nature portfolio

Peer Review File

A Molecular Video-derived Foundation Model for Scientific Drug Discovery

REVIEWER COMMENTS

Reviewer #2 (Remarks to the Author):

Summary:
The authors propose a molecular video-based foundation model, VideoMol, targeted at representation learning for 3D conformers downstream tasks. To learn useful representations of conformers, VideoMol is trained with three self-supervised learning strategies: Direction-Aware Pretraining Video-Aware Pretraining and Chemical-Aware pretraining, Then subsequently fine-tuned on different prediction tasks. An advantage the proposed method over similar deep learning approaches, is the provided interpretability through denoting key chemical substructures related to 3D conformational changes that according to the manuscript overlap with previous domain knowledge. The experiments show promising results of VideoMol in diverse drug discovery datasets for predicting molecular targets and properties.

The manuscript is well written and self contained. The authors made the effort to include diverse set of baselines including foundation models learnt on sequence, structure and image which gives a nice comparison across the current landscape of available methods. All experiments and results are described in detail and highlight multiple advantages of video-based features compared to current SOTA. It is encouraging that the inhibitors suggested by VideoMol overlap, to some extent, to results from other published domain studies.

Questions:
1. As the authors mentioned themselves, using videos increases complexity of conformer prediction models, which is an already challenging setup because of the 3D structure learning. Are the 60 frames the minimal or optimal number of frames to be used? Has this been explored?

2. Could the authors elaborate on how helpful will the features learned by VideoMol without fine tuning for a new dataset? Would it be beneficial for practiioners to use VideoMol as pretrained models to extract features and then use those for downstream tasks (without fine tuning)?

3. Why is there no standard error/deviation for any of the results? Have the experiments been repeated multiple times? Having some indication about the uncertainty intervals around the results would be appreciated.

4. It will help if the authors include some intuition on why videos make more sense and contribute to learning better features. It is not so easy to understands for readers that are less experienced with conformers. Some running/motivating example could help the presentation.

5. How sensitive is the framework to the source for video generation, in this case RDKit. I assume it is quite dependent on this platform, and thus, VideoMol probably does not allow for mixing videos from different sources. Does this pose any kind of limitation in real-world applications?

Reviewer #2 (Remarks on code availability):

The provided code is clean and well organised. The authors also included a docker image for setting up the environment. The code for training the model as well as reproducing the results is included. I believe only the code for VideoMol is provided, and not for the baseline methods. Not necessary, but if it is easy to include the related methods, the community might appreciate a full testbed.

Reviewer #3 (Remarks to the Author):

Noteworthy results:

The manuscript presents a video-based pretrained model that can be used to make downstream task predictions across multiple tasks with finetuning. The results are an improvement upon authors' previous work, ImageMol. The improvement is tied to use of video, which can be considered as an augmentation to static images, as well as the use of more comprehensive fingerprints that provide chemical, pharmacological and physicochemistry information. The authors show, through extensive testing, that the new model outperforms the previous and is at least as good or as better as some SOTA models for different tasks.

Impact to field:

The work is valuable to the field in multiple areas: it is a demonstration of technology transfer from video representation learning. It shows new self supervision tasks that are meaningful for molecule structure videos. It identifies a large set of benchmark cases. However I see major drawbacks or open questions that would limit its use beyond limited academic interest:

- Unlike stated, the model does not capture a dynamic conformation of the molecule. The videos are not generated to represent any physical dynamics, or conformer change, or changes to torsion angles etc. They are movies with standardized rotations around given axis. As such, they are only augmentations to enrich the model input about the 3d structure of the molecule. Authors should consider another wording than dynamics to prevent misleading the reader.

- if the manuscript's main aim was to inject more information about the 3d nature of the molecules, they could have considered an equivariant graph neural network or transformer. An equivariant neural network would remove the need to perform augmentation for different rotations.

