Suppl. Note 1. Description of modalities and model selection

The main focus of this work is the multimodality, and in this section we explain the modalities selected and the models used for each modality. We focus on multimodality over molecular modalities, so we use only one modality for proteins in the drug-target interaction tasks, and seven different modalities for molecules. For the other modalities (i.e. data properties of the measurements in the BDB MKG, 2.1, we infer the type from the data (either text, or number).

1.1 Numbers

For numbers, we create a tensor (a.k.a. embedding) directly from the value, resulting in an embedding of only one value, the number present in the data property.

1.2 Text

Following [6] we embed the texts using a language model, a sentence transformer [15] that maps sentences and paragraphs to a 768 dimensional dense vector space. We use the same checkpoint, which is publicly available: `https://huggingface.co/sentence-transformers/paraphrase-albert-small-v2`.

1.3 Evolutionary Scale Modeling - ESM

In our study, we use Evolutionary Scale Modeling - ESM [16] to characterise protein sequences. ESM is a transformer model trained using masked language modeling on the UniRef50 protein sequence database [20], learning contextual representations that capture evolutionary and structural information, enabling downstream tasks such as zero-shot mutation effect prediction and unsupervised structure inference.

In our experiments, we used the mean token embeddings extracted from the 33rd layer of the *esm1b_t33_650M_UR50S* checkpoint available via Hugging Face at `https://huggingface.co/facebook/esm1b_t33_650M_UR50S`, with all sequences are truncated to 1022 amino acids.

We did some experiments comparing the ESM-1B model against the ESM-2 (checkpoint *esm2_t33_650M_UR50D* available at `https://huggingface.co/facebook/esm2_t33_650M_UR50D`), to see which model performed better. The results, Suppl. Table 1, show that ESM-1B is better (on average) in 6 out of the 7 experiments, and thus we use it in the rest of the experiments.

1.4 Morgan Fingerprint

Morgan Fingerprint, equivalent to the ECFP4 fingerprint of the Extended-Connectivity Fingerprint family [17], is a circular topological fingerprint for

molecules that considers both the neighborhood of atoms in molecules and feature extraction kernels to generate a count vector for molecular characterisation purposes. Therefore, Morgan Fingerprint is a count-based fingerprint, and it has been extensively used in tasks such as substructure searching, similarity assessment, virtual screening for small molecules, and target prediction benchmarking, tasks in which it has proven to be highly performant.

In our experiments, we used the Morgan Fingerprint model available within the RDKit python library (`https://github.com/rdkit/rdkit`) with a shape of 2048 (nBits) and a radius of 2. However, whenever RDKit fails to generate a valid fingerprint, we use an embedding of a suitable shape filled with zeros.

1.5 ImageMol

ImageMol [23] is an unsupervised molecular image pretraining framework that combines comprehensive knowledge on molecular chemistry with an image processing infrastructure with the aim to extract pixel-level molecular features. ImageMol uses molecular images as the feature representation of compounds which allow for high accuracy and low computing costs, and exhibits chemical awareness concerning molecular structure enabled via large-scale unsupervised pretrained learning on ∼10 million drug-like compounds with diverse biological activities.

ImageMol's molecular encoder is pretrained via 5 strategies: i) structural classification, to predict chemical structural information; ii) rationality classification, to distinguish between rational and irrational molecules; iii) jigsaw classification, to predict rational permutations; iv) contrastive classification, to maximise the similarity between original and masked images; and v) image generation, to discriminate between real and fake molecular images. ImageMol's accuracy is demonstrated in several drug discovery tasks including drug-like property assessment and molecular target prediction.

In our experiments, we used the publicly available ImageMol's pretrained model *ImageMol.pth.tar*, referenced in the official ImageMol GitHub repository `https://github.com/HongxinXiang/ImageMol` and made available at `https://drive.google.com/file/d/1wQfby8JIhgo3DxPvFeHXPc14wS-b4KB5`. From such checkpoint, we select the molecular encoder's weights for the *ImageMol encoding-only* (IM) variant, and added the weights of the 100-neuron structural classification layer for the *ImageMol encoding with classification layer* (IMC) variant. Concerning the use of the IM and IMC variants, we preselected them for early, small-scale comparative evaluation in the context of 3 drug-target binding affinity datasets in the TDC benchmark suites, namely DTI DG, DAVIS, and KIBA, with initial embeddings for proteins obtained via ESM1b and initial embeddings for drugs obtained via the IM/IMC. From the comparative results in Suppl. Table 2, we observe the following:

- Number of local wins (italic entries): 6 times IM, 10 times IMC.
- Number of global first places (bold entries): 3 times IM, 5 times IMC.

- Number of global second places (underlined entries): 4 times IM, 4 times IMC.

From the above, IMC was the selected ImageMol variant for the large-scale experiments reported in this article.

1.6 ChemBERTa

ChemBERTa [1] is a RoBERTa-based transformer model with 12 attention heads and 6 layers, pretrained on PubChem data and fine-tuned on MoleculeNet data, suitable for molecular property prediction, drug design, chemical modelling. For pretraining, a dataset of 77M unique SMILES was extracted from PubChem and properly curated to facilitate large-scale pretraining on up to 10M subsets. On the other hand, fine-tuning was based on several MoleculeNet classifications tasks with a wide range of dataset sizes (1.5K to 41.K datapoints) and medicinal chemistry applications.

The base code of ChemBERTa is available via GitHub (`https://github.com/seyonechithrananda/bert-loves-chemistry`), and the pretrained checkpoints are available via Hugging Face. In our experiments, we used the checkpoint *ChemBERTa-zinc-base-v1*, which is available at `https://huggingface.co/seyonec/ChemBERTa-zinc-base-v1`.

1.7 MolFormer

MolFormer [18] is a large-scale chemical language model designed with the intention of learning a model trained on small molecules which are represented as SMILES strings. MolFormer leverges Masked Language Modeling and employs a linear attention Transformer combined with rotary embeddings.

In our experiments, we used the checkpoint *N-Step-Checkpoint_3_30000.ckpt* available in the official MolfFormer GitHub repository `https://github.com/IBM/molformer`.

1.8 SMI-TED

SMI-TED [19] is a SMILES-based Transformer Encoder-Decoder model developed by IBM, pre-trained on a curated dataset of 91 million canonicalized SMILES samples sourced from PubChem - equivalent to 4 billion molecular tokens. For pretraining, two strategies were employed: a masked language model approach to train the encoder, and an encoder-decoder strategy to refine SMILES reconstruction and improve the quality of the generated latent space. The pretrained model have demonstrated competitive performance on classification and regression benchmarks from MoleculeNet, including predicting molecular quantum mechanics.

