated with type I binding modes, with only a limited number of type II inhibitors described, including ponatinib and HG-12-6,⁵⁴ highlighting the relative scarcity of this binding mode.

312

313 **Identification of type II IRAK4 inhibitors with novel scaffolds**

314 Building on the drug repurposing analysis, pexidartinib was identified as a
315 candidate compound and predicted by ArcMol to adopt a type II binding mode to
316 IRAK4. Biochemical evaluation confirmed that pexidartinib exhibits measurable
317 IRAK4 inhibitory activity, with an IC_{50} of approximately 365 nM, supporting the
318 relevance of the ArcMol-based repurposing prediction. In parallel, ponatinib,⁵⁴ a
319 clinically approved kinase inhibitor previously reported to bind IRAK4 in a type II –
320 like manner, showed weaker but detectable inhibition under the same assay conditions
321 ($IC_{50} \approx 1.6 \mu M$), providing an established reference for inactive-state binding.

322 Encouraged by the experimental validation of a repurposed type II inhibitor, we
323 next sought to identify additional IRAK4 inhibitors with novel chemical scaffolds
324 associated with this binding mode. To this end, we used ArcMol latent representations
325 to perform similarity-based retrieval from a bioactivity-focused small-molecule
326 library. Pexidartinib and ponatinib were used as anchor compounds in the ArcMol
327 latent space, and candidate molecules were retrieved and ranked based on their
328 representation similarity. The resulting candidates were further filtered through
329 molecular docking and binding-mode inspection against an inactive IRAK4
330 conformation (see Supporting Information, Virtual Screening Setup and Biochemical
331 Assays; Fig. 6a and Fig. S9). Based on molecular docking performance and binding
332 mode consistency, 15 compounds were identified as candidates for experimental
333 evaluation, and their chemical structures are shown in Fig. S10. From this set, six

representative compounds were selected for biochemical testing, together with the reference compound ponatinib and the repurposed compound pexidartinib identified in a preceding analysis, whose chemical structures are shown in Fig. 6b. In biochemical assays, five of the six tested candidates exhibited inhibitory activity against IRAK4, with IC₅₀ values ranging from the low nanomolar to low micromolar regime (Fig. 6c,d). Among these, HY-128571 displayed the strongest inhibition (IC₅₀ $\approx$ 89 nM), followed by HY-12846 ($\approx$ 165 nM) and I3 ($\approx$ 444 nM). HY-12291 and HY-12462 showed weaker but detectable activity ($\approx$ 648 nM and $\approx$ 1.3 μ M, respectively), whereas HY-12038 showed no detectable inhibition at concentrations up to 10 μ M. Representative docking poses of HY-12846 and HY-128571 illustrate extended binding conformations engaging an auxiliary hydrophobic pocket adjacent to the ATP-binding site, consistent with inactive-state IRAK4 binding (Fig. 6e,f). Collectively, these results indicate that ArcMol-based latent-space similarity retrieval, combined with structure-guided filtering, can support the identification of scaffold-diverse IRAK4 inhibitors beyond direct analogues of known compounds, demonstrating its practical applicability in drug discovery settings.

Ablation studies on ArcMol

To assess the contribution of individual design components in ArcMol, we conducted a systematic ablation study in which key modules were removed or modified (Supplementary Table 3). Specifically, we evaluated model variants without the task-adaptive spherical readout, without multi-view descriptor fusion, without

explicit physicochemical descriptors, and without the task-adaptive mixture-of-experts mechanism. In addition, a dense mixture-of-experts variant was examined, in which task-specific expert routing was replaced by uniform expert aggregation. Across both classification and regression benchmarks, removal or modification of any component resulted in a consistent decrease in predictive performance, indicating that each module contributes to the overall effectiveness of the framework. Among the ablated variants, removing physicochemical descriptors led to the largest performance degradation, particularly for regression tasks. Variants without the task-adaptive spherical readout or the task-adaptive mixture-of-experts mechanism also exhibited substantial performance loss across multiple tasks. Furthermore, replacing sparse, task-adaptive expert routing with dense expert aggregation resulted in additional performance degradation.

Overall, ArcMol consistently outperformed all ablated variants, indicating that its performance arises from the combined contributions of representation fusion, task-adaptive supervision, physicochemical descriptors, and selective expert routing.

Discussion

In this study, we developed ArcMol, a task-adaptive molecular representation framework designed to integrate heterogeneous molecular views within a unified and structured latent space. By combining multi-view molecular representations and task-aware selection and fusion of molecular information, ArcMol provides a flexible and principled approach for aligning diverse chemical priors with heterogeneous downstream objectives. Crucially, this design allows molecular representations to be reshaped according to the prediction task, effectively shifting the representation space from one defined primarily by molecular structure to one organized around structure–property or structure–activity relationships. This task-dependent organization enables different prediction tasks to be addressed within a shared representational basis, while preserving task-specific distinctions. The resulting representations exhibit clear class separation, ordered property gradients, and locally coherent neighborhoods, supporting robust behavior across classification, regression, and data-limited scenarios.

To assess the practical relevance of ArcMol beyond methodological validation, we examined two complementary case studies centered on the therapeutically important kinase IRAK4. In a drug repurposing setting, ArcMol prioritized FDA-approved compounds with biologically and structurally plausible binding modes. Building on this analysis, ArcMol was further applied to the prioritization of type II–biased IRAK4 inhibitor candidates, where experimental evaluation of a subset of model-selected compounds led to the identification of multiple nanomolar type II

394 inhibitors. Together, these case studies indicate that task-aligned molecular
395 representations learned by ArcMol support consistent prioritization across
396 computational analysis and experimental assessment. More broadly, the ability to
397 learn task-aligned molecular representations that capture meaningful similarity
398 relationships may facilitate how chemical space is explored and interpreted in
399 practical discovery settings. In this context, ArcMol provides a foundation for
400 molecular property prediction and representation learning, and may inform future
401 extensions toward property-guided molecular optimization.

402

403

404

Methods

Multi-View Molecular Representation and Pooling

To integrate heterogeneous molecular information spanning structural topology and three-dimensional geometry, we developed a representation-agnostic pooling framework that projects diverse modalities into a unified latent space. ArcMol integrates 14 representative views across four primary categories: knowledge-based fingerprints⁵⁵ (including RDKit, MACCS, and EState), molecular language models (MolT5,²⁵ BioT5²⁶), 2D graph self-supervised embeddings (GraphMVP,²⁷ GROVER,³² etc.), and 3D geometry-aware representations (UniMol variants³³). For the explicit RDKit descriptors, we select the top k features based on their mutual information⁵⁶ (MI) with the target labels to optimize the information-to-noise ratio. To ensure numerical stability during the training of these heterogeneous features, a heuristic preprocessing pipeline is applied: missing or invalid values are imputed as 0.0, and all feature values are strictly clipped to the range $[-10^6, 10^6]$ to prevent gradient explosion in the initial layers. Recognizing that direct concatenation of these high-dimensional sources would lead to excessive feature dimensionality and noise accumulation, we first map each input vector x_m to a unified D -dimensional latent space through independent linear projection layers:

$$h_m = \text{ReLU}(W_m x_m + b_m)$$

In this formulation, $x_m \in \mathbb{R}^{d_m}$ denotes the raw input feature vector of the m -th modality with input dimension d_m ; $W_m \in \mathbb{R}^{D \times d_m}$ and $b_m \in \mathbb{R}^D$ represent the learnable weight matrix and bias vector specific to that modality, respectively; and

$h_m \in \mathbb{R}^D$ (where $D=128$) is the resulting aligned latent representation. This alignment stage projects diverse semantic and geometric priors into a consistent and computationally compatible manifold, providing a standardized latent foundation for the subsequent task-adaptive mixture-of-experts (MoE) fusion.