- This is where it gets interesting: if the success of the model was truly due to better representation of 3d structure, we would expect the model to be sensitive to different conformers, especially on tasks that provide a binding affinity proxy. While, in multiple places in the manuscript the opposite is claimed, that model is robust to molecule conformer choice. Perhaps authors can devise an experiment to understand why video information does not lead to sensitivity to conformer.

-Which makes me think the success of the model is not due to better 3d representation but one of the several other changes: 1-working with video frames and the new self-supervision tasks have

expanded the effective size of the data and complexity the network processes each molecule (perhaps this is why the number of frames seem to change the prediction accuracy) 2-the large number of domain information that is crafted into the fingerprint may be impactful in several tasks in this work. Further understanding from where exactly the accuracy improvement comes from, can be considered for future work. In the teamtime, the claims of impact of 3D could be dialed down.

-Feedback to the methodology: the splits used in this work would not stop data leaking from train to validation sets and scaffold balancing might not be enough. Indeed we see a hint of the issue in the COX examples where training data from ChEMBL in 8:1:1 split gave high ROC-AU >0.9, but when the model was tested against MedChemExpress data, only less than 40% of inhibitors are successfully identified. This difference may be due to data leak in the high ROC-AUC train-test data, inflating the apparent generalizability. In general if authors would like to claim generalizabilty, more attention to the split stragety, overlap between data points is needed according to certain similarity metric will be needed.

-Some of the tasks in the work are for high-throughout applications (e.g. virtual screening). In such cases the trade-off between accuracy and compute becomes important. The proposed method should clearly state the compute needs for pretraining and various downstream tasks. Because it works with video, compared to much smaller atomic position files, memory needs should be highlighted too.

-minor typos in text, highlighting one that is on figure in case it escapes proofreading: angel -> angle Fig 1b

Reviewer #3 (Remarks on code availability):

Lightly reviewed code. Checked the pretraining tools and the base encoder model definition. Looks rather standard, didnt see any weird libraries or so. Didn't run but I didnt see a reason why it wouldnt. Also there are links to pretraining data and to pretrained model to reproduce inference results in the paper.

Point-by-Point Response Letter

Manuscript #: NCOMMS-23-62188

We are grateful to the reviewers for their insightful and constructive feedback on our manuscript. In response to the feedback, we provide the detailed responses to address each reviewer's concerns point by point as follows.

Responses to the Reviewer #1

Overall Summary – "The authors made the effort to include diverse set of baselines, a **nice comparison** across the current landscape of available methods. All experiments and results are described in detail and highlight **multiple advantages** of video-based features**.** The manuscript is **well written" –**

Reviewer Comment	The authors propose a molecular video-based foundation model, VideoMol, targeted at representation learning for 3D conformers downstream tasks. To learn useful representations of conformers, VideoMol is trained with three self-supervised learning strategies: Direction-Aware Pretraining Video-Aware Pretraining and Chemical-Aware pretraining, Then subsequently fine-tuned on different prediction tasks. An advantage the proposed method over similar deep learning approaches, is the provided interpretability through denoting key chemical substructures related to 3D conformational changes that according to the manuscript overlap with previous domain knowledge. The experiments show promising results of VideoMol in diverse drug discovery datasets for predicting molecular targets and properties. The manuscript is well written and self contained. The authors made the effort to include diverse set of baselines including foundation models learnt on sequence, structure and image which gives a nice comparison across the current landscape of available methods. All experiments and results are described in detail and highlight multiple advantages of video-based features compared to current SOTA. It is encouraging that the inhibitors suggested by VideoMol overlap, to some extent, to results from other published domain studies.
Author Response	We thank the Reviewer for comprehensive summary and his/her positive support on the manuscript. We have made extensive revision to address the reviewer's critiques as below.