In our experiments, we used the SMI-TED checkpoint available via Hugging Face at `https://huggingface.co/ibm-research/materials.smi-ted/blob/main/smi-ted-Light_40.pt`

1.9 MolGen

MolGen-large [5] is a pre-trained molecular generative model built using SELFIES [11]. The training corpus comprises over 100 million molecules in SELFIES representation. MolGen-large employs a bidirectional Transformer as its encoder and an autoregressive Transformer as its decoder, and it can learn the intrinsic structural patterns of molecules by mapping corrupted SELFIES to their original forms.

In our experiments, we used the *MolGen-large* checkpoint available via Hugging Face at `https://huggingface.co/zjunlp/MolGen-large`.

1.10 Uni-Mol

Uni-Mol [24] is a powerful framework for molecular representation learning that emphasizes the use of 3D structural information. It employs two large-scale pretrained SE(3) Transformer models–one trained on molecular conformations and the other on protein pockets–to generate rich, spatially-informed representations. Equipped with tailored finetuning strategies, Uni-Mol adapts seamlessly to a broad spectrum of molecular tasks, ranging from property prediction to 3D structure-based applications. By incorporating three-dimensional geometry at its core, Uni-Mol sets new benchmarks in both conventional molecular property tasks and more complex 3D challenges such as binding pose prediction.

In our study, we utilized Uni-Mol to generate 3D conformations of drug molecules from their SMILES representations and extracted embeddings from the pretrained models. These embeddings provide compact and informative summaries of the small molecules' 3D structures.

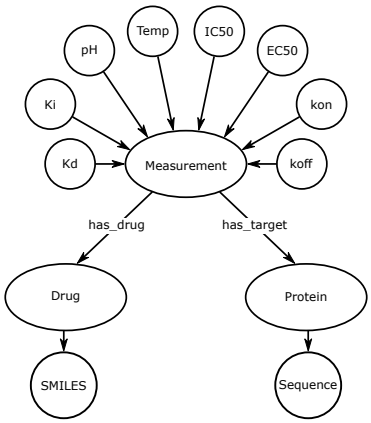
Uni-Mol's basecode is available at `https://github.com/deepmodeling/Uni-Mol`, while pre-trained models are available via Hugging Face. In our experiments, we used the model weights corresponding to the default *unimolv1* model, which are available at `https://huggingface.co/dptech/Uni-Mol-Models`.

Suppl. Note 2. Benchmarks and datasets

2.1 Pretraining Multimodal Knowledge Graphs

For pretraining the inductive graph neural network (Otter-Knowledge), we need a Multimodal Knowledge Graph (MKG). We create one from BindingDB [13] and UniProt [4]. BindingDB consists of 2,656,221 data points involving 1.2 million compounds and 9,000 targets. Following [6], we create triples for every drug-protein pair in an entity called Measurement. To avoid data redundancy, we remove duplicated triples from the TDC DTI dataset. Uniprot comprises 573,227 SwissProt proteins (from curated UniProt subset).

We consider three entities in our MKG: *Measurement*, *Drug* and *Protein*. We use the *Measurement* to relate the *Protein* and the *Drug*, using the object relationships *has_target* and *has_drug* respectively. The *Measurement* also has different attributes, like *Kd*, *Ki*, *IC50*, etc. These set of attributes is fixed, this is,



Supplementary Figure 1: Diagram of the BDB MKG used for the experiments.

we don't vary these attributes/modalities across A diagram of the MKG used is shown in Suppl. Fig. 1. For the *Drug*, the number of attributes will vary from experiment to experiment, using different modalities, including: fingerprints, SMILES, 2D image, 3D structure, SELFIES. *Protein* nodes contain only one fixed attribute, the protein sequence, coming from Uniprot, and represented using ESM-1b (see 1.3).

2.2 DTI Benchmarks

We evaluate all the approaches over the TDC DTI Datasets [3]: DG (based on BindingDB patents), KIBA and DAVIS. For KIBA and DAVIS we use three different splits, one random split with seed 0, one cold-start split based on the drug and one cold-start split based on the targets. Cold-start first splits on one entity type (either drug or target) into train/valid/test and then move all pairs associated with that entity in each set as the final splits. Suppl. Table 3 shows an overview of the datasets and splits we use.

2.3 ADMET Benchmarks

For each dataset in the ADMET benchmark group [2], we follow the standard scaffold split recommended in the TDC benchmark to partition the data into training, validation, and test sets. The training and validation sets are merged to form a single training set, while the test set provided by the TDC benchmark remains untouched until the final evaluation.

Four different metrics are used depending on the nature of the benchmark task. The Area Under the Receiver Operating Characteristic Curve (AUROC) is used for binary classification tasks where the numbers of positive and negative samples are relatively balanced. The Area Under the Precision-Recall Curve (AUPRC) is used for binary classification tasks with significant class im-

balance, particularly when positive samples are much fewer than negative ones. The Mean Absolute Error (MAE) is employed for most regression benchmarks. Spearman's rank correlation coefficient is used for regression tasks where target values depend on factors beyond chemical structure. Dataset statistics and additional information can be found in Suppl. Table 4.

Suppl. Note 3. Analysis of the impact of individual models vs diversity of modalities

To investigate whether the improvements observed in the multimodal setting arise primarily from the integration of distinct modalities rather than from the diversity of individual models, we conducted an additional controlled comparison in which models belonging to the same modality type were systematically substituted within the same experimental configurations.

Specifically, we compared the results of the experiments using sequence-based modalities (SMILES/SELFIES representations). This experiments include the following models: MOLF; SMI; MG. We keep MF as the main modality, and analyse the impact of adding other modalities to it. The goal of this analysis was to determine whether the observed performance trends are primarily driven by the modality type or by the specific model instance used to represent that modality.

The comparison was performed on the DG dataset, which is the largest dataset considered in this study and therefore provides the most reliable setting for controlled comparisons. In addition, we used the FFNN-IF architecture, which is the simplest fusion architecture in our benchmark. This choice minimizes potential confounding factors that could arise from more complex components such as knowledge graph integration or graph neural networks used in OK.

Suppl. Table 5 reports the results obtained when adding the sequence modality to the other experimental configurations. The first row corresponds to the experiment using only MF, while subsequent rows progressively introduce additional modalities (e.g., IMC, UM or both).

