Supplementary Information (SI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2026

CONFORFORMER: REPRESENTATION FOR MOLECULES THROUGH UNDERSTANDING OF CONFORMERS

Mas Pieter Klein[a], Irina Rudenko[b], Evgeny A. Pidko[a], Ivan Bushmarinov[c]

Supporting Information

[a] Department of Chemical Engineering, TU Delft, Delft, Netherlands
[b] Avride Inc., Austin, TX, USA
[c] Perplexity AI, Belgrade, Serbia

A Training details

A.1 General Remarks

All code is available at the GitHub repository `https://github.com/EPiCs-group/ConforFormer`. The model architecture throughout this whole document is that of the standard Uni-Mol up until contrastive learning is done. This can be found in Appendix C (Table 6) in Zhou *et al.*[1]. Set-up is kept identical, regardless of whether the Uni-Mol, OMol, or contrastive benchmark is used as training data. The pre-trained models parameters can be found in the following HuggingFace repository `https://huggingface.co/ConforFormer/ConforFormer` The settings for the Uni-Mol replication and ConforFormer models are listed in A.2 and A.3 respectively. A three-layer $512 \times 256 \times 128$ MLP was used for fine-tuning with exactly the same settings as in Zhou *et al.*[1]. See the Ablation Studies (section E in the Supporting Information) for details of other experiments.

A.2 Uni-Mol replication

Training was carried out using the Uni-Core framework (version 0.0.3).

Hyperparameters

- `masked_token_loss = 1`
- `masked_coord_loss = 5`
- `masked_dist_loss = 10`
- `x_norm_loss = 0.01`
- `delta_pair_repr_norm_loss = 0.01`
- `mask_prob = 0.15`
- `noise_type = "uniform"`
- `noise = 1.0`
- `only_polar = 0` (no hydrogens on the molecule)
- `dropout = 0.1` (applied to FFN, attention heads, etc.)
- Activation functions are always GeLU
- `batch size = 128`

Training details

- Linear learning rate schedule
- 10,000 warm-up steps
- 1,000,000 total steps
- Validation every 10,000 steps
- Adam optimizer
- $\epsilon = 1 \times 10^{-6}$
- $\beta = (0.9, 0.99)$
- Weight decay of 1×10^{-4}

A.3 ConforFormer models

All hyperparameters remained the same as in A.2 except for the learning rate and the learning rate scheduling:

- Learning rate schedule altered to `ReduceLRonPlateau`
  - Patience = 3
  - $\epsilon = 0.25$
  - lr-shrink = 0.5
- 5,000 warm-up steps
- Validation was done every 5000 steps, starting after warm-up
- Peak learning rate of 5×10^{-4}
- Batch size of 128

A.4 Reduced UniMol dataset

For quick iterations and experiments, we used a setup which could be trained to convergence overnight on a single NVIDIA H100 GPU. For that, we chose 1/8 of the Uni-Mol dataset by simply iterating over it and selecting every 8th data point:

```
for i, datapoint in enumerate(dataset):
    if i % 8 == 0:
        new_dataset.put(datapoint)
```

For these experiments, the hyperparameters remained the same as in A.2 except for the learning rate, learning rate schedule, and batch size, unless specified otherwise.

- Learning rate schedule altered to `ReduceLROnPlateau`
  - Patience = 2
  - $\epsilon = 0.25$
  - lr-shrink = 0.5
- 5,000 warm-up steps
- Validation was done every 5000 steps, starting after warm-up
- Peak learning rate of 5×10^{-4}
- batch size = 384

B Datasets and benchmarks

B.1 Uni-Mol

The training split of Uni-Mol[a], as detailed in Zhou *et al.*[1], consists of 18.8M unique molecules each with 10 conformations, resulting in ca. 190M datapoints. These 10 conformations were all generated using RDKit. On average, each datapoint has 27 heavy atoms. A large portion of the dataset consists of organic molecules. The dataset contains 67 unique heavy atoms, with C, O, and N making up greater than 95%. The remaining 5% consists almost exclusively of the halogens (F, Cl, Br, and I) along with P and S. Consequently, there are only 9 heavy atom types that have a share greater than 0.01%. The validation split consists of ≈100K unique molecules, again with 10 conformations each. Of the 21 unique heavy atoms, the same 9 atoms have a share greater than 0.01% of the dataset. After the reduction of the dataset, taking every eighth datapoint, the relative distribution of all heavy atoms remains the same.

B.2 OMol

The full Open Molecules (OMol) dataset[2] consists of various molecules which are relevant to homogenous catalysis, electrolytes, and biomolecular systems. Structures were calculated at the ωB97M-V/def2-TZVPD level of theory, resulting in unquestionably higher quality data than that found in Uni-Mol. In total there are 101M unique datapoints[b]. On average the molecules have 26 heavy atoms. C, O, and N make up 91% of all heavy atoms. OMol has a larger atom variety than Uni-Mol with 83 unique atom types, 59 having a share larger than 0.01% divided amongst various charge and spin states.

[a] Accessed on 2025/24/04
[b] Accessed on 2025/26/06

Reducing the dataset to that of only molecules with at least 2 heavy atoms and at least two conformations (the setup used to train ConforFormer–OMol) results in a dataset of 8.25M unique molecules, 55M total datapoints (further OMol-conf). On average, each molecule has 6-7 unique conformations. All 83 atom types remain in the dataset and now 64 have at least a 0.01% share of all heavy atoms.

B.3 MoleculeNet

MoleculeNet[2] is a collection of molecular benchmarks that contains tasks relevant to Physiology, Biophysics, Physical chemistry, or Quantum mechanics. While we acknowledge the documented limitations of the MoleculeNet benchmarks - for example, suboptimal dataset splitting, inconsistent molecular representations, and curation errors[3] — these datasets remain the most widely adopted standards in the field. Consequently, they provide an essential framework for benchmarking against a vast body of existing literature and established models, as well as the UniMol architecture we utilize. Classification tasks are always evaluated using ROC–AUC score, while regression tasks are either evaluated using RMSE or MAE. Generally, the benchmarks contain primarily organic molecules (compositions of C, N, O), with these atoms accounting for anywhere between 70% to almost 100% of all heavy atoms (all atoms excluding H) within a benchmark. For all tasks, SMILES strings are provided alongside the targets, with only a select few (QMx, $x \in \{7, 8, 9\}$, the quantum mechanics based benchmarks) having provided 3D coordinates. As such, the structures generated for every 3D coordinate are those already made by the team of UniMol. In the cases where 3D coordinates are provided, only one per molecule is present, and 9 more were generated. The benchmarks are used as provided by the UniMol team [c].