Ref 1.1 – "Are the 60 frames the minimal or optimal number of frames to be used? Has this been explored?" –

<table>
<tr><td>Reviewer Comment</td><td>As the authors mentioned themselves, using videos increases complexity of conformer prediction models, which is an already challenging setup because of the 3D structure learning. Are the 60 frames the minimal or optimal number of frames to be used? Has this been explored?</td></tr>
<tr><td>Author Response</td><td>We thank the Reviewer for this great point about the optimized number of frames. We used the 60 frames after balancing both the optimized model performance and the computational cost. We explored the impact of frame number on the model performance in Supplementary Table 26, including 5 frames, 10 frames, 20 frames, 30 frames and 60 frames. We found that the increased number of frames improve the performance of the models. After balancing computing time and the optimized model performance by the increased number of frames, we selected 60 frames. We have added these new results and more detailed explanations in the revised manuscript.

Supplementary Table 26: Effect of frame number on VideoMol on 6 regression datasets with balanced scaffold split. #frame indicates the number of frames. All means and standard deviations are reported through three independent runs.
<table>
<tr><th></th><th colspan="2">5HT1A</th><th colspan="2">AA1R</th><th colspan="2">AA2AR</th></tr>
<tr><th>#frame</th><th>RMSE</th><th>MAE</th><th>RMSE</th><th>MAE</th><th>RMSE</th><th>MAE</th></tr>
<tr><td>5</td><td>0.873±0.023</td><td>0.693±0.012</td><td>0.847±0.013</td><td>0.656±0.025</td><td>0.824±0.009</td><td>0.659±0.012</td></tr>
<tr><td>10</td><td>0.800±0.025</td><td>0.624±0.017</td><td>0.773±0.010</td><td>0.584±0.006</td><td>0.766±0.002</td><td>0.609±0.005</td></tr>
<tr><td>20</td><td>0.765±0.026</td><td>0.591±0.010</td><td>0.728±0.003</td><td>0.546±0.005</td><td>0.744±0.005</td><td>0.586±0.005</td></tr>
<tr><td>30</td><td>0.742±0.014</td><td>0.573±0.011</td><td>0.704±0.004</td><td>0.527±0.001</td><td>0.736±0.019</td><td>0.570±0.013</td></tr>
<tr><td>60</td><td>0.708±0.017</td><td>0.547±0.015</td><td>0.655±0.007</td><td>0.496±0.006</td><td>0.712±0.011</td><td>0.543±0.005</td></tr>
<tr><th></th><th colspan="2">CNR2</th><th colspan="2">DRD2</th><th colspan="2">HRH3</th></tr>
<tr><th>#frame</th><th>RMSE</th><th>MAE</th><th>RMSE</th><th>MAE</th><th>RMSE</th><th>MAE</th></tr>
<tr><td>5</td><td>0.991±0.032</td><td>0.811±0.027</td><td>0.875±0.016</td><td>0.660±0.012</td><td>0.733±0.017</td><td>0.567±0.009</td></tr>
<tr><td>10</td><td>0.950±0.032</td><td>0.762±0.039</td><td>0.822±0.014</td><td>0.620±0.003</td><td>0.696±0.013</td><td>0.540±0.014</td></tr>
<tr><td>20</td><td>0.910±0.016</td><td>0.721±0.011</td><td>0.792±0.001</td><td>0.600±0.004</td><td>0.679±0.019</td><td>0.521±0.014</td></tr>
<tr><td>30</td><td>0.890±0.013</td><td>0.700±0.009</td><td>0.769±0.005</td><td>0.580±0.003</td><td>0.686±0.019</td><td>0.531±0.015</td></tr>
<tr><td>60</td><td>0.864±0.005</td><td>0.679±0.010</td><td>0.742±0.004</td><td>0.556±0.005</td><td>0.668±0.008</td><td>0.506±0.002</td></tr>
</table></td></tr>
<tr><td>Excerpt from Revised Manuscript</td><td>The impact of the video frame number. To explore the impact of different frame numbers on VideoMol, we sampled 5, 10, 20, 30, and 60 molecular frames from 5HT1A, AA1R, AA2AR, CNR2, DRD2, and HRH3 datasets at equal time intervals. We found that the performance of VideoMol is positively correlated with the number of frames with an average performance improvement of 6.5% (5→10 frames), 3.9% (10→20 frames), 2.0% (20→30 frames), 3.9% (30→60 frames) on RMSE metric and 7.6% (5→10 frames), 4.7% (10→20 frames), 2.4% (20→30 frames), 4.4% (30→60 frames) on MAE metric, which shows that the increase of frame number enriches the 3D information extracted by VideoMol and its performance may be expected to be further increased by expanding the frame number (Supplementary Table 26).</td></tr>
</table>