The results show that models belonging to the same modality family produce highly consistent trends across configurations. For example, when a sequence modality is added to the MF baseline, all evaluated sequence-based models produce a comparable improvement in performance. In contrast, adding spatial (IMC) or geometric (UM) modalities to the MF baseline does not produce a similar improvement in this setting (Suppl. Table 5, second and third rows, Original column). A comparable pattern is observed when introducing a third modality in the IMC–MF–UM configuration (fourth row).

To facilitate interpretation, these results are also visualized in Suppl. Fig. 2. The figure highlights that sequence-based encoders follow a consistent performance trend across model variants, whereas the addition of spatial or geometric modalities produces different effects. This behavior suggests that the dominant

factor influencing the performance trend is the modality type rather than the specific encoder used to implement that modality.

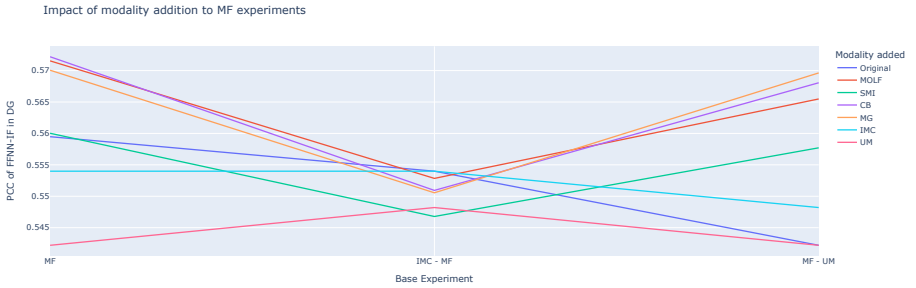
To further quantify this observation, we conducted a pairwise sign analysis. When comparing models within the same modality, the results are mixed (e.g., 2–2 and 3–1 splits across pairs), with no model consistently outperforming others. This lack of transitivity indicates that performance differences are configuration-dependent and do not reflect a systematic advantage of any specific model.

In contrast, comparisons across configurations with different modality compositions exhibit a fully consistent pattern: all pairwise comparisons result in unanimous outcomes (4–0), indicating that modality changes induce systematic performance differences across all model variants.

The detailed pairwise comparisons are as follows:

- MOLF vs SMI: 4 - 0
- MOLF vs CB: 2 - 2
- MOLF vs MG: 3 - 1
- SMI vs CB: 1 - 3
- SMI vs MG: 1 - 3
- CB vs MG: 3 - 1
- MF vs IMC - MF: 4 - 0
- MF vs MF - UM: 4 - 0
- MF vs IMC - MF - UM: 4 - 0
- IMC - MF vs MF - UM: 0 - 4
- IMC - MF vs IMC - MF - UM: 0 - 4
- MF - UM vs IMC - MF - UM: 4 - 0

Overall, this analysis indicates that while individual model choices can produce small variations in performance, the primary trends observed in our multimodal experiments are largely driven by the modality being introduced rather than by the diversity of models themselves. These results therefore support the interpretation that the improvements reported in the manuscript reflect the integration of complementary molecular representations rather than simply the aggregation of heterogeneous model architectures.



Supplementary Figure 2: **Impact of sequence-based model substitution on multimodal performance.** Performance obtained adding modalities to experiments containing Morgan Fingerprint (MF) in the DG dataset using the Feed-Forward Neural Network architecture with Intermediate Fusion (FFNN-IF). Each curve corresponds to a different modality added, while the x-axis represents the experimental configuration. Models belonging to the same modality family exhibit consistent performance trends across configurations, whereas configurations involving spatial (Imagemol, IMC) and geometric (Unimol, UM) modalities display different behaviors. This suggests that the dominant performance trend is primarily associated with the modality type rather than the specific model used to implement that modality.

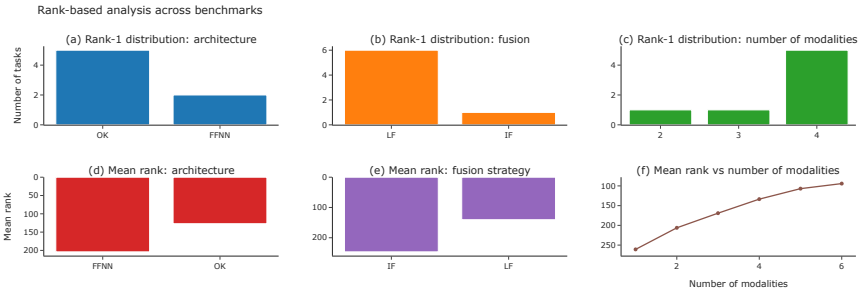
Suppl. Note 4. Ranking

To complement the comparison based on raw performance values, we additionally analyze the relative ranking of experimental configurations across benchmarks. For each benchmark, experiments are ranked according to their performance, allowing a comparison that is independent of the scale of the evaluation metric. Suppl. Fig. 3 summarizes the resulting rank distributions across benchmarks for DTI.

Panels (a) and (b) show the number of benchmarks for which each architecture and fusion strategy achieves the best rank. The OK architecture obtains the top rank in the majority of benchmarks, while late fusion (LF) consistently outperforms intermediate fusion (IF). Panels (d) and (e) further confirm this trend when considering the mean rank across all experiments, where lower values indicate better relative performance.

Finally, panels (c) and (f) analyze the effect of multimodality. The number of modalities used in an experiment shows a clear relationship with performance: configurations incorporating more modalities tend to achieve better ranks across benchmarks. In particular, the mean rank decreases monotonically as the number of modalities increases, suggesting that integrating multiple molecular representations systematically improves predictive performance across heterogeneous benchmarks.