B.3.1 BACE

BACE is a classification benchmark with 1 target and contains slightly over 1500 molecules. The target is a binary label which qualitatively describes a molecule's ability to inhibit the human beta-secretase 1 (BACE-1). The molecules within the benchmark are purely organic, containing on average 34 heavy atoms, primarily C, N, O, and S. The halogens account for slightly over 1% of heavy atoms.

B.3.2 BBBP

BBBP is a classification benchmark with labels indicating whether a molecule can or cannot penetrate the blood-brain barrier. The $\approx$ 2000 molecules are primarily organics and occasionally halogenated. There are small amounts of salts, specifically alkai (earth) metals with 21 Na atoms and 1 Ca atom. On average, they contain 24 heavy atoms.

[c] Accessed on 2025/25/04

B.3.3 Clintox

Clintox is a classification task, describing drug-like molecules using qualitative data of those approved by the FDA or failed due to toxicity. It contains $\approx$ 1500 molecules, primarily of organic nature, with 26 heavy atoms. There are small amounts of main-group and d-block atoms. Curiously, all of this atomic variety is found in either the train or test splits.

B.3.4 ESOL

ESOL is a benchmark of 1100 small ($\approx$ 13 heavy atoms) organic molecules, occasionally halogenated (F, Cl, Br and I accounting for $\approx$ 5% of heavy atoms, primarily Cl). The target is the log solubility of a molecule in water (in mol/L). It is evaluated using RMSE. Disproportionally few heavier halogens (Br and I) are in the test benchmark, specifically none.

B.3.5 FreeSolv

FreeSolv is a benchmark of small organic molecules and their experimental or calculated solvation energy in water (in kcal/mol). Performance is evaluated using RMSE. It contains 642 small organic molecules (on average 9 heavy atoms), occasionally halogenated.

B.3.6 HIV

HIV is a classification benchmark asking a model to distinguish between molecules which do or do not inhibit HIV replication. It is primarily organic molecules (C, O, N accounting for 95% of all heavy atoms), however still contains a large variety of alkali (earth) metals, d-block metals (frequently containing almost all occurrences of a d-block metal through the whole of MoleculeNet), and main-group elements. On average, the 41K datapoints contain approx. 25 heavy atoms.

B.3.7 Lipo

Lipo is a regression benchmark requiring a model to predict experimentally determined log(P) (octanol/water partition coefficient) at a pH of 7.4. It 4,200 organic molecules with on average 27 heavy atoms. The performance of a model is evaluated using RMSE.

B.3.8 MUV

MUV is a benchmark of 93K organic molecules of 24 heavy atoms on average. It contains the classification task of 17 targets, determined through high-through put experiments on BioAssays. It is a subset of datapoints contained in PCBA, refined through nearest-neighbor analysis and is meant to validated virtual screening methods.

B.3.9 PCBA

PCBA is a benchmark of selected PubChem BioAssay consisting of results of high-throughput experiments on the biological activity of small molecules. It has 128 targets. On average, each

datapoint has 26 heavy atoms, primarily C, N ,O (accounting for 95% of all heavy atoms in the benchmark), and various main-group and d-block elements. It contains 438K datapoints. The fine-tuning result of ConforFormer–OMol on this benchmark was 0.829 (on par with literature data) but due to its size we did not run it for most of the models in the study and do not include it in the tables.

B.3.10 QM7, QM8, QM9

The QMx benchmarks consist of small organic (QM8 and QM9 occasionally halogenated) molecules with 7, 8, or 9 heavy atoms. They are regression benchmarks, performance measured in MAE, with the aim to predict various quantum-mechanically properties, such as atomization energy and HOMO-LUMO gap. They contain approx. 7K, 22K, and 130K datapoints, respectively.

B.3.11 Sider

SIDER is a classification benchmark with 27 targets, aiming to predict if a marketed drug has adverse drug reactions to 27 system organ classes. It consists of generally large molecules (average of 33 heavy atoms), primarily of organic nature. However, it does contain a variety of main-group and d-block elements. This atomic variety finds itself almost exclusively in the training split, frequently over 95% of these atoms. It contains 1427 datapoints.

B.3.12 Tox21

Tox21 is a classification benchmark of 7831 datapoints and has 12 targets. These targets are binary labels as to whether a molecule has any qualitative toxicity on 12 biological systems. Datapoints contain 18.5 heavy atoms on average, primarily organic in nature. There are also small amount of main-group and d-block elements, primarily found within the training data ($> 90\%$ occurrence in training split).

B.3.13 Toxcast

Toxcast has 617 binary classification targets representing qualitative toxicology data generated using in-vitro high-throughput experimentation. The ca. 8600 datapoints usually contain 19 heavy atoms, primarily C, O, N, and halogens. Main-group and d-block elements make up about 1% of all heavy atoms.

C PharmaIsomer dataset

To construct this benchmark, we used a portion of ZINC20[4] [d]. This database consists of commercially available medicinal molecules for virtual screening. For our purposes, it provides a large selection of relevant and valid molecular structrues from which to generate molecular conformations.

[d] Accessed on 2025/14/07

From the ZINC20 dataset, we took mildly reactive, relatively easy to purchase molecules, which in the database are marked as having reactivity up to and including "standard" and purchase up to and including "wait OK." We removed all overlap with Uni-Mol and OMol datasets using SMILES-identified unique formulas and also removed molecules that do not have a chemical formula. Ten conformers of each SMILES string were generated using ETKDGv1,[5] implemented in rdKit, and optimized with the MMFF94 force field.[6] To filter conformations, a root-mean-squared threshold of 0.5 was used. Only structurally distinct conformers were used and special attention was put to avoid duplicate conformers.

To ensure that only isomeric structures are assessed for similarity, simplify inference, and make metrics between runs comparable, batches were statically constructed beforehand rather than dynamically produced at inference time. Specifically, each batch contained 128 unique molecules, all isomers to each other. Each isomer had exactly 2 conformers, resulting in 256 datapoints per batch. An 80/10/10 train/test/validation split was employed so that the performance of models trained on this dataset could be evaluated; metrics in the main text are reported on the validation split.