Task-Adaptive Mixture-of-Experts (MoE) Fusion

The Task-Adaptive Mixture-of-Experts (MoE) Fusion module dynamically aggregates the M aligned representations by evaluating their semantic density and cross-modal interdependencies. For each projected representation $h_m \in \mathbb{R}^D$, a content-based attention score a_m^c is evaluated through a non-linear transformation:

$$a_m^c = v^T \text{ReLU}(W_c h_m + b_c)$$

where $W_c \in \mathbb{R}^{H \times D}$ and $b_c \in \mathbb{R}^H$ denote the weight matrix and bias vector of the attention layer, respectively, and $v \in \mathbb{R}^H$ is a learnable projection vector that maps the hidden state to a scalar. To penalize information redundancy and prioritize unique chemical priors, we incorporate a complementarity bias a_m^p , defined as the average dissimilarity of representation m relative to all other branches $n \neq m$:

$$a_m^p = \frac{1}{M-1} \sum_{n \neq m} (1 - \text{CKA}_{\text{RBF}}(h_m, h_n; \gamma))$$

This term utilizes Radial Basis Function-based Centered Kernel Alignment³⁹ (CKA_{RBF}) with a kernel bandwidth γ to capture high-order structural correlations. To align fusion logic with specific downstream objectives, a task-aware gating⁵⁷ term $g_m = \Phi(\mathbf{t})$ is derived from a task context vector $\mathbf{t} \in \mathbb{R}^{d_t}$, where $\Phi(\cdot)$ denotes a task-specific projection layer. These facets are aggregated into the total score $S_m =$

$a_m^c + \beta \cdot a_m^p + g_m$, where β is a learnable scaling coefficient regulating the influence of the complementarity bias.

This total score S_m drives an expert routing mechanism⁵⁸ that employs sparse Top-K gating to activate the K most relevant representation experts:

$$w_m = \frac{\mathbb{I}(S_m \in \text{Top-K}) \cdot \exp(S_m)}{\sum_{j \in \text{Top-K}} \exp(S_j)}$$

where $\mathbb{I}(\cdot)$ is the indicator function that restricts the softmax operation to the selected subset of experts. The model is regularized by a sparsity entropy loss⁵⁸ $\mathcal{L}_{\text{sparse}} = -\lambda_s \sum w_m \log w_m$ to concentrate attention on task-relevant experts and enhance interpretability. The final fused representation $\mathbf{z}_{\text{fused}}$ is synthesized as:

$$\mathbf{z}_{\text{fused}} = \text{Dropout}\left(\sum_{m=1}^M w_m \mathbf{h}_m\right)$$

This task-optimized distillation captures the holistic latent consensus of the molecular structure.

Task-aware descriptor pooling module

To extract task-specific chemical information, we implement a Task-Aware Descriptor Attention module that adaptively filters and compresses high-dimensional molecular descriptors. A salient feature vector $\mathbf{d}_{\text{sel}} \in \mathbb{R}^k$ is constructed by selecting the top k descriptors based on their mutual information with target labels, followed by a non-linear projection into a hidden space:

$$\mathbf{h}_d = \text{LayerNorm}(\text{ReLU}(W_f \mathbf{d}_{\text{sel}} + b_f))$$

Where W_f and b_f are learnable parameters. To prioritize features critical to the current task, an attention score⁵⁹ α is computed using a learnable query vector $\mathbf{q}$

that captures the task context.

$$\alpha = \sigma\left(\frac{\sum (h_d \cdot q)}{\sqrt{H}}\right)$$

This attention mechanism allows the model to selectively emphasize descriptors critical to the current target context, such as lipophilicity for membrane permeability or electronic properties for metabolic stability. The resulting descriptor embedding, $z_{desc} = \text{ReLU}(W_o(\alpha \odot W_v h_d) + b_o)$, is integrated with structural fusion embeddings to form a comprehensive molecular representation $X = [z_{fused} \parallel z_{desc}]$. By harmonizing task-aware chemical knowledge with global topologies, this synergized latent space yields a highly discriminative and target-aligned representation for robust property prediction.

Spherical Manifold Projection

To enhance latent discriminability and ensure numerical stability during the optimization of heterogeneous modalities, the joint molecular representation X is projected onto a unit hypersphere where similarity is governed by cosine distance.⁶⁰⁻⁶² This mapping restricts the embedding to a spherical manifold where molecular proximity is determined by directional semantics rather than feature magnitude.⁶¹ The projection is defined as follows:

$$z = \frac{f(W_a X + b_a)}{\|f(W_a X + b_a)\|_2}$$

where X is the joint representation synthesized from the multi-view fusion and descriptor modules, while $W_a X$ and b_a denote the learnable weight matrix and bias vector of the projection layer, respectively. The term $f(\cdot)$ represents the Parametric

Rectified Linear Unit (PReLU) activation function⁶³, which provides the necessary non-linearity while maintaining healthy gradient flow across the manifold. This spherical constraint restricts the latent vectors to a unit hypersphere, ensuring that the latent space is governed by cosine similarity rather than Euclidean distance.

Dual-Head Learning with Spherical Proxies and Distillation

ArcMol employs a synchronous Teacher-Student architecture to balance discriminability and generalization.^{64,65} This distillation framework extracts robust task-specific features while mitigating overfitting risks inherent in multi-modal heterogeneity. In classification tasks, the architecture integrates a proxy-based⁶² student head f_s and a multilayer perceptron (MLP) teacher head f_t . The teacher head generates class logits l_t directly from the joint representation X via non-linear mapping:

$$l_t = f_t(X) = W_2 \cdot \text{ReLU}(W_1 X + b_1) + b_2$$

Where W_1 , W_2 and b_1 , b_2 represent the learnable weight matrices and bias vectors of the teacher’s MLP layers, respectively. Simultaneously, the student head introduces a proxy learning mechanism^{61,62} by calculating the cosine similarity between the projected spherical embedding z and a set of trainable proxies $P = [p_1, ..., p_C] \in \mathbb{R}^{C \times D}$, where C is the number of classes and D is the embedding dimension. To maximize inter-class separation, an additive angular margin m is applied to the ground-truth class y during training:

$$l_{s,y} = \kappa \cdot (\cos \theta_y - m), \quad l_{s,j} = \kappa \cdot z \cdot \cos \theta_j \quad (j \neq y)$$

where κ represents a scaling factor⁶⁰ that re-scales the logits for optimal cross-entropy optimization and $\cos \theta_j = z \cdot p_j$ denotes the cosine similarity between the embedding and the j-th proxy. To prevent centroid collapse and maintain diversity within the latent space, the ArcProxyOrthogonality module enforces an orthogonality constraint⁶⁶ on the proxy matrix:

$$\mathcal{L}_{ortho} = \lambda_o \| PP^T - I \|_F^2$$

where I is the identity matrix, $\|\cdot\|_F$ denotes the Frobenius norm, and λ_o is a regularization weight that penalizes redundancy among class proxies. The comprehensive classification loss $\mathcal{L}_{cls}$ synergizes cross-entropy supervision with label-smoothing⁶⁷ and a Kullback-Leibler (KL) divergence distillation⁶⁸ term:

$$\mathcal{L}_{kd} = KL(\sigma(l_s/\tau) \parallel \sigma(l_t/\tau)) \cdot \tau^2$$

Here, $\sigma(\cdot)$ is the softmax function, and τ is a temperature hyperparameter that softens the probability distributions to facilitate knowledge transfer from the teacher to the student. The total classification objective is defined as

$$\mathcal{L}_{cls} = 0.5(\mathcal{L}_{CE}(l_t, y) + \mathcal{L}_{CE}(l_s, y)) + \lambda_{kd} \cdot \mathcal{L}_{kd} + \lambda_o \mathcal{L}_{ortho}$$

where λ_{kd} regulates the distillation intensity.

For regression tasks, both student head g_s and teacher head g_t utilize MLP structures to project the joint representation X into continuous scalar predictions $y_s = g_s(X)$ and $y_t = g_t(X)$. To supplement these values with ordinal manifold constraints, the ArcProxyAlignment module discretizes the target property range into K anchors⁶⁹ $[b_1, \dots, b_K]$, where each anchor b_k represents a distinct property level. Each ground-truth label y is mapped to a soft-target distribution⁷⁰ y_k via a Gaussian

524 kernel:

$$\hat{y}_k = \frac{\exp(-(y - b_k)^2 / (2\sigma_{soft}^2))}{\sum_{j=1}^K \exp(-(y - b_j)^2 / (2\sigma_{soft}^2))}$$

525 where σ_{soft} is a bandwidth parameter that controls the smoothness of the
526 probability assignment over property bins. The model then optimizes an Arc
527 regularization loss $\mathcal{L}_{arc}$, which enforces an angular margin^{41,62} m_{arc} on the anchor
528 proxy v_t nearest to the true value:

$$\mathcal{L}_{arc} = - \sum_{k=1}^K \hat{y}_k \log \frac{\exp(\kappa_{arc} \cdot (\cos \phi_k - \mathbb{I}_{[k=t]} \cdot m_{arc}))}{\sum_{j=1}^K \exp(\kappa_{arc} \cdot (\cos \phi_j - \mathbb{I}_{[j=t]} \cdot m_{arc}))}$$

529 In this loss function, $\cos \phi_k = z \cdot v_k$ is the similarity between the embedding
530 and the k-th bin proxy, κ_{arc} is the scaling factor, while $\mathbb{I}_{[k=t]}$ is an indicator function
531 that applies the margin m_{arc} only to the target bin t to enhance local compactness.
532 The final regression objective integrates Mean Squared Error (MSE) supervision,
533 consistency distillation (λ_{kd}), and the sparsity entropy regularization $\mathcal{L}_{sparse}$ derived
534 from the MoE module:

$$\mathcal{L}_{reg} = 0.5(MSE(\hat{y}_t, y) + MSE(\hat{y}_s, y)) + \lambda_{kd} \cdot MSE(\hat{y}_s, \hat{y}_t) + \lambda_{arc} \mathcal{L}_{arc} + \mathcal{L}_{sparse}$$

535 where λ_{arc} is the weighting coefficient for the geometric manifold constraint.
536 This collaborative optimization ensures that ArcMol captures global chemical trends
537 while maintaining a topologically consistent latent space.