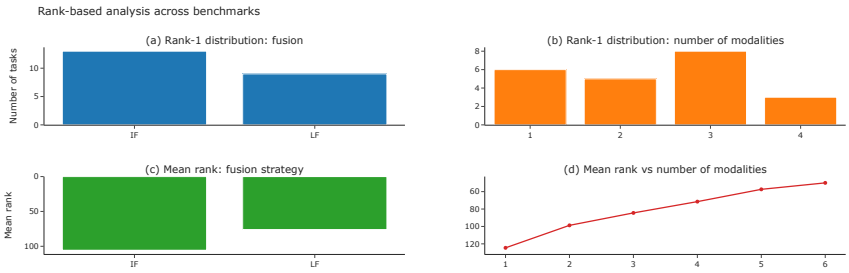
We further analyze the ADMET benchmark suite to assess the generality



Supplementary Figure 3: **Rank-based analysis of architectural and multimodal design choices across benchmarks.** (a) Number of tasks in which each architecture achieves the top rank. (b) Number of tasks in which each fusion strategy achieves the top rank. (c) Number of tasks in which experiments with a given number of molecular modalities achieve the top rank. (d,e) Mean rank across benchmarks for architectures and fusion strategies, respectively (lower rank indicates better performance). (f) Mean rank as a function of the number of modalities used in the experiment. Together, these analyses complement the comparison based on raw performance values by highlighting the relative consistency of different design choices across heterogeneous benchmarks.

of the trends observed in drug–target interaction tasks. Since only a single architecture is used in this case, we focus on the impact of fusion strategy and the number of molecular modalities. Panels (a) and (b) of Suppl. Fig. 4 show the number of tasks in which each fusion strategy achieves the top rank and the distribution of top-ranked tasks across different numbers of modalities, respectively. Late fusion (LF) consistently outperforms intermediate fusion (IF), achieving a lower mean rank across tasks (75.98 vs. 105.67), confirming the advantage of combining representations at a later stage.

Panels (c) and (d) illustrate the relationship between the number of modalities and performance. The mean rank decreases monotonically from 124.56 for unimodal experiments to 49.95 when six modalities are integrated, demonstrating a clear benefit of multimodal integration. These results indicate that the performance gains observed in ADMET property prediction are primarily driven by the incorporation of complementary molecular modalities rather than by the specific choice of fusion strategy alone, providing strong evidence that the positive effects of multimodality are robust across diverse molecular prediction tasks.



Supplementary Figure 4: **Rank-based analysis of fusion and multimodal design choices across ADMET benchmarks.** (a) Number of tasks in which each fusion strategy achieves the top rank. (b) Number of tasks in which experiments with a given number of molecular modalities achieve the top rank. (c) Mean rank across benchmarks for the fusion strategies (lower rank indicates better performance). (d) Mean rank as a function of the number of modalities used in the experiment. Together, these analyses complement the comparison based on raw performance values by highlighting the relative consistency of different design choices across heterogeneous benchmarks.

Suppl. Note 5. State of the art

Supplementary Tables 6, 7, 8, 9 10 and 11 compare the best result we obtain with the state of the art.

Suppl. Note 6. Statistical details

We use statistical tests to detect any significant difference between experiments. As we are comparing different models across the same benchmarks (i.e. paired experiments), we use Wilcoxon test to measure any significant difference between the distribution of the results. We first run a two-sided Wilcoxon test, with a *p-value threshold* of 0.05. The null hypothesis is that the two distributions are similar. If the result is positive (i.e. the *p-value* is less than the threshold), we can reject the null hypothesis and say that the two distributions are different, and in that case we conduct the Wilcoxon test again to detect which distribution is greater (or smaller in the case of ADMET MAE benchmarks. We also asume a threshold of 0.05. If the p-value ¡ $\alpha = 0.05$, $H0$ can be rejected, meaning that the second population is greater (or smaller in the case of ADMET MAE benchmarks) than the first one. If one of the experiments is not reported is because the wilcoxon test was not positive (i.e. the *p-value* is greater than 0.05). We use the SciPy package to run the tests [22].

6.1 DTI

We compute the mean result of each benchmark for all the models grouped by the number of molecular modalities; g_N is the group of models combining N molecular modalities.

6.1.1 Intermediate fusion

For the knowledge-enhanced multimodal neural network (Otter-Knowledge), results of the Wilcoxon Signed-Rank test indicate that there is

- a significant large difference between g_1 (Mdn = 0.5990, n = 7) and g_2 (Mdn = 0.6076, n = 7), W+ = 27.0, p = 0.016, r = 0.914
- a significant large difference between g_1 (Mdn = 0.5990, n = 7) and g_3 (Mdn = 0.6128, n = 7), W+ = 27.0, p = 0.016, r = 0.914
- a significant large difference between g_1 (Mdn = 0.5990, n = 7) and g_4 (Mdn = 0.6217, n = 7), W+ = 27.0, p = 0.016, r = 0.914
- a significant large difference between g_2 (Mdn = 0.6076, n = 7) and g_3 (Mdn = 0.6128, n = 7), W+ = 26.0, p = 0.023, r = 0.857
- a significant large difference between g_2 (Mdn = 0.6076, n = 7) and g_4 (Mdn = 0.6217, n = 7), W+ = 26.0, p = 0.023, r = 0.857

For the simple neural network architecture, results of the Wilcoxon Signed-Rank test indicate that there is no statistically significant difference between any pair of groups.

We also compare Otter-Knowledge (OK-IF) against the simple neural network (SNN-IF), taking the average score across benchmarks for all the experiments, both using a intermediate fusion approach. Results of the Wilcoxon Signed-Rank test indicate that there is a significant large difference between $SNN - IF$ (Mdn = 0.5477, n = 39) and $OK - IF$ (Mdn = 0.6205, n = 39), W+ = 780.0, p = 0.000, r = 1.129.

6.1.2 Late fusion

For the knowledge-enhanced multimodal neural network (Otter-Knowledge), results of the Wilcoxon Signed-Rank test indicate that there is

- a significant large difference between g_1 (Mdn = 0.5990, n = 7) and g_2 (Mdn = 0.6399, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.5990, n = 7) and g_3 (Mdn = 0.6575, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.5990, n = 7) and g_4 (Mdn = 0.6653, n = 7), W+ = 28.0, p = 0.008, r = 1.005

- a significant large difference between g_1 (Mdn = 0.5990, n = 7) and g_5 (Mdn = 0.6697, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.5990, n = 7) and g_6 (Mdn = 0.6726, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.6399, n = 7) and g_3 (Mdn = 0.6575, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.6399, n = 7) and g_4 (Mdn = 0.6653, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.6399, n = 7) and g_5 (Mdn = 0.6697, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.6399, n = 7) and g_6 (Mdn = 0.6726, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_3 (Mdn = 0.6575, n = 7) and g_4 (Mdn = 0.6653, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_3 (Mdn = 0.6575, n = 7) and g_5 (Mdn = 0.6697, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_3 (Mdn = 0.6575, n = 7) and g_6 (Mdn = 0.6726, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_4 (Mdn = 0.6653, n = 7) and g_5 (Mdn = 0.6697, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_4 (Mdn = 0.6653, n = 7) and g_6 (Mdn = 0.6726, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_5 (Mdn = 0.6697, n = 7) and g_6 (Mdn = 0.6726, n = 7), W+ = 28.0, p = 0.008, r = 1.005