The dataset contains 3,261,807,960 datapoints in 12,741,440 batches and is freely available under a CC-BY license[7].

D LLM usage

Large language models were used for grammar checking, LaTeX formatting, initial literature search, generation of the SQLite processing code used to support isomer classification analysis, and generation of code for the CatBoost/FP4 and XGBoost/ECFP4 (1024-bit) baselines. The initial literature search was performed with OpenAI o1 (Deep Research), code generation and LaTeX editing with OpenAI GPT 5.1 (initial draft), 5.2 (second draft) and 5.4 (final draft). The edits to the paper text performed by LLMs were limited to LaTeX syntax corrections and table formatting.

E Ablation studies

E.1 Overview

The table S3 contains details of the specific experiments we ran, with the following table S4 containing benchmark results for various pre-training and post-training setups with different number of unfrozen layers (15 corresponding to a fully unfrozen model). The"ConforFormer" objective refers to the loss $\mathcal{L}_{total}$ as described in 3.2. Abbreviated names for experiments are used throughout these tables, with the following mapping to the main text entries:

- **U**: Uni-Mol replicate
- **U-no-flat**: Uni-Mol no "flat"
- **O-c**: Uni-Mol, OMol data

- **CF-O-c**: ConforFormer–OMol
- **CF-U**: ConforFormer–UniMol
- **Random-w**: Uni-Mol no pretrain

An important note from the ablation study is that the addition of contrastive learning in-post provides no improvement to the model. In fact, it actually worsens performance. Comparing the Cpost runs in Table S4, an increase in temperature (τ) is seen to worsen finetuning results. Taking the results of BBBP as an example, the ROC–AUC score changes as $0.656 \rightarrow 0.562 \rightarrow 0.544$ as τ goes from 0.01 to 0.1 and finally 0.5. We used these results to guide our choice of τ for ConforFormer, but we checked that increasing the temperature of the NT-Xent loss in pre-training worsens the performacne as well (see **CF-U-r-0.25**). As the model is made to differentiate more, it performs worse. Compared to the results of "ConforFormer–OMol", the best performing model with contrastive pose-training results in worse metrics after finetuning.

A contrastive-only pre-training on the Unimol dataset (**Contrast-U-r**), on the other hand, results in a below average quality of the embeddings but still performs surprisingly well, suggesting that contrastive objective alone could also be potentially viable strategy for model pre-training.

Sanity checks were run to validate that the ConforFormer loss and not changes in the training setup were actually driving the metric improvements. However, changing the batch size (**U-r-128**) or adding more conformers in each batch (**U-r-256-conf**) did not lead to any noticeable improvements. Training the Uni-Mol model without additional contrastive loss on the full OpenMolecules dataset (**O**) did not improve benchmarks beyond ConforFormer–OMol **CF-O-c** or Uni-Mol replicate **U** either.

Given the results of models trained using contrastive learning on PharmaIsomer, it was hypothesized that doing training that mimics this benchmark would improve performance. For this, the Uni-Mol data set was tailored to allow for the construction of batches that guarantee a minimum number of isomers. The results of these pre-training can be found in the **CF-U-**x**I** setups, $x \in \{10, 25, 40\}$ in Table S4 . These 3 rows represent a minimum of 40%, 25%, and 10% of datapoints within a batch having an isomer pair. No noticeable improvement is observed within the classification tasks and no significant trend is observed for the regression tasks. Thus, the hypothesis of additional isomers in the batch is not seen to be true for the Uni-Mol dataset with the contrastive training set-up outlined in Section A.3.

Each line in Table S4 corresponds to a single fine-tuning run. For the stability-analysis context, refer to the transfer-learning discussion in Section 2.3 of the main text.

As a final ablation, the number of conformers used during finetuning was reduced from 10 to 1, 2, or 5. The results are shown in Tables S5 and S6. Regardless of setup,

E.2 Full run data

The labels for pre-training runs from Table S3 are used throughout Table S4, which contains fine-tuning results with different number of model layers unfrozen. Layers were frozen starting from the last. Finetuning was performed using the hyperparameters specified in the Uni-Mol GitHub repository `https://github.com/deepmodeling/Uni-Mol/tree/main/unimol`. A typical fine-tuning job took $\approx$ 3 hours on an A100 GPU. For the fine-tuning starting from **Random-w**, all batch sizes were set to 128, and training was stopped after no improvement was made after 40 epochs (18 hours on an H100). The remaining hyperparameters are identical to those under standard finetunings.

Table S1: Performance on quantum-chemical and physicochemical regression benchmarks. Values denote root-mean-square deviation (RMSD), with lower values indicating better agreement with reference data. Literature results are reproduced from the Uni-Mol benchmark study. Results from the current work are grouped into unfrozen models, where the encoder is fine-tuned for each task, frozen models, where representations from a fixed pre-trained encoder are used without backbone fine-tuning, and 2D fingerprint baselines. Dataset sizes (N) indicate the number of molecules included in each benchmark. See Section B for dataset descriptions.