538

539 Training Strategy and Implementation Details

540 ArcMol is implemented in PyTorch and trained end-to-end with supervised

learning. For each dataset, all modality-specific molecular representations are precomputed and treated as fixed inputs, while the modality projection layers, task-adaptive mixture-of-experts⁷¹ fusion module, spherical projection, proxy parameters, and teacher–student prediction heads are jointly optimized. Classification and regression tasks share the same backbone architecture, with task-specific losses and angular proxy-based supervision applied accordingly. Models are optimized using AdamW⁷² with decoupled weight decay and trained for up to 100 epochs with early stopping based on validation performance (patience of 20 epochs). Mini-batch sizes are selected from {48, 64, 96} depending on feature dimensionality, and training is stabilized using gradient clipping (maximum norm 1.0) and dropout (rate 0.1–0.3). Hyperparameters are selected via automated search using Optuna⁷³ on the validation set (as shown in Supplementary Table 4), with separate searches for classification and regression settings.

Code availability

The source code of ArcMol is publicly available at <https://github.com/zyrlia1018/ArcMol>. All codes and processed data associated with this study are openly accessible via Zenodo at <https://zenodo.org/records/18373814>

Data availability

All data supporting the key findings of this study are publicly available via Zenodo at <https://zenodo.org/records/18373814>. The deposited materials include

processed datasets derived from multiple public benchmarks, including the ADMET benchmark (originally from <https://biosig.lab.uq.edu.au/deeppk/data>), the MoleculeNet benchmark (originally from <https://github.com/deepchem/deepchem>), the MoleculeACE benchmark (originally from <https://github.com/molML/MoleculeACE>), and the IRAK4 dataset used in this study. The reference protein structures are available in the Protein Data Bank under accession codes 6EGD and 6EG9. Source data are provided with this paper.

Acknowledgements

This work was supported by National Key R&D Program of China (grant no. 2023YFF1204905, S.Y.). We also thank Chengdu Zenitar Biotechnology Co., Ltd. for their excellent support with computing resources.

Author contributions

Taijin Wang and Lijuan Chen supervised the study. Yurong Zou conceived and designed the overall network framework and code, and performed the experimental analysis and writing. Other contributors provided valuable suggestions and discussions.

Competing interests

The authors declare no competing interests.

Reproducibility statement

All reported results are averaged over five independent runs with different random seeds. Publicly available datasets were used throughout this study, and sufficient details on data processing, model architecture, training procedures, and

589 hyperparameter search spaces are provided in the Methods and Supplementary
590 Information to enable reproduction of the experiments. All experiments were
591 performed on NVIDIA A100 GPUs (80 GB).

592

593

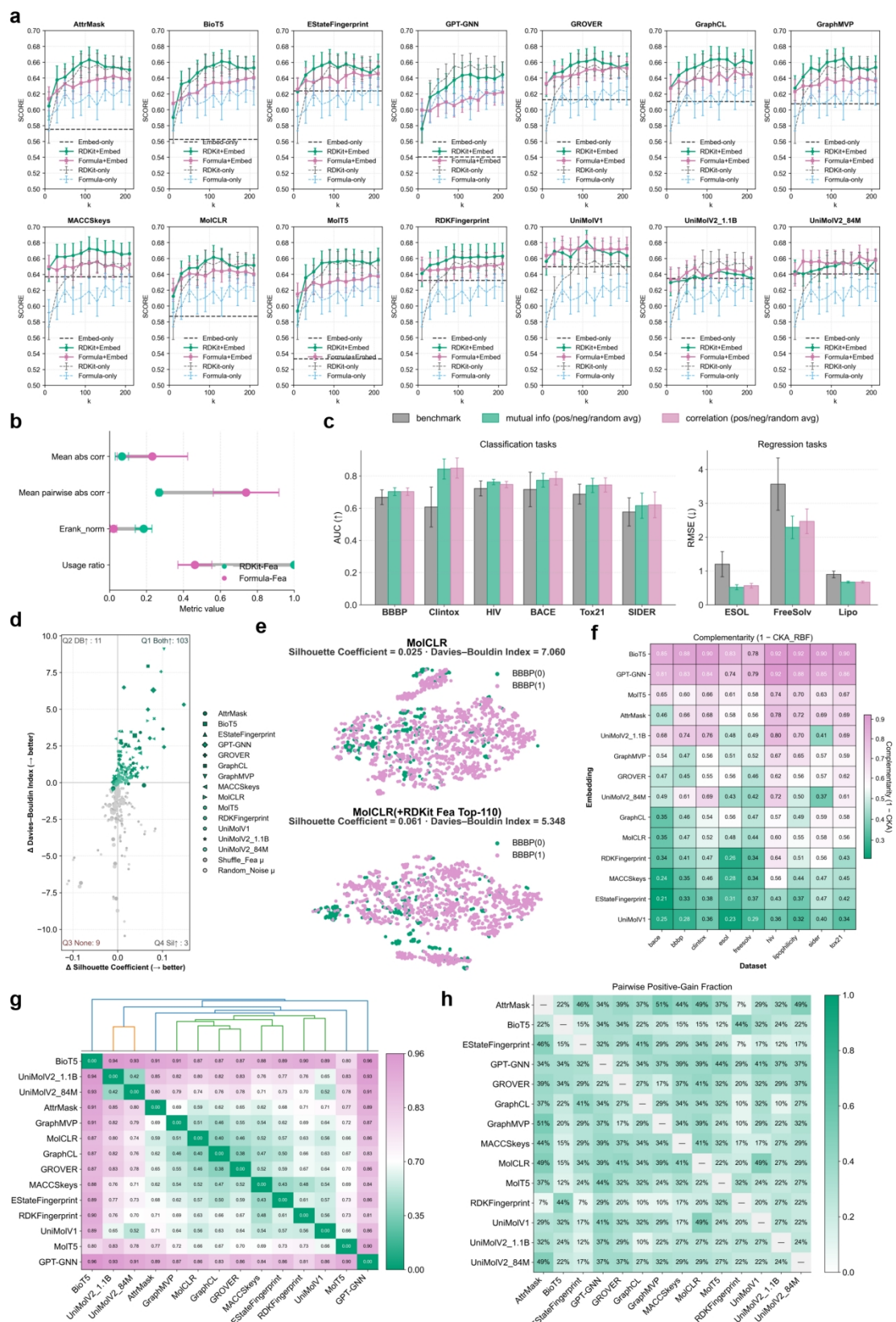


Figure 1. Unveiling representational complementarity and task-dependent organization in molecular representation spaces. a, Ablation study on feature dimension, contrasting saturation dynamics between symbolic regression features and RDKit descriptors. **b**, Performance gain comparison between mutual information and correlation-based

feature screening strategies. **c**, Permutation invariance test evaluating prediction stability across different feature input sequences. **d**, Volcano plot characterizing latent manifold reshaping, where axes denote improvements in cluster compactness (silhouette coefficient) and separation (Davies-Bouldin index). **e**, Visualization of latent space clustering improvements for MolCLR before and after semantic injection. **f**, Complementarity heatmap between 14 molecular embeddings and RDKit descriptors across 47 tasks. **g**, Pairwise semantic orthogonality matrix via centered kernel alignment, revealing divergence between representation paradigms. **h**, Distribution of synergistic performance gains achieved by fusing orthogonal representation pairs.