For the simple neural network architecture, results of the Wilcoxon Signed-Rank test indicate that there is

- a significant large difference between g_1 (Mdn = 0.6001, n = 7) and g_2 (Mdn = 0.6134, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.6001, n = 7) and g_3 (Mdn = 0.6174, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.6001, n = 7) and g_4 (Mdn = 0.6194, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.6001, n = 7) and g_5 (Mdn = 0.6206, n = 7), W+ = 28.0, p = 0.008, r = 1.005

- a significant large difference between g_1 (Mdn = 0.6001, n = 7) and g_6 (Mdn = 0.6279, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.6134, n = 7) and g_3 (Mdn = 0.6174, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.6134, n = 7) and g_4 (Mdn = 0.6194, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.6134, n = 7) and g_5 (Mdn = 0.6206, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.6134, n = 7) and g_6 (Mdn = 0.6279, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_3 (Mdn = 0.6174, n = 7) and g_4 (Mdn = 0.6194, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_3 (Mdn = 0.6174, n = 7) and g_5 (Mdn = 0.6206, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_3 (Mdn = 0.6174, n = 7) and g_6 (Mdn = 0.6279, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_4 (Mdn = 0.6194, n = 7) and g_5 (Mdn = 0.6206, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_4 (Mdn = 0.6194, n = 7) and g_6 (Mdn = 0.6279, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_5 (Mdn = 0.6206, n = 7) and g_6 (Mdn = 0.6279, n = 7), W+ = 28.0, p = 0.008, r = 1.005

We also compare Otter-Knowledge (OK-LF) against the simple neural network (SNN-LF), taking the average score across benchmarks for all the experiments, both using a late fusion approach. Results of the Wilcoxon Signed-Rank test indicate that there is a significant large difference between $SNN-LF$ (Mdn = 0.6341, n = 126) and $OK-LF$ (Mdn = 0.6736, n = 126), W+ = 7,869.0, p = 0.000, r = 0.846.

6.1.3 Intermediate Fusion vs Late Fusion

We compare each of the approaches (Otter-Knowledge, Simple Neural Network) with intermediate fusion and with late fusion, taking the average score across benchmarks for all the experiments.

Comparing Otter-Knowledge Intermediate Fusion (OK-IF) against Otter-Knowledge Late Fusion (OK-LF), results of the Wilcoxon Signed-Rank test indicate that there is a significant large difference between $OK-IF$ (Mdn = 0.6205, n = 39) and $OK-LF$ (Mdn = 0.6704, n = 39), W+ = 528.0, p = 0.000, r = 0.812.

Comparing the Simple Neural Network Intermediate Fusion (SNN-IF) against Simple Neural Network Late Fusion (SNN-LF), results of the Wilcoxon Signed-Rank test indicate that there is a significant large difference between $SNN-IF$ (Mdn = 0.5477, n = 39) and $SNN-LF$ (Mdn = 0.6233, n = 39), W+ = 780.0, p = 0.000, r = 1.129.

6.2 Wilcoxon Test - ADMET

We compute the mean result of each benchmark for all the models grouped by the number of molecular modalities; g_N is the group of models combining N molecular modalities.

6.2.1 Intermediate Fusion

We report the results for the simple neural network architecture with intermediate fusion.

For the benchmarks using SCC as metric, results of the Wilcoxon Signed-Rank test indicate that there is no statistically significant difference between any pair of groups.

For the benchmarks using MAE as metric, results of the Wilcoxon Signed-Rank test indicate that there is

- a significant large difference between g_1 (Mdn = 0.7600, n = 5) and g_2 (Mdn = 0.6744, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_1 (Mdn = 0.7600, n = 5) and g_3 (Mdn = 0.6667, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_1 (Mdn = 0.7600, n = 5) and g_4 (Mdn = 0.6591, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_2 (Mdn = 0.6744, n = 5) and g_3 (Mdn = 0.6667, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_2 (Mdn = 0.6744, n = 5) and g_4 (Mdn = 0.6591, n = 5), W+ = 0.0, p = 0.031, r = 0.963

Note that, for the benchmarks using MAE, the lower is the better, so results above mean that, taking as example the first row, g_2 is lower than g_1.

For the benchmarks using AUROC as metric, results of the Wilcoxon Signed-Rank test indicate that there is

- a significant large difference between g_1 (Mdn = 0.8541, n = 7) and g_2 (Mdn = 0.8847, n = 7), W+ = 27.0, p = 0.016, r = 0.914
- a significant large difference between g_1 (Mdn = 0.8541, n = 7) and g_3 (Mdn = 0.8984, n = 7), W+ = 27.0, p = 0.016, r = 0.914

- a significant large difference between g_4 (Mdn = 0.8947, n = 7) and g_3 (Mdn = 0.8984, n = 7), W+ = 27.0, p = 0.016, r = 0.914

For the benchmarks using AUPRC as metric, results of the Wilcoxon Signed-Rank test indicate that there is

- a significant large difference between g_1 (Mdn = 0.5926, n = 6) and g_2 (Mdn = 0.6331, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_1 (Mdn = 0.5926, n = 6) and g_3 (Mdn = 0.6354, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_1 (Mdn = 0.5926, n = 6) and g_4 (Mdn = 0.6246, n = 6), W+ = 19.0, p = 0.047, r = 0.811
- a significant large difference between g_4 (Mdn = 0.6246, n = 6) and g_3 (Mdn = 0.6354, n = 6), W+ = 20.0, p = 0.031, r = 0.879

6.2.2 Late Fusion

Here, we report the results for the simple neural network architecture with late fusion.

For the benchmarks using SCC as metric, results of the Wilcoxon Signed-Rank test indicate that there is no statistically significant difference between any pair of groups.