Model	**ESOL**	**FreeSolv**	**Lipo**	**QM7**	**QM8**	**QM9**
N points	1128	642	4200	6830	21786	133885
Literature data						
D-MPNN	1.05(1)	2.08(8)	0.683(16)	103.5(86)	0.0190(1)	0.00814(1)
Attentive FP	0.88(3)	2.07(18)	0.721(1)	72.0(27)	0.0179(10)	0.00812(1)
N-GramRF	1.07(11)	2.69(8)	0.812(28)	92.8(40)	0.0236(6)	0.01037(16)
N-GramXGB	1.08(8)	5.06(74)	2.072(30)	81.9(19)	0.0215(5)	0.00964(31)
PretrainGNN	1.10(1)	2.76(0)	0.739(3)	113.2(6)	0.0200(1)	0.00922(4)
$\text{GROVER}_{\text{base}}$	0.98(9)	2.18(5)	0.817(8)	94.5(38)	0.0218(4)	0.00984(55)
$\text{GROVER}_{\text{large}}$	0.90(2)	2.27(5)	0.823(10)	92.0(9)	0.0224(3)	0.00986(25)
GraphMVP	1.03(3)	–	0.681(10)	–	–	–
MolCLR	1.27(4)	2.59(25)	0.691(4)	66.8(23)	0.0178(3)	–
GEM	0.80(3)	1.88(9)	0.660(8)	58.9(8)	0.0171(1)	0.00746(1)
Uni-Mol	0.79(3)	1.48(5)	0.603(10)	41.8(2)	0.0156(1)	0.00467(4)
Current work						
Unfrozen models						
Uni-Mol replicate	0.83(3)	1.80(11)	0.608(9)	58.8(30)	0.0160(1)	0.00520(0)
ConforFormer–OMol	0.91(2)	1.99(5)	0.642(11)	53.8(18)	0.0159(0)	0.00542(4)
Uni-Mol no pretrain	0.98(5)	2.49(23)	0.787(22)	83.6(156)	0.0186(6)	0.00618(7)
Frozen models						
Uni-Mol replicate	1.15(3)	**2.64(6)**	0.916(4)	**82.6(44)**	0.0264(5)	0.0184(12)
Uni-Mol no "flat"	1.23(3)	2.92(4)	0.935(4)	88.5(32)	0.0263(2)	0.01910(19)
Uni-Mol, OMol data	1.18(1)	3.00(6)	0.949(6)	89.9(57)	0.0274(2)	0.0202(3)
ConforFormer–UniMol	1.17(3)	3.38(5)	0.807(5)	104.9(90)	0.0223(2)	0.01258(17)
ConforFormer–OMol	**1.12(2)**	3.53(7)	**0.752(7)**	99.9(112)	**0.0219(2)**	**0.01172(31)**
2D baselines						
XGBoost ECFP4 (1024-bit) baseline	1.562(19)	3.967(24)	0.864(5)	155.5(33)	0.02341(5)	0.01258(3)
CatBoost FP4 baseline	1.786(64)	3.400(32)	1.045(8)	129.2(58)	0.02632(18)	0.01704(62)

Table S2: Performance on biological activity classification benchmarks. Values report the mean ROC–AUC over multiple random splits (standard deviation in parentheses), with higher values indicating better performance. Literature results are taken from the original Uni-Mol study for reference. Results from the current work are reported separately for unfrozen models, where the encoder is fine-tuned together with the task-specific prediction head, frozen models, where molecular representations are extracted from a fixed pre-trained encoder and only lightweight task-specific models are trained, and 2D fingerprint baselines. Dataset sizes (N) correspond to the number of labeled molecules in each benchmark. See Section B for dataset descriptions.

Model	**BBBP**	**BACE**	**ClinTox**	**Tox21**	**ToxCast**	**SIDER**	**HIV**	**MUV**
N points	2039	1513	1478	7831	8575	1427	41127	93087
Literature data								
D-MPNN	0.710(3)	0.809(6)	0.906(6)	0.759(7)	0.655(3)	0.570(7)	0.771(5)	0.786(14)
Attentive FP	0.643(18)	0.784(0)	0.847(3)	0.761(5)	0.637(2)	0.606(32)	0.757(14)	0.766(15)
N-GramRF	0.697(6)	0.779(15)	0.775(40)	0.743(4)	–	0.668(7)	0.772(1)	0.769(7)
N-GramXGB	0.691(8)	0.791(13)	0.875(27)	0.758(9)	–	0.655(7)	0.787(4)	0.748(2)
PretrainGNN	0.687(13)	0.845(7)	0.726(15)	0.781(6)	0.657(6)	0.627(8)	0.799(7)	0.813(21)
$\text{GROVER}_{\text{base}}$	0.700(1)	0.826(7)	0.812(30)	0.743(1)	0.654(4)	0.648(6)	0.625(9)	0.673(18)
$\text{GROVER}_{\text{large}}$	0.695(1)	0.810(14)	0.762(37)	0.735(1)	0.653(5)	0.654(1)	0.682(11)	0.673(18)
GraphMVP	0.724(16)	0.812(9)	0.791(28)	0.759(5)	0.631(4)	0.639(12)	0.770(12)	0.777(6)
MolCLR	0.722(21)	0.824(9)	0.912(35)	0.750(2)	–	0.589(14)	0.781(5)	0.796(19)
GEM	0.724(4)	0.856(11)	0.901(13)	0.781(1)	0.692(4)	0.672(4)	0.806(9)	0.817(5)
Uni-Mol	0.729(6)	0.857(2)	0.919(18)	0.796(5)	0.696(1)	0.659(13)	0.808(3)	0.821(13)
Current work								
Unfrozen models								
Uni-Mol replicate	0.705(26)	0.832(25)	0.857(22)	0.788(3)	0.685(7)	0.644(14)	0.784(8)	0.784(9)
ConforFormer–OMol	0.691(21)	0.820(29)	0.686(26)	0.787(4)	0.689(7)	0.634(13)	0.786(4)	0.758(30)
Uni-Mol, no pretrain	0.655(11)	0.775(39)	0.639(46)	0.735(15)	0.635(12)	0.607(18)	0.739(18)	0.616(11)
Frozen models								
Uni-Mol replicate	0.640(4)	0.775(2)	**0.767(10)**	0.709(4)	**0.653(2)**	0.606(13)	0.734(3)	0.755(10)
Uni-Mol no "flat"	0.651(1)	0.778(1)	0.725(11)	0.711(1)	0.637(1)	0.606(6)	0.746(2)	0.742(8)
Uni-Mol, OMol data	0.664(4)	**0.783(4)**	0.698(11)	0.710(1)	0.629(1)	0.610(6)	**0.756(5)**	0.695(7)
ConforFormer–UniMol	0.665(7)	0.731(55)	0.533(14)	0.753(1)	0.644(3)	**0.649(2)**	0.711(4)	0.716(5)
ConforFormer–OMol	**0.673(6)**	0.751(13)	0.716(9)	**0.755(1)**	0.638(3)	0.640(6)	0.751(4)	**0.774(6)**
2D baselines								
XGBoost ECFP4 (1024-bit)	0.682(8)	0.805(35)	0.839(21)	0.724(4)	0.589(2)	0.618(11)	0.762(14)	0.774(26)
CatBoost FP4	0.675(6)	0.800(8)	0.816(53)	0.682(7)	0.577(5)	0.616(8)	0.720(16)	0.641(31)

Table S3: Training setups used in the experiments. Batch size refers to the pre-training procedure; when written as n_u (n_t), it denotes n_t geometries total in the batch, with n_u used to compute the Uni-Mol losses $\mathcal{L}_{token}$, $\mathcal{L}_{coord}$, and $\mathcal{L}_{distance}$. PharmaIsomer post-training batch size, where applicable, is always 256.