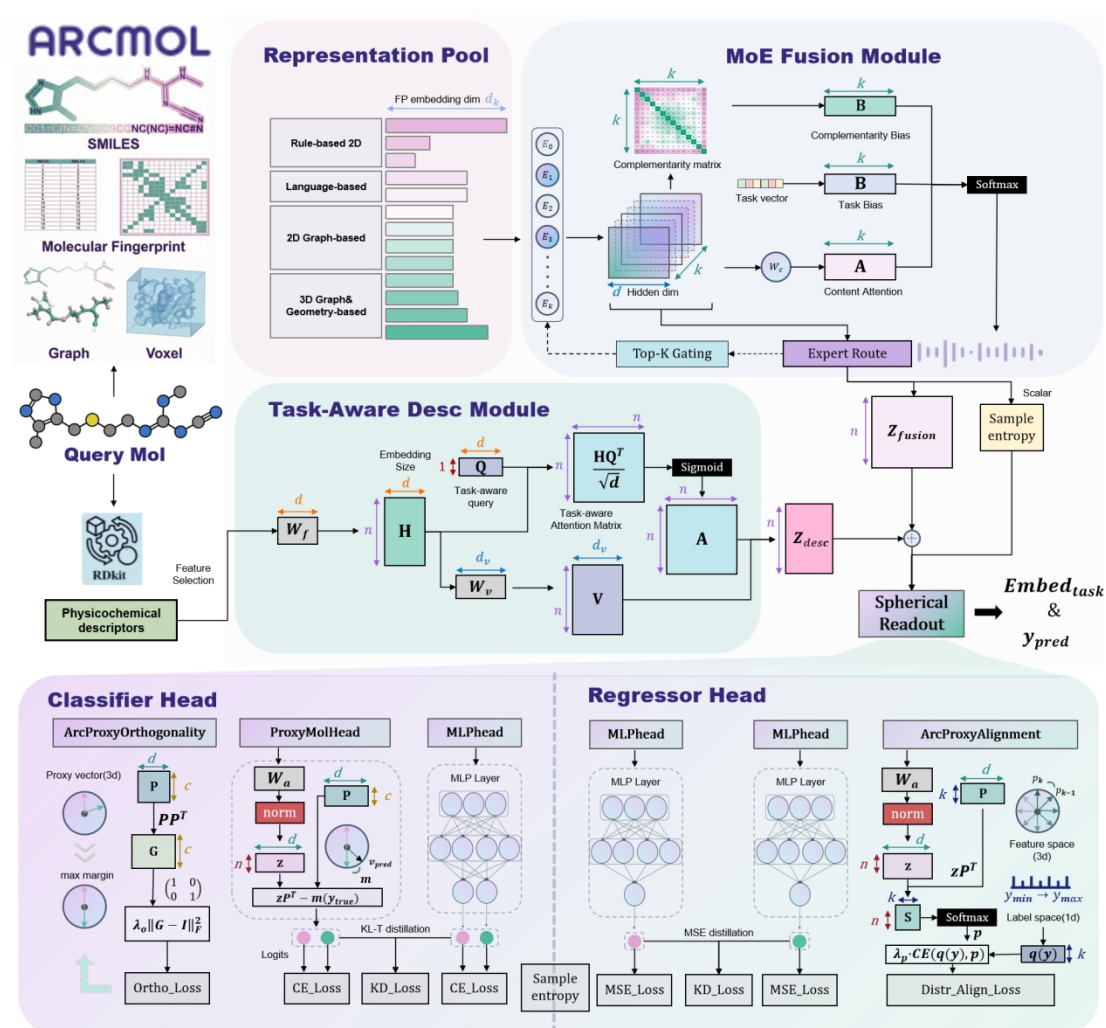


Figure 2. Architecture of the ArcMol framework. ArcMol integrates structural and task-aware descriptor views through task-adaptive MoE fusion. A dual-head distillation readout then projects these features onto a spherical manifold to synchronize task spaces for optimized representation and molecular property prediction.

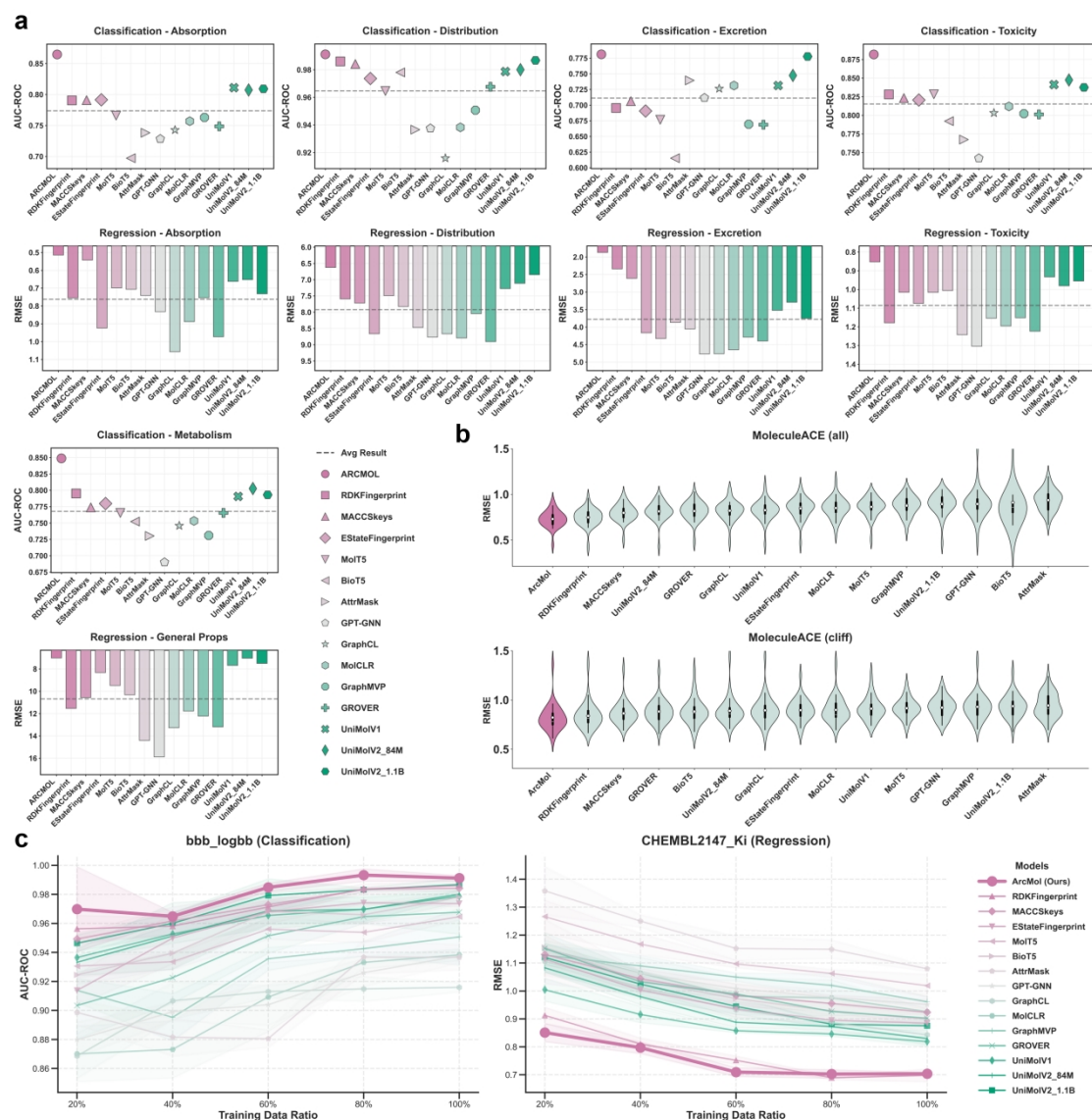


Figure 3. ArcMol maintains predictive stability across ADMET tasks, activity cliffs, and low-data regimes. a, Predictive performance of ArcMol across 73 ADMET-related molecular property prediction tasks, spanning general physicochemical properties, absorption, distribution, metabolism, excretion, and toxicity (G/A/D/M/E/T). Scatter - bar plots report task-wise evaluation metrics. **b,** Violin plots showing regression error distributions on the MoleculeACE benchmark, including all tasks and activity-cliff subsets. **c,** Few-shot learning curves for the classification task bbb_logbb and the regression task CHEMBL2147_Ki, evaluated under increasing fractions of training data (20% - 100%).