For the benchmarks using MAE as metric, results of the Wilcoxon Signed-Rank test indicate that there is

- a significant large difference between g_1 (Mdn = 0.7598, n = 5) and g_2 (Mdn = 0.7115, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_1 (Mdn = 0.7598, n = 5) and g_3 (Mdn = 0.6948, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_1 (Mdn = 0.7598, n = 5) and g_4 (Mdn = 0.6864, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_1 (Mdn = 0.7598, n = 5) and g_5 (Mdn = 0.6812, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_1 (Mdn = 0.7598, n = 5) and g_6 (Mdn = 0.6776, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_2 (Mdn = 0.7115, n = 5) and g_3 (Mdn = 0.6948, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_2 (Mdn = 0.7115, n = 5) and g_4 (Mdn = 0.6864, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_2 (Mdn = 0.7115, n = 5) and g_5 (Mdn = 0.6812, n = 5), W+ = 0.0, p = 0.031, r = 0.963

- a significant large difference between g_2 (Mdn = 0.7115, n = 5) and g_6 (Mdn = 0.6776, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_3 (Mdn = 0.6948, n = 5) and g_4 (Mdn = 0.6864, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_3 (Mdn = 0.6948, n = 5) and g_5 (Mdn = 0.6812, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_3 (Mdn = 0.6948, n = 5) and g_6 (Mdn = 0.6776, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_4 (Mdn = 0.6864, n = 5) and g_5 (Mdn = 0.6812, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_4 (Mdn = 0.6864, n = 5) and g_6 (Mdn = 0.6776, n = 5), W+ = 0.0, p = 0.031, r = 0.963
- a significant large difference between g_5 (Mdn = 0.6812, n = 5) and g_6 (Mdn = 0.6776, n = 5), W+ = 0.0, p = 0.031, r = 0.963

Note that, for the benchmarks using MAE, the lower is the better, so results above mean that, taking as example the first row, g_2 is lower than g_1.

For the benchmarks using AUROC as metric, results of the Wilcoxon Signed-Rank test indicate that there is

- a significant large difference between g_1 (Mdn = 0.8545, n = 7) and g_2 (Mdn = 0.8886, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.8545, n = 7) and g_3 (Mdn = 0.9005, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.8545, n = 7) and g_4 (Mdn = 0.9068, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.8545, n = 7) and g_5 (Mdn = 0.9106, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_1 (Mdn = 0.8545, n = 7) and g_6 (Mdn = 0.9134, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.8886, n = 7) and g_3 (Mdn = 0.9005, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.8886, n = 7) and g_4 (Mdn = 0.9068, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.8886, n = 7) and g_5 (Mdn = 0.9106, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_2 (Mdn = 0.8886, n = 7) and g_6 (Mdn = 0.9134, n = 7), W+ = 28.0, p = 0.008, r = 1.005

- a significant large difference between g_3 (Mdn = 0.9005, n = 7) and g_4 (Mdn = 0.9068, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_3 (Mdn = 0.9005, n = 7) and g_5 (Mdn = 0.9106, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_3 (Mdn = 0.9005, n = 7) and g_6 (Mdn = 0.9134, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_4 (Mdn = 0.9068, n = 7) and g_5 (Mdn = 0.9106, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_4 (Mdn = 0.9068, n = 7) and g_6 (Mdn = 0.9134, n = 7), W+ = 28.0, p = 0.008, r = 1.005
- a significant large difference between g_5 (Mdn = 0.9106, n = 7) and g_6 (Mdn = 0.9134, n = 7), W+ = 28.0, p = 0.008, r = 1.005

For the benchmarks using AUPRC as metric, results of the Wilcoxon Signed-Rank test indicate that there is

- a significant large difference between g_1 (Mdn = 0.6171, n = 6) and g_2 (Mdn = 0.6700, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_1 (Mdn = 0.6171, n = 6) and g_3 (Mdn = 0.6899, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_1 (Mdn = 0.6171, n = 6) and g_4 (Mdn = 0.7007, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_1 (Mdn = 0.6171, n = 6) and g_5 (Mdn = 0.7076, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_1 (Mdn = 0.6171, n = 6) and g_6 (Mdn = 0.7122, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_2 (Mdn = 0.6700, n = 6) and g_3 (Mdn = 0.6899, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_2 (Mdn = 0.6700, n = 6) and g_4 (Mdn = 0.7007, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_2 (Mdn = 0.6700, n = 6) and g_5 (Mdn = 0.7076, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_2 (Mdn = 0.6700, n = 6) and g_6 (Mdn = 0.7122, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_3 (Mdn = 0.6899, n = 6) and g_4 (Mdn = 0.7007, n = 6), W+ = 21.0, p = 0.016, r = 0.987

- a significant large difference between g_3 (Mdn = 0.6899, n = 6) and g_5 (Mdn = 0.7076, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_3 (Mdn = 0.6899, n = 6) and g_6 (Mdn = 0.7122, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_4 (Mdn = 0.7007, n = 6) and g_5 (Mdn = 0.7076, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_4 (Mdn = 0.7007, n = 6) and g_6 (Mdn = 0.7122, n = 6), W+ = 21.0, p = 0.016, r = 0.987
- a significant large difference between g_5 (Mdn = 0.7076, n = 6) and g_6 (Mdn = 0.7122, n = 6), W+ = 21.0, p = 0.016, r = 0.987

6.2.3 Intermediate Fusion vs Late Fusion

When comparing the simple neural network architecture using intermediate fusion against the simple neural network architecture using late fusion, the Wilcoxon Signed-Rank test indicate that there is no statistically significant difference between any pair of groups, for any of the metrics.

Supplementary References

[1] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*, 2022.

[2] Therapeutics Data Commons. ADMET Benchmark Group — tdcommons.ai. `https://tdcommons.ai/benchmark/admet_group/overview/`. [Accessed 14-05-2025].

[3] Therapeutics Data Commons. Drug-Target Interaction — tdcommons.ai. `https://tdcommons.ai/multi_pred_tasks/dti/`. [Accessed 14-05-2025].

[4] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 11 2022.

[5] Yin Fang, Ningyu Zhang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Domain-agnostic molecular generation with chemical feedback. In *ICLR*. OpenReview.net, 2024.

[6] Thanh Lam Hoang, Marco Luca Sbodio, Marcos Martinez Galindo, Mykhaylo Zayats, Raul Fernandez-Diaz, Victor Valls, Gabriele Picco, Cesar Berrospi, and Vanessa Lopez. Knowledge enhanced representation learning for drug discovery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):10544–10552, Mar. 2024.

[7] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay S. Pande, and Jure Leskovec. Pre-training graph neural networks. *CoRR*, abs/1905.12265, 2019.

[8] David Huang, Sauhaarda (Raunak) Chowdhuri, Andrew Li, Alex Li, Ayush Agrawal, Kameron Gano, and Andy Zhu. A unified system for molecular property predictions: Oloren chemengine and its applications. *ChemRxiv*, 2022(1020), 2022.

[9] Nan jiang, Mohammed Quazi, Christina Schweikert, D. Frank Hsu, Tudor Oprea, and suman sirimulla. Enhancing admet property models performance through combinatorial fusion analysis. *ChemRxiv*, 2023(1129), 2023.