Ref 1.2 – "Effectiveness of features learned by VideoMol without fine-tuning on new datasets" –

Reviewer Comment	Could the authors elaborate on how helpful will the features learned by VideoMol without fine tuning for a new dataset? Would it be beneficial for practitioners to use VideoMol as pretrained models to extract features and then use those for downstream tasks (without fine tuning)?
Author Response	We thank the reviewer for this great point. We conducted new experiments and reported new results on downstream tasks using features extracted by pretrained VideoMol (called VideoMolFeat) in **Supplementary Table 25**. We also evaluated features extracted by ensemble fingerprints for comparison (called EnsembleFP). To be fair, we did not fine-tune VideoMol and directly input VideoMolFeat and EnsembleFP into a structurally identical multi-layer perceptron (MLP). New experimental results showed that VideoMolFeat achieved the best performance with 17.2% average RMSE improvement and 19.4% average MAE improvement compared with EnsembleFP on 10 kinases datasets. These new findings (without fine tuning) show that the features extracted by VideoMol are superior alternative compared with traditional molecular fingerprinting.

Table S25: The performance of molecular fingerprint on 10 kinases datasets with balanced scaffold split. EnsembleFP-MLP indicates the performance of integrating traditional molecular fingerprints in CAP and training an MLP.VideoMolFeat-MLP represents the performance of using VideoMol to extract molecular features and train an MLP.

	1. 5HT1A		2. 5HT2A		3. AA1R	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
EnsembleFP-MLP	1.030±0.000	0.837±0.001	1.207±0.005	0.975±0.006	0.915±0.005	0.734±0.005
VideoMolFeat-MLP	**0.818±0.031**	**0.653±0.033**	**0.910±0.014**	**0.695±0.017**	**0.838±0.007**	**0.647±0.005**
—— continue ——						
	4. AA2AR		5. AA3R		6. CNR2	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
EnsembleFP-MLP	1.076±0.003	0.871±0.002	0.991±0.001	0.791±0.002	1.228±0.004	1.027±0.002
VideoMolFeat-MLP	**0.850±0.005**	**0.691±0.004**	**0.862±0.014**	**0.695±0.012**	**1.003±0.070**	**0.789±0.066**
—— continue ——						
	7. DRD2		8. DRD3		9. HRH3	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
EnsembleFP-MLP	0.979±0.000	0.785±0.001	1.096±0.004	0.933±0.002	0.888±0.003	0.681±0.002
VideoMolFeat-MLP	**0.885±0.003**	**0.660±0.002**	**0.837±0.012**	**0.663±0.011**	**0.741±0.001**	**0.578±0.002**
—— continue ——						
	10. OPRM					
	RMSE	MAE				
EnsembleFP-MLP	1.023±0.006	0.815±0.002				
VideoMolFeat-MLP	**0.875±0.004**	**0.672±0.003**				

Ref 1.3 – "Standard error/deviation of experimental results repeated multiple times" –

Reviewer Comment	Why is there no standard error/deviation for any of the results? Have the experiments been repeated multiple times? Having some indication about the uncertainty intervals around the results would be appreciated.
Author Response	We repeated all molecular property prediction experiments for 10 times with random seeds 0 to 9 and the remaining experiments were repeated with 3 random seeds 0 to 2. All these new standard deviations have been provided in the **Supplementary Tables 3-7**. We found that overall standard error/deviation and 95% CI are very small, indicating robustness of VideoMol models.