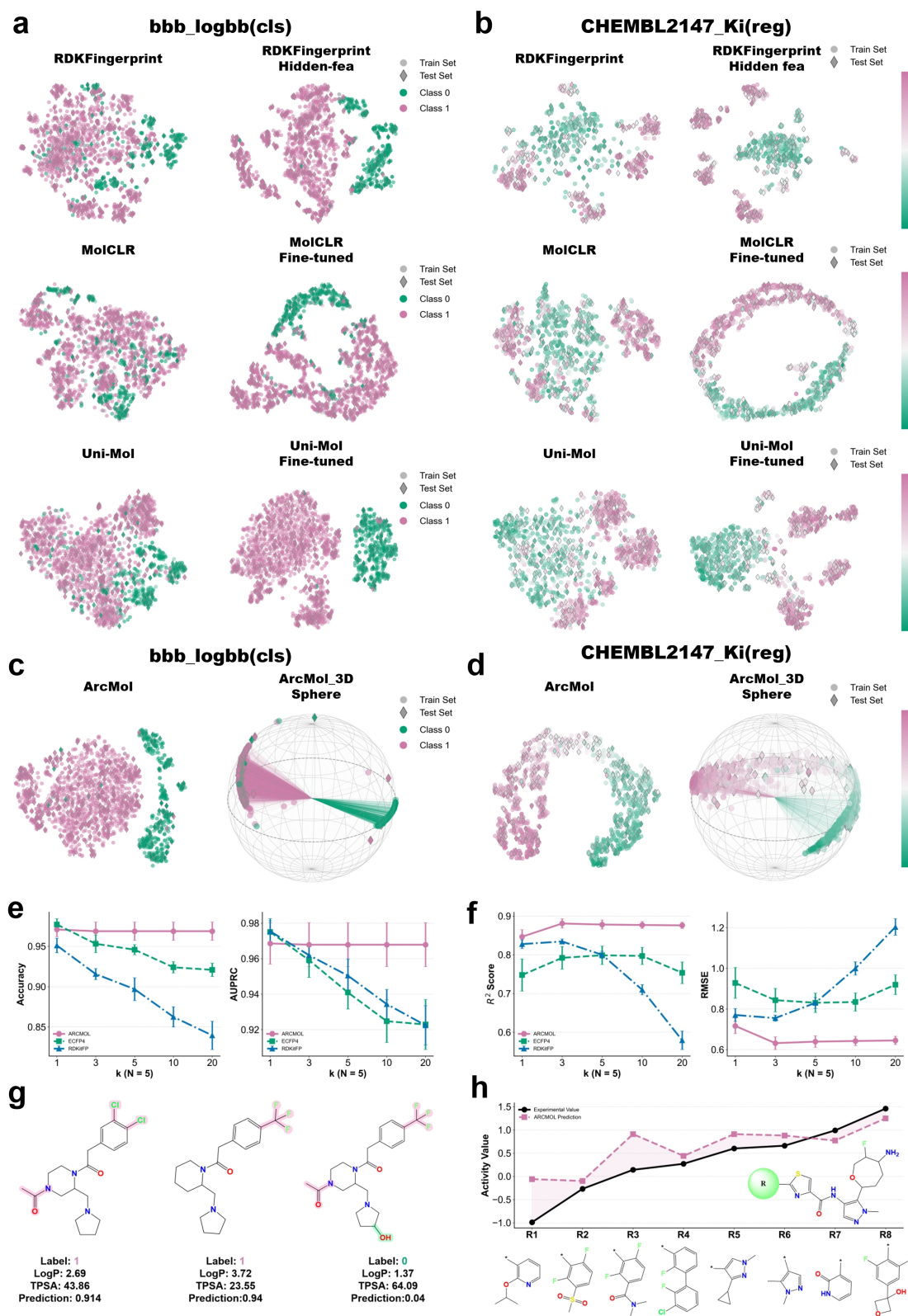


Figure 4. Geometry of the latent representation space. **a**, t-SNE visualizations of latent representations for the classification task *bbb_logbb* using different molecular representations. Results are shown for RDGFingerprint, MolCLR, and Uni-Mol before training (left) and after fine-tuning (right). For RDGFingerprint, hidden features from the downstream MLP are used. Pink denotes class 1 and teal denotes class 0; training and test samples are shown together. **b**,

t-SNE visualizations of latent representations for the regression task CHEMBL2147_Ki before training (left) and after fine-tuning (right). Colors range from teal to pink, indicating activity from low to high. **c**, Latent space of ArcMol for the bbb_logbb classification task. Left: t-SNE visualization of the learned latent representations. Right: example visualization of the same representations projected to three dimensions by PCA and mapped onto a unit sphere. **d**, Same as c, but for the CHEMBL2147_Ki regression task; color gradients indicate activity values. **e**, Classification performance based on k-nearest-neighbor (k-NN) retrieval, showing accuracy and AUPRC for ArcMol, MolCLR, and Uni-Mol across different values of k. **f**, Regression performance based on k-NN retrieval, showing R^2 and RMSE as a function of the number of neighbors k. **g**, Representative examples from the bbb_logbb classification task illustrating ArcMol predictions for structurally similar molecules. **h**, Representative examples from the CHEMBL2147_Ki regression task, comparing predicted and experimental activity trends across a series of R-group modifications.

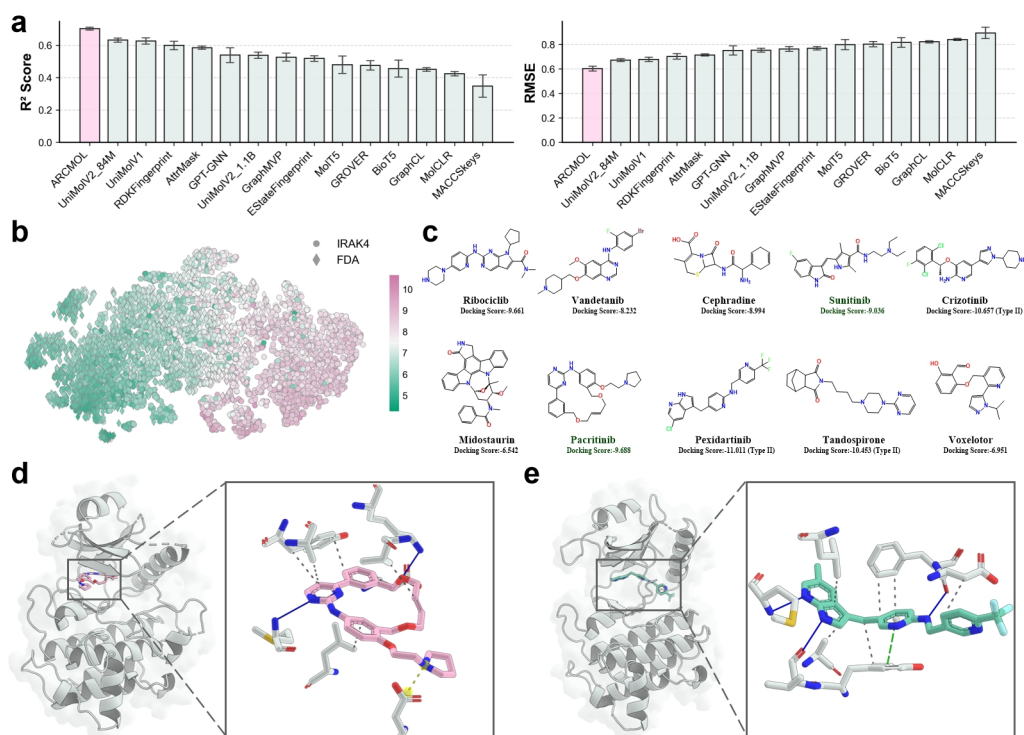


Figure 5. ArcMol enabled drug repurposing identifies candidate IRAK4 inhibitors. **a**, Predictive performance of ArcMol and baseline models on the IRAK4 activity prediction task, evaluated on a scaffold-split test set using R² (left) and RMSE (right). **b**, Two-dimensional projection of the learned latent representations for the IRAK4 dataset, with FDA-approved drugs highlighted. **c**, Top-ranked FDA-approved drugs predicted by ArcMol as potential IRAK4 inhibitors. Compounds previously reported to interact with IRAK4 are highlighted in dark green. **d**, Representative binding pose of a predicted type I IRAK4 inhibitor, illustrating interactions within the active kinase conformation. **e**, Representative binding pose of a predicted type II IRAK4 inhibitor, showing binding to the inactive conformation of IRAK4. Molecular docking was performed using both the active crystal structure of IRAK4 (PDB ID: 6EGD⁵⁴) and the inactive DFG-out conformation of IRAK4 (PDB ID: 6EG9⁵⁴).