[10] Kerstin Kläser, Błażej Banaszewski, Samuel Maddrell-Mander, Callum McLean, Luis Müller, Ali Parviz, Shenyang Huang, and Andrew Fitzgibbon. `MiniMol`: A parameter-efficient foundation model for molecular learning, 2024.

[11] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100string representation. *Machine Learning: Science and Technology*, 1(4):045024, oct 2020.

[12] Huong Van Le, Weibin Ren, Junhong Kim, Yukyung Yun, Young Bin Park, Young Jun Kim, Bok Kyung Han, Inho Choi, Jong IL Park, Hwi-Yeol Yun, and Jae-Mun Choi. Caliciboost: Performance-driven evaluation of molecular representations for caco-2 permeability prediction, 2025.

[13] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

[14] James H. Notwell and Michael W. Wood. Admet property prediction through combinations of molecular fingerprints, 2023.

[15] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[16] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019.

[17] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[18] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

[19] Eduardo Soares, Emilio Vital Brazil, Victor Yukio Shirasuna, Dmitry Zubarev, Renato Cerqueira, and Kristin Schmidt. Smi-ted: A large-scale foundation model for materials and chemistry. *The Thirteenth International Conference on Learning Representations (ICRL 2025)*, 2025. Under review.

[20] Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

[21] Gemma Turon, Jason Hlozek, John G. Woodland, Kelly Chibale, and Miquel Duran-Frigola. First fully-automated ai/ml virtual screening cascade implemented at a drug discovery centre in africa. *bioRxiv*, 2022.

[22] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[23] Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022.

[24] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *The Eleventh International Conference on Learning Representations*, 2023.

Supplementary Table 1: Comparison between ESM1b and ESM2.

Modalities		DG	DAVIS			KIBA			AVG PCC
Drug	Protein		Drug	Rand	Target	Drug	Rand	Target	
IMC-SMI	ESM1b	0.531	0.237	0.839	0.593	0.625	0.856	0.592	**0.611**
	ESM2	0.518	0.204	0.828	0.535	0.622	0.843	0.632	0.598
MF-IMC-MOLF	ESM1b	0.622	0.234	0.845	0.581	0.616	0.862	0.588	**0.621**
	ESM2	0.606	0.208	0.830	0.527	0.653	0.851	0.632	0.615
MF-IMC-SMI	ESM1b	0.610	0.200	0.848	0.580	0.653	0.867	0.631	**0.627**
	ESM2	0.598	0.123	0.842	0.530	0.646	0.855	0.634	0.604
MF-MOLF	ESM1b	0.625	0.186	0.838	0.548	0.654	0.855	0.609	**0.616**
	ESM2	0.620	0.192	0.832	0.522	0.655	0.852	0.615	0.613
MF-SMI	ESM1b	0.595	0.205	0.830	0.552	0.672	0.843	0.608	**0.615**
	ESM2	0.587	0.171	0.815	0.506	0.668	0.830	0.618	0.599
MOLF-IMC	ESM1b	0.576	0.246	0.842	0.602	0.637	0.864	0.591	**0.622**
	ESM2	0.553	0.261	0.834	0.558	0.599	0.856	0.594	0.608
SMI	ESM1b	0.504	0.143	0.703	0.584	0.598	0.796	0.606	0.562
	ESM2	0.505	0.163	0.724	0.551	0.595	0.802	0.620	**0.566**

Supplementary Table 2: Benchmark results for the preselected ImageMol variants. **Bold** and underlined entries denote the best and second best global results, respectively. In *italic* the best local score, i.e. the best score for single/multiple drug modality cases.

Modalities		DG	DAVIS			KIBA			AVG PCC
Drug	Protein		Drug	Rand	Target	Drug	Rand	Target	
IM	ESM1b	*0.5461*	0.2084	***0.8462***	0.5795	*0.6305*	*0.8618*	0.5863	0.6084
IMC	ESM1b	0.5310	***0.2810***	0.8381	***0.5934***	0.6092	0.8543	*0.5940*	*0.6144*
IM-MF	ESM1b	0.6152	0.2288	0.8399	0.5720	***0.6545***	0.8600	***0.6126***	0.6261
IMC-MF	ESM1b	***0.6251***	*0.2371*	*0.8433*	*0.5865*	0.6303	***0.8656***	0.6038	***0.6274***

Supplementary Table 3: Statistics of DTI benchmarks used.

Dataset	Train			Test		
	DTI Pairs	Drugs	Proteins	DTI Pairs	Drugs	Proteins
DG	183430	117,103	428	49,028	34,185	248
Davis Drug	17,813	47	379	14	68	379
Davis Target	18,020	68	265	5,168	68	76
Davis Random	18,041	68	379	5,154	68	379
Kiba Drug	82,539	1447	229	23,095	414	228
Kiba Target	82,326	2,068	160	23,196	46	229
Kiba Random	82,360	2,068	229	23,531	2021	228

Supplementary Table 4: Overview of ADMET benchmarks. Absorption measures how a drug travels from the site of administration to site of action. Drug distribution refers to how drug moves to and from the various tissues of the body and the amount of drugs in the tissues. Drug metabolism measures how specialized enzymatic systems breakdown the drugs and it determines the duration and intensity of a drug's action. Drug excretion is the removal of drugs from the body using various different routes of excretion, including urine, bile, sweat, saliva, tears, milk, and stool. Toxicity measures how much damage a drug could cause to organisms.

Class	Dataset	Size	Task	Metric	Dataset Split
Absorption	Caco2	906	Regression	MAE	Scaffold
	HIA	578	Binary	AUROC	Scaffold
	Pgp	1,212	Binary	AUROC	Scaffold
	Bioav	640	Binary	AUROC	Scaffold
	Lipo	4,200	Regression	MAE	Scaffold
	AqSol	9,982	Regression	MAE	Scaffold
Distribution	BBB	1,975	Binary	AUROC	Scaffold
	PPBR	1,797	Regression	MAE	Scaffold
	VDss	1,130	Regression	Spearman	Scaffold
Metabolism	CYP2C9 Inhib.	12,092	Binary	AUPRC	Scaffold
	CYP2D6 Inhib.	13,130	Binary	AUPRC	Scaffold
	CYP3A4 Inhib.	12,328	Binary	AUPRC	Scaffold
	CYP2C9 Subs.	666	Binary	AUPRC	Scaffold
	CYP2D6 Subs.	664	Binary	AUPRC	Scaffold
	CYP3A4 Subs.	667	Binary	AUROC	Scaffold
Excretion	Half Life	667	Regression	Spearman	Scaffold
	CL-Hepa	1,020	Regression	Spearman	Scaffold
	CL-Micro	1,102	Regression	Spearman	Scaffold
Toxicity	LD50	7,385	Regression	MAE	Scaffold
	hERG	648	Binary	AUROC	Scaffold
	Ames	7,255	Binary	AUROC	Scaffold
	DILI	475	Binary	AUROC	Scaffold

Supplementary Table 5: Impact of adding sequence-based modalities (SMILES/SELFIES) across different experimental configurations in the DG dataset using the FFNN-IF architecture. Columns correspond to different sequence-based models.