Training	Dataset	GPU	Hours	Objective	Batch size
CF-2U-r	1/8 Uni-Mol	H100	14	ConforFormer + Uni-Mol loss on full batch, $\tau = 0.07$	256
CF-O-c	OMol-conf	4×H100	24	ConforFormer, $\tau = 0.07$	128 (256)
CF-U	Uni-Mol	H100	48	ConforFormer, $\tau = 0.07$	128 (256)
CF-U-10I	Uni-Mol, 10% isomers	H100	48	ConforFormer, $\tau = 0.07$	128 (256)
CF-U-25I	Uni-Mol, 25% isomers	H100	48	ConforFormer, $\tau = 0.07$	128 (256)
CF-U-40I	Uni-Mol, 40% isomers	H100	48	ConforFormer, $\tau = 0.07$	128 (256)
CF-U-H	Uni-Mol with H	A100	72	ConforFormer, $\tau = 0.07$	128 (256)
CF-U-r	1/8 Uni-Mol	A100	14	ConforFormer, $\tau = 0.07$	128 (256)
CF-U-r-0.25	1/8 Uni-Mol	A100	14	ConforFormer, $\tau = 0.25$	128 (256)
Contrast-U-r	1/8 Uni-Mol	A100	10	Contrast loss only, $\tau = 0.07$	256
O	OMol	A100	120	Uni-Mol	384
O-c	OMol-conf	A100	120	Uni-Mol	128
O-Cpost-0.01	OMol	A100	120	Uni-Mol + contrast on PharmaIsomer in post-train, $\tau = 0.01$	384
O-Cpost-0.05	OMol	A100	120	Uni-Mol + contrast on PharmaIsomer in post-train, $\tau = 0.05$	384
Random-w	–	A100	–	Random weight initialization	–
U	Uni-Mol	A100	72	Uni-Mol	128
U-no-flat	Uni-Mol, no "flat"	A100	72	Uni-Mol	128
U-r	1/8 Uni-Mol	A100	10	Uni-Mol	384
U-r-128	1/8 Uni-Mol	A100	14	Uni-Mol	128
U-r-256	1/8 Uni-Mol	H100	14	Uni-Mol	256
U-r-256-conf	1/8 Uni-Mol, 2 conformers each	H100	14	Uni-Mol	256
U-r-Cpost-0.01	1/8 Uni-Mol	H100	14	Uni-Mol + contrast on PharmaIsomer in post-train, $\tau = 0.01$	384
U-r-Cpost-0.1	1/8 Uni-Mol	H100	14	Uni-Mol + contrast on PharmaIsomer in post-train, $\tau = 0.1$	384
U-r-Cpost-0.5	1/8 Uni-Mol	H100	14	Uni-Mol + contrast on PharmaIsomer in post-train, $\tau = 0.5$	384

Table S4: Benchmark results on MoleculeNet as detailed in B.3. "Unfrozen" refers to the number of model layers unfrozen diring the fine-tuning procedure. Left block: classification benchmarks. Right block: regression benchmarks.

Training	Unfrozen	ROC-AUC, ↑								RMSE, ↓					
		BBBP	BACE	ClinTox	Tox21	ToxCast	SIDER	HIV	MUV	ESol	FreeSolv	Lipo	QM7	QM8	QM9
CF-2U-r	0	0.671	0.710	0.589	0.715	0.603	0.646	0.678	0.771	1.24	3.52	0.94	88.2	0.0236	0.0143
CF-O-c	0	0.673	0.763	0.724	0.753	0.638	0.643	0.724	0.739	1.04	3.67	0.75	163.0	0.0223	0.0123
CF-U	0	0.626	0.785	0.596	0.736	0.617	0.623	0.603	0.747	1.06	3.66	0.87	93.4	0.0243	0.0140
CF-U-10I	0	0.657	0.754	0.586	0.732	0.620	0.603	0.738	0.744	1.34	3.39	0.91	105.4	0.0262	0.0163
CF-U-25I	0	0.646	0.752	0.651	0.709	0.612	0.593	0.702	0.746	1.48	3.87	0.88	98.7	0.0250	0.0152
CF-U-40I	0	0.660	0.732	0.518	0.694	0.606	0.596	0.685	0.780	1.51	4.24	0.91	103.5	0.0254	0.0170
CF-U-H	0	0.641	0.799	0.430	0.721	0.626	0.638	0.707	0.709	1.18	3.60	0.84	119.3	0.0248	0.0140
CF-U-r	0	0.658	0.760	0.530	0.752	0.645	0.650	0.707	0.714	1.13	3.36	0.80	97.3	0.0227	0.0129
CF-U-r-0.25	0	0.583	0.657	0.685	0.674	0.564	0.608	0.722	0.644	1.57	3.93	1.00	97.3	0.0296	0.0242
Contrast-U-r	0	0.635	0.751	0.566	0.688	0.590	0.589	0.695	0.684	1.35	4.02	0.98	117.2	0.0259	0.0171
O	0	0.655	0.775	0.630	0.693	0.639	0.583	0.723	0.742	1.17	2.95	1.03	109.5	0.0309	0.0257
O-c	0	0.659	0.787	0.680	0.710	0.629	0.613	0.758	0.695	1.18	3.01	0.95	85.6	0.0278	0.0200
O-Cpost-0.01	0	0.672	0.722	0.562	0.679		0.609	0.666	0.573	1.52	4.29	1.03	91.4	0.0270	0.0191
O-Cpost-0.05	0	0.613	0.783	0.768	0.682	0.574	0.548	0.576	0.603	1.71	4.30	1.06	96.9	0.0280	0.0221
U	0	0.633	0.778	0.753	0.717	0.652	0.583	0.732	0.738	1.16	2.59	0.92	89.3	0.0274	0.0208
U-no-flat	0	0.650	0.778	0.727	0.712	0.636	0.598	0.744	0.734	1.27	2.89	0.93	91.4	0.0264	0.0191
U-r	0		0.728	0.742	0.671	0.626		0.757	0.603	1.19	2.86	1.03	113.9	0.0297	0.0241
U-r-128	0	0.673	0.777	0.746	0.655		0.582	0.743	0.534	1.51	3.19	1.04	127.1	0.0315	0.0266