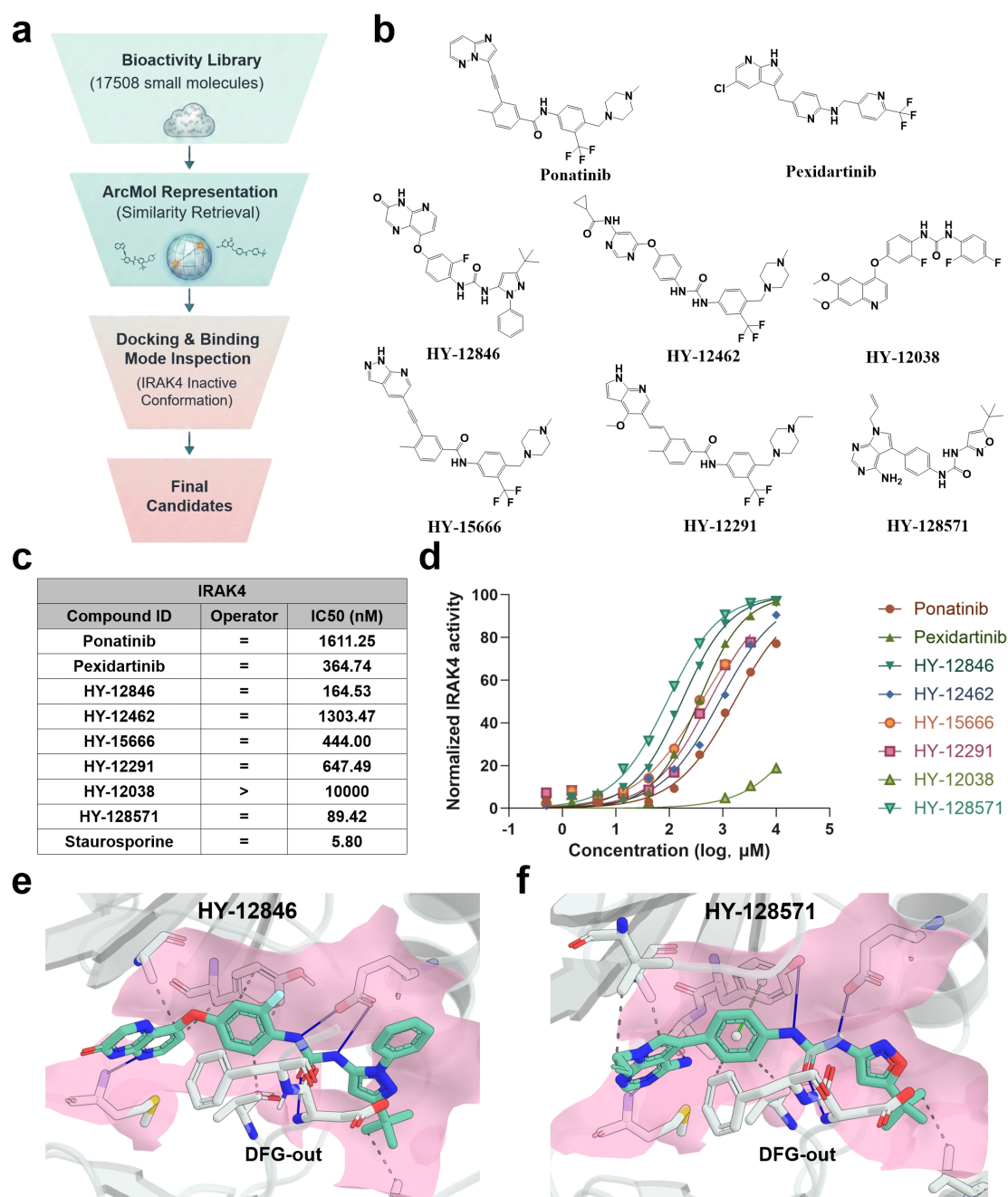


Figure 6. **a**, Screening protocol for IRAK4. Screening was performed on a bioactivity library comprising 17,508 drug-like small molecules. Molecular docking was conducted using the inactive (DFG-out) crystal structure of IRAK4 (PDB ID: 6EG9⁵⁴), followed by ArcMol-based prioritization and structure-guided filtering. **b**, Chemical structures of ponatinib, the repurposed FDA-approved drug pexidartinib, and the six commercially available candidate compounds selected for experimental evaluation. **c**, Summary of half-maximal inhibitory concentration (IC₅₀) values of the indicated compounds, determined using IRAK4 enzyme kinetics assays. **d**, Dose-response curves of seven active compounds measured in IRAK4 enzyme inhibition assays. **e,f**, Representative docking poses of HY-12846 and HY-128571 bound to the inactive (DFG-out) conformation of IRAK4. The protein surface is shown in pink. Both compounds adopt extended binding poses, with terminal substituents occupying an auxiliary hydrophobic pocket adjacent to the ATP-binding site, a feature characteristic of type II kinase inhibitors.

689 **Table 1 Benchmark performance comparison on molecular property prediction tasks.**

type	Classification (AUC↑)						Regression (RMSE↓)		
Methods	BBBP	Clintox	HIV	BACE	Tox21	SIDER	ESOL	FreeSolv	Lipo
#Molecules	2039	1478	41127	1513	7831	1427	1128	642	4200
#task	1	2	1	1	12	27	1	1	1
D-MPNN ¹²	71.0(0.3)	90.6(0.6)	77.1(0.5)	80.9(0.6)	75.9(0.7)	57.0(0.7)	1.050(0.008)	2.082(0.082)	0.683(0.016)
AttentiveFP ⁷⁴	64.3(1.8)	84.7(0.3)	75.7(1.4)	78.4(2.2)	76.1(0.5)	60.6(3.2)	0.877(0.029)	2.073(0.183)	0.721(0.001)
N-GramRF ⁷⁵	69.7(0.6)	77.5(4.0)	77.2(0.1)	77.9(1.5)	74.3(0.4)	66.8(0.7)	1.074(0.107)	2.688(0.085)	0.812(0.028)
N-GramXGB ⁷⁵	69.1(0.8)	87.5(2.7)	78.7(0.4)	79.1(1.3)	75.8(0.6)	65.5(0.7)	1.083(0.082)	5.061(0.744)	2.072(0.030)
PretrainGNN ⁷⁶	68.7(1.3)	72.6(1.5)	79.9(0.7)	84.5(0.7)	78.1(0.6)	62.7(0.8)	1.100(0.006)	2.764(0.002)	0.739(0.003)
GEM ⁷⁷	72.4(0.4)	90.1(1.3)	80.6(0.9)	85.6(1.1)	78.1(0.1)	67.2(0.4)	0.798(0.029)	1.877(0.094)	0.660(0.008)
ImageMol ⁷⁸	73.9(0.2)	85.1(1.4)	79.7(0.2)	83.9(0.5)	77.3(0.1)	66.0(0.1)	0.970(0.070)	2.020(0.070)	0.720(0.010)
Uni-Mol ³³	72.9(0.6)	91.9(1.8)	80.8(0.3)	85.7(0.2)	79.6(0.5)	65.9(1.3)	0.788(0.029)	1.480(0.048)	0.603(0.010)
MolMCL _{GIN} ⁴¹	74.1(0.6)	95.7(1.2)	79.8(0.3)	85.0(1.1)	77.5(0.3)	66.7(0.8)			
MolMCL _{GPS} ⁴¹	73.6(0.7)	95.1(0.5)	80.2(0.5)	86.1(1.3)	79.0(0.6)	68.7(0.2)			
MolT5 ²⁵	64.9(0.4)	51.3(0.8)	73.6(0.1)	72.2(0.7)	62.7(7.1)	49.6(7.4)	1.965(0.025)	4.561(0.048)	1.093(0.006)
BioT5 ²⁶	62.3(0.5)	62.0(1.0)	69.9(0.7)	75.6(0.2)	67.4(5.5)	51.5(8.1)	1.204(0.031)	3.608(0.053)	0.960(0.009)
GPT-GNN ³¹	55.1(0.3)	45.6(1.1)	69.5(0.4)	58.4(14.2)	60.3(7.6)	52.8(7.7)	1.594(0.012)	4.386(0.014)	1.060(0.014)
AttrMask ³⁰	65.2(1.4)	73.5(4.3)	75.3(1.5)	77.8(1.8)	72.2(0.9)	63.3(0.9)	1.112(0.048)	2.922(0.033)	0.730(0.004)
GraphCL ²⁹	67.8(2.4)	77.5(3.4)	75.1(0.7)	74.6(2.1)	74.3(0.7)	60.6(1.3)	1.313(0.071)	3.598(0.315)	0.776(0.022)
GraphMVP ²⁷	70.8(0.5)	79.1(2.8)	76.0(0.1)	79.3(1.5)	71.9(0.8)	62.4(1.1)	1.064(0.019)	2.951(0.190)	0.691(0.013)
MolCLR ²⁸	73.1(1.6)	91.6(2.7)	77.3(1.3)	81.5(1.6)	73.4(0.6)	59.4(0.7)	1.305(0.025)	2.727(0.077)	0.735(0.018)
GROVER ³²	65.1(2.5)	74.1(11.8)	57.7(4.3)	81.1(2.3)	66.3(4.3)	56.8(5.1)	1.128(0.080)	2.377(0.390)	0.905(0.086)
LLM4SD ⁴²	75.6(0.2)	91.3(3.4)	79.0(0.2)	83.8(0.3)	74.7(0.8)	65.4(1.4)	0.520(0.040)	2.620(0.006)	0.680(0.002)
ArcMol	77.8(1.2)	94.3(2.6)	79.1(1.8)	87.9(2.5)	82.2(0.3)	73.0(0.6)	0.517(0.026)	1.872(0.023)	0.578(0.031)