	Original	MOLF	SMI	CB	MG
MF	0.559	0.572	0.560	0.572	0.570
IMC – MF	0.554	0.553	0.547	0.551	0.551
MF – UM	0.542	0.565	0.558	0.568	0.570
IMC – MF – UM	0.548	0.559	0.557	0.557	0.554

Supplementary Table 6: SotA results compared to our experiments - DTI Group. [OK: Otter Knowledge, IF: Intermediate Fusion, LF: Late Fusion]

Dataset	Our results	Reported SotA
DG (PCC; Max is better)	**0.6400** For OK-IF with CB-IMC-UM	0.588 (0.002) Reported in [6]
Kiba Target (PCC; Max is better)	**0.693** For FFNN-LF with IMC-MF-MG-MOLF	No leaderboard.
Kiba Drug (PCC; Max is better)	**0.722** For OK-LF with IMC-MF-MOLF-SMI-UM	No leaderboard.
Kiba Random (PCC; Max is better)	**0.901** For OK-LF with CB-IMC-MOLF-UM	No leaderboard.
Davis Target (PCC; Max is better)	**0.661** For OK-LF with CB-IMC-MF-MOLF-SMI-UM	No leaderboard.
Davis Drug (PCC; Max is better)	**0.464** For FFNN-LF with MG-UM	No leaderboard.
Davis Random (PCC; Max is better)	**0.876** For OK-LF with IMC-MF-MOLF-UM	No leaderboard.

Supplementary Table 7: SotA results compared to our experiments - Absorption [OK: Otter Knowledge, IF: Intermediate Fusion, LF: Late Fusion]

Dataset	Our results	Reported SotA
Bioavailability ma (AUROC; Max is better)	0.72 For FFNN-LF with IMC-MOLF-UM	**0.942 (0.002)** Reported in [10]
Bbb martins (AUROC; Max is better)	0.92 For FFNN-LF with CB-IMC-MOLF-UM	**0.924 (0.003)** Reported in [10]
Caco2 wang (MAE; Min is better)	0.76 For FFNN-IF with MF	**0.256 (0.006)** Reported in [12]
Hia hou (AUROC; Max is better)	0.99 For FFNN-LF with MF-MG-MOLF	**0.993 (0.005)** Reported in [10]
Pgp broccatelli (AUROC; Max is better)	0.93 For FFNN-IF with MF-MOLF	**0.938 (0.002)** Reported in [14]

Supplementary Table 8: SotA results compared to our experiments - Metabolism [OK: Otter Knowledge, IF: Intermediate Fusion, LF: Late Fusion]

Dataset	Our results	Reported SotA
Cyp2c9 substrate carbonmangels (AUPRC; Max is better)	0.46 For FFNN-IF with MF-SMI	**0.474 (0.025)** Reported in [10]
Cyp2c9 veith (AUPRC; Max is better)	0.78 For FFNN-IF with MF-MOLF	**0.859 (0.001)** Reported in [14]
Cyp2d6 substrate carbonmangels (AUPRC; Max is better)	0.73 For FFNN-IF with MF	**0.736 (0.024)** Reported in [7]
Cyp2d6 veith (AUPRC; Max is better)	0.70 For FFNN-LF with IMC-MF-MG-MOLF-UM	**0.790 (0.001)** Reported in [14]
Cyp3a4 substrate carbonmangels (AUROC; Max is better)	0.75 For FFNN-LF with IMC-MF-SMI	**0.667 (0.019)** Reported in [9]
Cyp3a4 veith (AUPRC; Max is better)	0.87 For FFNN-LF with CB-IMC-MF-MOLF	**0.916 (0.000)** Reported in [14]

Supplementary Table 9: SotA results compared to our experiments - Excretion [OK: Otter Knowledge, IF: Intermediate Fusion, LF: Late Fusion]

Dataset	Our results	Reported SotA
Clearance hepatocyte az (SCC; Max is better)	0.50 For FFNN-IF with MF-MOLF-UM	**0.536 (0.020)** Reported in [9]
Clearance microsome az (SCC; Max is better)	0.62 For FFNN-IF with MF-MOLF-UM	**0.630 (0.010)** Reported in [14]
Half life obach (SCC; Max is better)	0.62 For FFNN-LF with MF-MOLF	**0.576 (0.025)** Reported in [9]

Supplementary Table 10: SotA results compared to our experiments - Physicochemical and distribution. [OK: Otter Knowledge, IF: Intermediate Fusion, LF: Late Fusion]

Dataset	Our results	Reported SotA
Lipophilicity astrazeneca (MAE; Min is better)	0.99 For FFNN-IF with SMI	**0.456 (0.008)** Reported in [10]
Ppbr az (MAE; Min is better)	16.70 For FFNN-IF with SMI	**7.526 (0.106)** Reported in [14]
Solubility aqsoldb (MAE; Min is better)	1.77 For FFNN-IF with SMI	**0.741 (0.013)** Reported in [10]
Vdss lombardo (SCC; Max is better)	0.69 For FFNN-LF with CB-MF-SMI	**0.713 (0.007)** Reported in [14]

Supplementary Table 11: SotA results compared to our experiments - Toxicity [OK: Otter Knowledge, IF: Intermediate Fusion, LF: Late Fusion]

Dataset	Our results	Reported SotA
Ames (AUROC; Max is better)	0.85 For FFNN-IF with MF-MOLF	**0.871 (0.002)** Reported in [21]
Dili (AUROC; Max is better)	0.94 For FFNN-LF with MF-MOLF-UM	**0.956 (0.006)** Reported in [10]
Herg (AUROC; Max is better)	0.87 For FFNN-IF with MF-MG-UM	**0.880 (0.002)** Reported in [14]
Ld50 zhu (MAE; Min is better)	0.75 For FFNN-IF with SMI	**0.552 (0.009)** Reported in [8]