Continued on next page

Training	Unfrozen	ROC-AUC, ↑								RMSE, ↓					
		BBBP	BACE	ClinTox	Tox21	ToxCast	SIDER	HIV	MUV	ESol	FreeSolv	Lipo	QM7	QM8	QM9
U-r-256	0	0.654	0.771	0.626	0.677	0.612	0.602	0.734	0.641	1.29	2.75	1.02	143.8	0.0297	0.0255
U-r-256-conf	0	0.654	0.764	0.626	0.677	0.612	0.607	0.734	0.641	1.33	2.75	1.04	116.1	0.0287	0.0232
U-r-Cpost-0.01	0	0.656	0.730	0.438	0.692	0.601		0.660	0.626	1.24	3.37	0.99	81.1	0.0273	0.0207
U-r-Cpost-0.1	0	0.562	0.679	0.574	0.591	0.522		0.559	0.539	1.89	4.00	1.12	126.5	0.0318	0.0274
U-r-Cpost-0.5	0	0.544	0.590	0.500	0.603	0.540	0.540	0.518	0.640	1.99	4.39	1.11	126.4	0.0320	0.0283
CF-O-c	1	0.676	0.811	0.682	0.774	0.667	0.655	0.773	0.824	1.08	2.38	0.67	96.0	0.0188	0.0083
CF-U	1	0.659	0.802	0.650	0.773	0.662	0.638	0.763	0.793	0.95	2.92	0.75	101.9	0.0216	0.0100
CF-U-H	1	0.641	0.828	0.727	0.751	0.663	0.652	0.766	0.805	1.00	2.62	0.73	98.9	0.0217	0.0101
CF-U-r	1	0.683	0.805	0.656	0.775	0.682	0.590	0.757	0.775	1.05	2.16	0.74	88.1	0.0187	0.0082
CF-U-r-0.25	1	0.643	0.673	0.641	0.718	0.624	0.624	0.766	0.698	1.31	2.63	0.85	78.9	0.0188	0.0091
O	1	0.651	0.799	0.596	0.751	0.654	0.671	0.736	0.730	1.10	2.64	0.80	105.9	0.0232	0.0131
O-Cpost-0.01	1	0.700		0.672	0.685	0.618	0.615	0.668	0.609	1.34	3.94	0.97	79.8	0.0251	0.0137
O-Cpost-0.05	1	0.614	0.831	0.722	0.722	0.598	0.559	0.598	0.622		4.24	1.02	91.8	0.0251	0.0145
U	1	0.727	0.779	0.826	0.780	0.695	0.632	0.780	0.813	0.83	1.88	0.65	85.5	0.0181	0.0078
U-no-flat	1	0.723	0.788	0.889	0.789	0.683	0.635	0.790	0.787	0.86	2.12	0.65	62.8	0.0174	0.0072
U-r	1		0.747	0.708	0.755	0.665		0.766	0.704	1.00	2.90	0.77	113.3	0.0219	0.0118
U-r-Cpost-0.01	1	0.674	0.787	0.543	0.735	0.630	0.614	0.667		1.06	2.84		77.3	0.0235	0.0119
U-r-Cpost-0.1	1	0.650	0.786	0.584	0.650	0.594	0.577	0.568	0.497	1.20	3.55		85.8	0.0266	0.0163
U-r-Cpost-0.5	1	0.650	0.786	0.584	0.710	0.602	0.596	0.568	0.497	1.20	3.55		96.0	0.0254	0.0148
CF-O-c	3	0.702	0.816	0.670	0.779	0.668	0.654	0.784	0.791	1.02	2.21	0.67	72.7	0.0180	0.0067
CF-U	3	0.665	0.803	0.671	0.777	0.668	0.641	0.766	0.795	0.94	2.59	0.74	101.2	0.0192	0.0079
CF-U-H	3	0.639	0.829	0.723	0.760	0.672	0.647	0.762	0.790	0.96	2.40	0.70	90.5	0.0191	0.0079
CF-U-r	3	0.689	0.821	0.646	0.777	0.684	0.630	0.769	0.760	0.95	2.25	0.72	78.7	0.0177	0.0069

Continued on next page

Training	Unfrozen	ROC-AUC, ↑								RMSE, ↓					
		BBBP	BACE	ClinTox	Tox21	ToxCast	SIDER	HIV	MUV	ESol	FreeSolv	Lipo	QM7	QM8	QM9
CF-U-r-0.25	3	0.680	0.740	0.643	0.724	0.646	0.647	0.762	0.675	1.13	2.20	0.77	69.7	0.0176	0.0068
O	3	0.728	0.837	0.630	0.751	0.654	0.671	0.745	0.797	1.11	2.25	0.71	82.4	0.0200	0.0087
O-Cpost-0.01	3		0.752	0.681	0.713	0.622	0.604	0.687	0.663	1.20	3.79	0.93	64.4	0.0225	0.0098
O-Cpost-0.05	3	0.653	0.825	0.775	0.720	0.609	0.566	0.664	0.679	1.23	4.13	0.97	68.5	0.0263	0.0102
U	3	0.717	0.785	0.842	0.782	0.692	0.636	0.791	0.807	0.88	1.63	0.61	74.3	0.0165	0.0060
U-no-flat	3	0.719	0.791	0.902	0.785	0.690	0.633	0.794	0.798	0.91	1.84	0.62	56.7	0.0168	0.0057
U-r	3	0.698	0.832	0.749	0.779	0.676	0.638	0.778	0.795	0.88	2.30	0.69	90.6	0.0185	0.0080
U-r-Cpost-0.01	3	0.696	0.785	0.579		0.629	0.624	0.693		0.98	2.42		80.8	0.0205	0.0088
U-r-Cpost-0.1	3	0.671	0.785		0.683	0.608	0.584	0.613	0.582	1.10	3.09	1.02	78.0	0.0234	0.0105
U-r-Cpost-0.5	3	0.661	0.661	0.503	0.729	0.605	0.595	0.578	0.602	1.08	3.74	0.93	93.6	0.0231	0.0100
CF-O-c	15	0.650	0.807	0.716	0.778	0.691	0.649	0.782	0.772	0.94	2.15	0.64	60.3	0.0159	0.0054
CF-U-r	15	0.699	0.815	0.745				0.789	0.819	0.88	1.89	0.66	59.4	0.0168	
O	15	0.680	0.859	0.704	0.785	0.680	0.618	0.785	0.711	0.87	1.95	0.66	52.0	0.0159	0.0054
Random-w	15	0.639	0.823	0.592	0.746	0.648	0.617	0.761	0.619	0.98	2.83	0.77	97.3	0.0178	0.0063
U	15	0.723	0.811	0.861	0.786	0.684	0.646	0.771	0.789	0.81	1.93	0.62	59.2	0.0161	0.0052
U-no-flat	15	0.694	0.796	0.883	0.797	0.688	0.648	0.800	0.883	0.74	3.08	0.38	64.2	0.0160	0.0053
U-r	15	0.725	0.825	0.854	0.782	0.677	0.637	0.780	0.792	0.85	2.15	0.64	58.1	0.0166	0.0060