690

691

692

693

References

1. Shen, C., *et al.* DrugFlow: An AI-Driven One-Stop Platform for Innovative Drug Discovery. *Journal of Chemical Information and Modeling* **64**, 5381-5391 (2024).
2. He, G., *et al.* Prototype-based contrastive substructure identification for molecular property prediction. *Briefings in Bioinformatics* **25**, bbae565 (2024).
3. Zhang, S., *et al.* BioM2: biologically informed multi-stage machine learning for phenotype prediction using omics data. *Briefings in Bioinformatics* **25**, bbae384 (2024).
4. Yi, J., *et al.* Pushing the boundaries of few-shot learning for low-data drug discovery with a Bayesian meta-learning hypernetwork framework. *Briefings in Bioinformatics* **26**, bbaf408 (2025).
5. Han, X., Zhang, Z., Bai, C. & Wu, Z. Integrated multimodal hierarchical fusion and meta-learning for enhanced molecular property prediction. *Briefings in Bioinformatics* **26**, bbaf251 (2025).
6. Li, H., *et al.* A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications* **14**, 7568 (2023).
7. Rives, A., *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).
8. Jumper, J., *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
9. Xu, Y., *et al.* Pretrained E(3)-equivariant message-passing neural networks with multi-level representations for organic molecule spectra prediction. *npj Computational Materials* **11**, 203 (2025).
10. Gómez-Bombarelli, R., *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science* **4**, 268-276 (2018).
11. Wu, Z., *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **9**, 513-530 (2018).
12. Yang, K., *et al.* Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **59**, 3370-3388 (2019).
13. Wang, Y., *et al.* Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing. *Nature Communications* **15**, 313 (2024).
14. Stokes, J.M., *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688-702.e613 (2020).
15. Altae-Tran, H., Ramsundar, B., Pappu, A.S. & Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Central Science* **3**, 283-293 (2017).
16. Liu, R., Wang, H., Glover, K.P., Feasel, M.G. & Wallqvist, A. Dissecting Machine-Learning Prediction of Molecular Activity: Is an Applicability Domain Needed for Quantitative Structure–Activity Relationship Models Based on Deep Neural Networks? *Journal of Chemical Information and Modeling* **59**, 117-126 (2019).
17. Janet, J.P. & Kulik, H.J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *The Journal of Physical Chemistry A* **121**, 8939-8954 (2017).
18. Tetko, I.V., Tanchuk, V.Y., Kasheva, T.N. & Villa, A.E.P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *Journal of Chemical Information and Computer*

738 *Sciences* **41**, 1488-1493 (2001).

739 19. Klebe, G. Applying thermodynamic profiling in lead finding and optimization. *Nature Reviews*
740 *Drug Discovery* **14**, 95-110 (2015).

741 20. Wildman, S.A. & Crippen, G.M. Prediction of Physicochemical Parameters by Atomic
742 Contributions. *Journal of Chemical Information and Computer Sciences* **39**, 868-873 (1999).

743 21. Boulougouri, M., Vandergheynst, P. & Probst, D. Molecular set representation learning.
744 *Nature Machine Intelligence* **6**, 754-763 (2024).

745 22. Zang, X., Zhao, X. & Tang, B. Hierarchical Molecular Graph Self-Supervised Learning for
746 property prediction. *Communications Chemistry* **6**, 34 (2023).

747 23. Fedik, N., et al. Extending machine learning beyond interatomic potentials for predicting
748 molecular properties. *Nature Reviews Chemistry* **6**, 653-672 (2022).

749 24. Zou, Y., et al. A structure-based framework for selective inhibitor design and optimization.
750 *Communications Biology* **8**, 422 (2025).

751 25. Edwards, C., et al. Translation between Molecules and Natural Language. in *Proceedings of*
752 *the 2022 Conference on Empirical Methods in Natural Language Processing* 375-413
753 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).

754 26. Pei, Q., et al. BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge
755 and Natural Language Associations. in *Proceedings of the 2023 Conference on Empirical*
756 *Methods in Natural Language Processing* 1102-1123 (Association for Computational
757 Linguistics, 2023).

758 27. Liu, S., et al. Pre-training Molecular Graph Representation with 3D Geometry. in *International*
759 *Conference on Learning Representations* (2022).

760 28. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of
761 representations via graph neural networks. *Nature Machine Intelligence* **4**, 279-287 (2022).

762 29. You, Y., et al. Graph Contrastive Learning with Augmentations. in *Advances in Neural*
763 *Information Processing Systems*, Vol. 33 (eds. Larochelle, H., Ranzato, M., Hadsell, R., Balcan,
764 M.F. & Lin, H.) 5812-5823 (Curran Associates, Inc., 2020).

765 30. Lin, W. & Fung, C.C.A. Rethinking the Masking Strategy for Pretraining Molecular Graphs from
766 a Data-Centric View. *ACS Omega* **9**, 20832-20838 (2024).

767 31. Hu, Z., Dong, Y., Wang, K., Chang, K.-W. & Sun, Y. GPT-GNN: Generative Pre-Training of Graph
768 Neural Networks. in *Proceedings of the 26th ACM SIGKDD Conference on Knowledge*
769 *Discovery and Data Mining* (2020).

770 32. Rong, Y., et al. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Advances in*
771 *Neural Information Processing Systems* **33**(2020).

772 33. Lu, S., Gao, Z., He, D., Zhang, L. & Ke, G. Data-driven quantum chemical property prediction
773 leveraging 3D conformations with Uni-Mol+. *Nature Communications* **15**, 7104 (2024).

774 34. Landrum, G. RDKit: A software suite for cheminformatics, computational chemistry, and
775 predictive modeling. <https://www.rdkit.org/> (2013).

776 35. Udrescu, S.-M. & Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression.
777 *Science Advances* **6**, eaay2631 (2019).

778 36. Arora, S., Hu, W. & Kothari, P.K. An Analysis of the t-SNE Algorithm for Data Visualization. in
779 *Proceedings of the 31st Conference On Learning Theory*, Vol. 75 (eds. Sébastien, B., Vianney, P.
780 & Philippe, R.) 1455--1462 (PMLR, Proceedings of Machine Learning Research, 2018).

781 37. Lai, H., Huang, T., Lu, B., Zhang, S. & Xiaog, R. Silhouette coefficient-based weighting k-means

algorithm. *Neural Computing and Applications* **37**, 3061-3075 (2025).

38. Xiao, J., Lu, J. & Li, X. Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis* **21**, 1327-1338 (2017).

39. Wang, T., Zhao, D. & Tian, S. An overview of kernel alignment and its applications. *Artificial Intelligence Review* **43**, 179-192 (2015).

40. Heid, E., *et al.* Chemprop: A Machine Learning Package for Chemical Property Prediction. *Journal of Chemical Information and Modeling* **64**, 9-17 (2024).

41. Wan, Y., Wu, J., Hou, T., Hsieh, C.-Y. & Jia, X. Multi-channel learning for integrating structural hierarchies into context-dependent molecular representation. *Nature Communications* **16**, 413 (2025).

42. Zheng, Y., *et al.* Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence* **7**, 437-447 (2025).

43. Myung, Y., de Sá, A.G.C. & Ascher, D.B. Deep-PK: deep learning for small molecule pharmacokinetic and toxicity prediction. *Nucleic Acids Research* **52**, W469-W475 (2024).

44. van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *Journal of Chemical Information and Modeling* **62**, 5938-5951 (2022).

45. De Nardo, D., *et al.* Interleukin-1 receptor–associated kinase 4 (IRAK4) plays a dual role in myddosome formation and Toll-like receptor signaling. *Journal of Biological Chemistry* **293**, 15195-15207 (2018).

46. Mendez, D., *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930-D940 (2019).

47. Ammann, S.E., *et al.* Discovery of Edecisertib (GS-5718): A Potent, Selective Inhibitor of IRAK4. *Journal of Medicinal Chemistry* **68**, 10619-10630 (2025).

48. Pfaffenbach, M., *et al.* Discovery of BIO-8169—A Highly Potent, Selective, and Brain-Penetrant IRAK4 Inhibitor for the Treatment of Neuroinflammation. *Journal of Medicinal Chemistry* **67**, 8383-8395 (2024).

49. Kiseleva, A.A., *et al.* Unexpected Activities in Regulating Ciliation Contribute to Off-target Effects of Targeted Drugs. *Clinical Cancer Research* **25**, 4179-4193 (2019).

50. Campbell, G.R., Rawat, P. & Spector, S.A. Pacritinib Inhibition of IRAK1 Blocks Aberrant TLR8 Signalling by SARS-CoV-2 and HIV-1-Derived RNA. *Journal of Innate Immunity* **15**, 96-106 (2022).

51. Hall, M.D., *et al.* Synthesis, Activity, and Pharmacophore Development for Isatin- β -thiosemicarbazones with Selective Activity toward Multidrug-Resistant Cells. *Journal of Medicinal Chemistry* **52**, 3191-3204 (2009).

52. Lin, F., *et al.* Real-World Patient Experience of Pexidartinib for Tenosynovial Giant-Cell Tumor. *The Oncologist* **29**, e535-e543 (2024).

53. Zuccotto, F., Ardini, E., Casale, E. & Angiolini, M. Through the “Gatekeeper Door”: Exploiting the Active Kinase Conformation. *Journal of Medicinal Chemistry* **53**, 2681-2694 (2010).

54. Wang, L., *et al.* Conformational flexibility and inhibitor binding to unphosphorylated interleukin-1 receptor–associated kinase 4 (IRAK4). *Journal of Biological Chemistry* **294**, 4511-4519 (2019).

55. Cereto-Massagué, A., *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58-63 (2015).

- 826 56. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Physical Review E*
827 **69**, 066138 (2004).
- 828 57. Standley, T., *et al.* Which Tasks Should Be Learned Together in Multi-task Learning? in
829 *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 (eds. Hal, D.,
830 III & Aarti, S.) 9120--9132 (PMLR, Proceedings of Machine Learning Research, 2020).
- 831 58. Shazeer, N., *et al.* Outrageously Large Neural Networks: The Sparsely-Gated
832 Mixture-of-Experts Layer. in *International Conference on Learning Representations (ICLR*,
833 Toulon, France, 2017).
- 834 59. Vaswani, A., *et al.* Attention is All you Need. in *Advances in Neural Information Processing*
835 *Systems*, Vol. 30 (eds. Guyon, I., *et al.*) (Curran Associates, Inc., 2017).
- 836 60. Wang, F., Xiang, X., Cheng, J. & Yuille, A.L. NormFace: L2 Hypersphere Embedding for Face
837 Verification. in *IEEE Conference on Computer Vision and Pattern Recognition (IEEE*, Honolulu,
838 HI, USA, 2017).
- 839 61. Liu, W., *et al.* SphereFace: Deep Hypersphere Embedding for Face Recognition. in *IEEE*
840 *International Conference on Computer Vision (IEEE*, Venice, Italy, 2017).
- 841 62. Deng, J., Guo, J., Xue, N. & Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face
842 Recognition. in *IEEE Conference on Computer Vision and Pattern Recognition (IEEE*, Long
843 Beach, CA, USA, 2019).
- 844 63. He, K., Zhang, X., Ren, S. & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level
845 Performance on ImageNet Classification. in *IEEE International Conference on Computer Vision*
846 (IEEE, Santiago, Chile, 2015).
- 847 64. Hinton, G., Vinyals, O. & Dean, J. Distilling the Knowledge in a Neural Network. in *NeurIPS*
848 *Deep Learning Workshop* (Montreal, Canada, 2015).
- 849 65. Beiran, M. & Litwin-Kumar, A. Prediction of neural activity in connectome-constrained
850 recurrent networks. *Nature Neuroscience* **28**, 2561-2574 (2025).
- 851 66. Xie, P., Wu, W., Zhu, Y. & Xing, E.P. Orthogonality-Promoting Distance Metric Learning: Convex
852 Relaxation and Theoretical Analysis. in *Proceedings of the 35th International Conference on*
853 *Machine Learning*, Vol. 80 5403-5412 (Proceedings of Machine Learning Research, 2018).
- 854 67. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception
855 Architecture for Computer Vision. in *2016 IEEE Conference on Computer Vision and Pattern*
856 *Recognition (CVPR)* 2818-2826 (2016).
- 857 68. Wu, T., *et al.* Rethinking Kullback-Leibler Divergence in Knowledge Distillation for Large
858 Language Models. 5737-5755 (Association for Computational Linguistics, Abu Dhabi, UAE,
859 2025).
- 860 69. Niu, Z., Zhou, M., Wang, L., Gao, X. & Hua, G. Ordinal Regression with Multiple Output CNN
861 for Age Estimation. in *2016 IEEE Conference on Computer Vision and Pattern Recognition*
862 *(CVPR)* 4920-4928 (2016).
- 863 70. Zhao, X., Qi, L., An, Y. & Geng, X. Generalizable Label Distribution Learning. in *Proceedings of*
864 *the 31st ACM International Conference on Multimedia* 8932--8941 (Association for Computing
865 Machinery, Ottawa ON, Canada, 2023).
- 866 71. Masoudnia, S. & Ebrahimpour, R. Mixture of experts: a literature survey. *Artificial Intelligence*
867 *Review* **42**, 275-293 (2014).
- 868 72. Diederik P. Kingma, J.B. Adam: A Method for Stochastic Optimization. *arXiv*, 1412.6980
869 (2017).

- 870 73. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation
871 Hyperparameter Optimization Framework. in *Proceedings of the 25th ACM SIGKDD*
872 *International Conference on Knowledge Discovery & Data Mining* 2623–2631 (Association for
873 Computing Machinery, Anchorage, AK, USA, 2019).
- 874 74. Xiong, Z., *et al.* Pushing the Boundaries of Molecular Representation for Drug Discovery with
875 the Graph Attention Mechanism. *Journal of Medicinal Chemistry* **63**, 8749-8760 (2020).
- 876 75. Liu, S., Demirel, M.F. & Liang, Y. N-gram Graph: Simple Unsupervised Representation for
877 Graphs, with Applications to Molecules. in *Advances in Neural Information Processing*
878 *Systems*, Vol. 32 8464-8476 (Curran Associates, Inc., Vancouver, Canada, 2019).
- 879 76. Hu, W., *et al.* Strategies for Pre-training Graph Neural Networks. in *International Conference*
880 *on Learning Representations* (ICLR, 2020).
- 881 77. Fang, X., *et al.* Geometry-enhanced molecular representation learning for property prediction.
882 *Nature Machine Intelligence* **4**, 127-134 (2022).
- 883 78. Zeng, X., *et al.* Accurate prediction of molecular properties and drug targets using a
884 self-supervised image representation learning framework. *Nature Machine Intelligence* **4**,
885 1004-1016 (2022).
- 886

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformationArcMol.docx](#)
- [ArcMolnrsoftwarepolicy.pdf](#)
- [ArcMolmachinelearningchecklist.pdf](#)
- [sourcedata.zip](#)