Table S5: Performance on quantum-chemical and physicochemical regression benchmarks when reducing the number of conformations during finetuning from 10 to 1, 2, or 5. Values denote root-mean-square deviation (RMSD), with lower values indicating better agreement with reference data. Standard deviation in parentheses.

Model	Unfrozen	num. conf.	ESOL	Lipo	QM7	QM8	QM9
CF-O-c	0	1	1.14(3)	0.746(4)	105.9(136)	0.0218(1)	0.01160(7)
CF-O-c	0	2	1.14(3)	0.746(4)	105.9(136)	0.0218(1)	0.01160(7)
CF-O-c	0	5	1.12(2)	0.743(3)	98.6 (115)	0.0218(1)	0.01158(11)
U	0	1	1.15(2)	0.925(8)	80.8 (35)	0.0262(4)	0.01776(11)
U	0	2	1.15(2)	0.925(8)	80.8 (35)	0.0262(4)	0.01776(11)
U	0	5	1.13(2)	0.919(6)	80.8 (39)	0.0262(4)	0.01776(5)
U	15	1	0.84(4)	0.609(13)	56.2 (10)	0.0159(5)	0.00524(5)
U	15	2	0.84(4)	0.609(13)	56.2 (10)	0.0159(5)	0.00524(5)
U	15	5	0.85(3)	0.606(11)	57.8 (28)	0.0159(4)	0.00522(4)

Table S6: Performance on biological activity classification benchmarks when reducing the number of conformations during finetuning from 10 to 1, 2, or 5. Values report the mean ROC–AUC over multiple random splits (standard deviation in parentheses), with higher values indicating better performance.

Model	Unfrozen	num. conf.	BBBP	BACE	ClinTox	Tox21	ToxCast	SIDER	HIV	MUV
CF-O-c	0	1	0.662(9)	0.744(22)	0.711(9)	0.755(1)	0.638(2)	0.636(2)	0.749(3)	0.778(4)
CF-O-c	0	2	0.662(9)	0.744(22)	0.712(9)	0.755(1)	0.638(2)	0.636(1)	0.749(3)	0.778(4)
CF-O-c	0	5	0.666(6)	0.744(13)	0.712(10)	0.755(1)	0.638(2)	0.636(2)	0.754(3)	0.776(5)
U	0	1	0.639(5)	0.775(4)	0.786(8)	0.705(1)	0.652(5)	0.608(3)	0.735(4)	0.748(10)
U	0	2	0.639(5)	0.775(4)	0.786(7)	0.705(1)	0.652(5)	0.608(3)	0.735(4)	0.748(10)
U	0	5	0.640(4)	0.777(4)	0.788(8)	0.708(1)	0.652(3)	0.609(5)	0.735(3)	0.763(10)
U	15	1	0.705(17)	0.812(12)	0.819(31)	0.777(10)	0.677(10)	0.629(13)	0.787(7)	0.792(29)
U	15	2	0.705(18)	0.812(12)	0.865(36)	0.777(10)	0.677(10)	0.636(11)	0.781(12)	0.772(61)
U	15	5	0.700(23)	0.799(53)	0.878(31)	0.786(7)	0.688(6)	0.643(23)	0.790(10)	0.799(26)

F More visual examples of model performance

F.1 Isomers

Figure S1: A pair of isomers (distinct molecules) having similarity of 0.99 in the Uni-Mol embedding space and 0.04 in ConforFormer–OMol

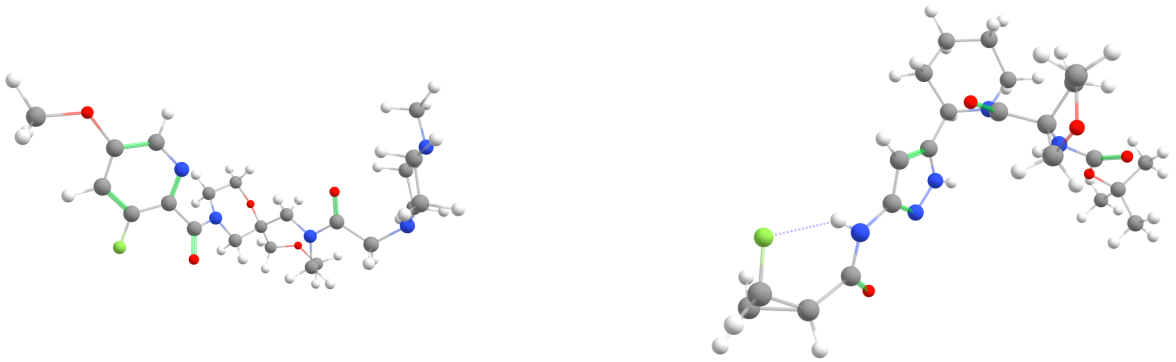

Figure S2: A pair of isomers (distinct molecules) having similarity of 0.99 in the Uni-Mol embedding space and 0.20 in ConforFormer–OMol

Figure S3: A pair of isomers (distinct molecules) having similarity of 0.99 in the Uni-Mol embedding space and 0.70 in ConforFormer–OMol

Figure S4: A pair of isomers (distinct molecules) having similarity of 0.98 in the Uni-Mol embedding space and 0.40 in ConforFormer–OMol

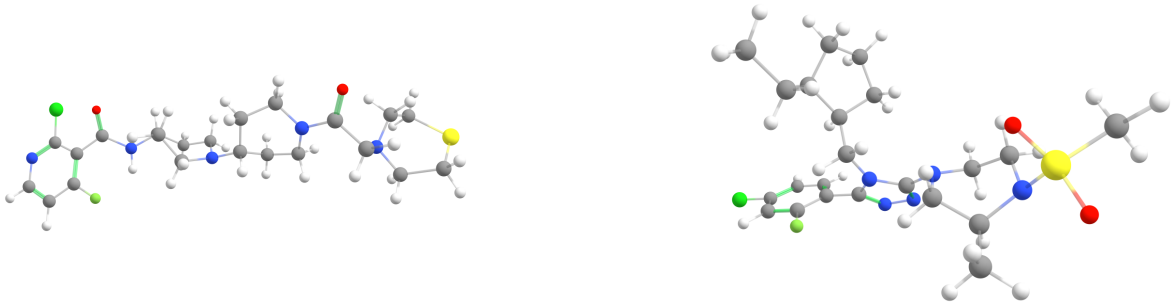

Figure S5: A pair of isomers (distinct molecules) having similarity of 0.98 in the Uni-Mol embedding space and 0.20 in ConforFormer–OMol

Figure S6: A pair of isomers (distinct molecules) having similarity of 0.93 in the Uni-Mol embedding space and 0.14 in ConforFormer–OMol

F.2 Conformers

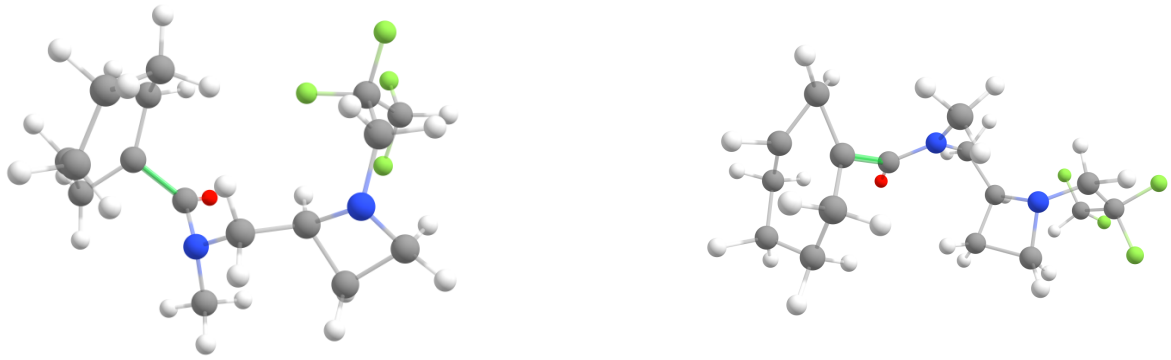

Figure S7: A pair of conformers of the same molecule having similarity of 0.95 (very low, below the 2nd percentile of all pairs) in the Uni-Mol embedding space and 0.99 in ConforFormer–OMol

Figure S8: A pair of conformers of the same molecule having similarity of 0.95 (very low, below the 2d percentile ofd all pairs) in the Uni-Mol embedding space and 0.99 in ConforFormer–OMol

Figure S9: A pair of conformers of the same molecule having similarity of 0.96 in the Uni-Mol embedding space and 0.64 in ConforFormer–OMol. *Note* the distorted ring in the right image is an indication of improperly generated data and is technically not a conformer of the image left. This chemically significant distortion in the structure is not detected by Uni-Mol.

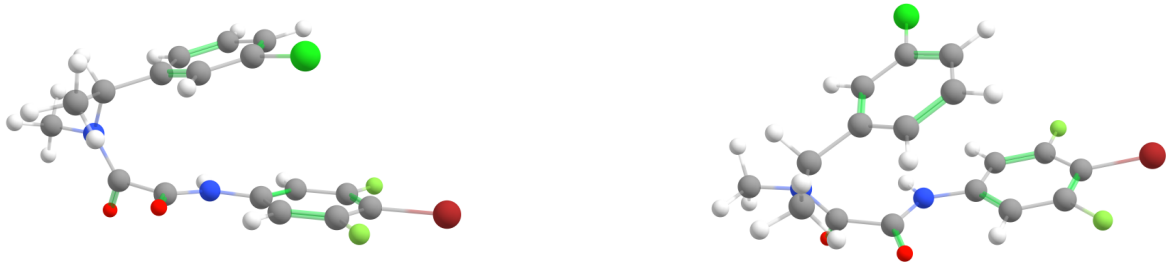

Figure S10: A pair of conformers of the same molecule having similarity of 0.94 in the Uni-Mol embedding space and 0.99 in ConforFormer–OMol.

References

[1] G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang and G. Ke, *Uni-Mol: A Universal 3D Molecular Representation Learning Framework*, 2022, https://chemrxiv.org/doi/full/10.26434/chemrxiv-2022-jjm0j-v4, ChemRxiv.

[2] D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba, P. Eastman, N. C. Frey, X. Fu, V. Gharakhanyan, A. S. Krishnapriyan, J. A. Rackers, S. Raja, A. Rizvi, A. S. Rosen, Z. Ulissi, S. Vargas, C. L. Zitnick, S. M. Blau and B. M. Wood, *The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models*, 2025, http://arxiv.org/abs/2505.08762, arXiv:2505.08762 [physics].

[3] P. Walters, *We need better benchmarks for machine learning in drug discovery*, Practical Cheminformatics Blog, 2023, https://practicalcheminformatics.blogspot.com/2023/08/we-need-better-benchmarks-for-machine.html, Accessed: 2026-04-29.

[4] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *Journal of Chemical Information and Modeling*, 2020, **60**, 6065–6073.

[5] S. Riniker and G. A. Landrum, *Journal of Chemical Information and Modeling*, 2015, **55**, 2562–2574.

[6] T. A. Halgren, *Journal of Computational Chemistry*, 1996, **17**, 490–519.

[7] M. Klein, I. Rudenko, E. Pidko and I. Bushmarinov, *PharmaIsomer*, 2026, https://zenodo.org/records/18739668.