Table S3: The ROC-AUC performance of different methods on 10 main types of biochemical kinases from KinomeScan datasets with balanced scaffold split. All compared results are obtained from ImageMol.

	BTK	CDK4-cyclinD3	EGFR	FGFR1	FGFR2
MoCLR$_{GIN}$	0.556±0.118	0.778±0.171	0.583±0.067	0.695±0.249	0.667±0.132
MoCLR$_{GCN}$	0.602±0.129	0.944±0.039	0.750±0.051	0.619±0.378	0.667±0.052
RNN_LR	0.611±0.000	0.667±0.000	0.536±0.000	0.771±0.000	0.741±0.000
TRFM_LR	0.694±0.000	0.750±0.000	0.821±0.000	0.743±0.000	0.704±0.000
RNN_MLP	0.556±0.023	0.833±0.000	0.536±0.029	0.848±0.059	0.716±0.046
TRFM_MLP	0.537±0.013	0.639±0.039	0.667±0.061	0.643±0.031	0.741±0.000
RNN_RF	0.546±0.013	0.917±0.000	0.548±0.017	0.476±0.027	0.685±0.055
TRFM_RF	0.639±0.039	0.639±0.039	0.607±0.000	0.476±0.013	0.556±0.030
CHEM-BERT	0.648±0.013	0.583±0.297	0.845±0.094	0.429±0.117	0.765±0.106
ImageMol	0.843±0.026	0.917±0.068	0.857±0.000	**0.857±0.023**	0.852±0.052
VideoMol	**0.861±0.023**	**0.972±0.039**	**0.905±0.017**	0.848±0.027	**0.988±0.017**
			------ Continue ------		
	FGFR3	**FGFR4**	**FLT3**	**KPCD3**	**MET**	**Average**
MoCLR$_{GIN}$	0.760±0.039	0.773±0.121	0.722±0.091	0.571±0.107	0.611±0.236	0.6716
MoCLR$_{GCN}$	0.792±0.106	0.537±0.013	0.722±0.208	0.505±0.067	0.574±0.052	0.6712
RNN_LR	0.646±0.015	0.528±0.000	0.778±0.000	0.457±0.000	0.796±0.026	0.6531
TRFM_LR	0.812±0.000	0.639±0.000	0.611±0.000	0.438±0.027	0.778±0.000	0.6990
RNN_MLP	0.469±0.026	0.269±0.035	0.630±0.105	0.410±0.036	0.667±0.045	0.5934
TRFM_MLP	0.802±0.015	0.676±0.035	0.667±0.136	0.219±0.027	0.556±0.000	0.6147
RNN_RF	0.312±0.000	0.389±0.000	0.519±0.026	0.343±0.000	0.500±0.000	0.5235
TRFM_RF	0.646±0.078	0.602±0.013	0.546±0.035	0.262±0.058	0.593±0.026	0.5566
CHEM-BERT	0.438±0.077	0.528±0.060	0.574±0.189	0.557±0.091	0.944±0.000	0.6311
ImageMol	0.854±0.064	0.833±0.045	0.722±0.120	0.762±0.088	**0.963±0.026**	0.8460
VideoMol	**0.896±0.039**	**0.852±0.080**	**0.981±0.026**	**0.867±0.036**	**0.963±0.026**	**0.9133**

Table S4: The RMSE and MAE performance of different methods on 10 GPCR with balanced scaffold split. The lower the value, the better the performance. All compared results are obtained from ImageMol.

	1. 5HT1A		2. 5HT2A		3. AA1R	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
MoCLR$_{GIN}$	0.850±0.021	0.670±0.012	0.853±0.019	0.642±0.014	0.786±0.015	0.588±0.009
MoCLR$_{GCN}$	0.949±0.027	0.764±0.014	0.875±0.008	0.681±0.024	0.856±0.026	0.662±0.026
RNN_LR	1.574±0.091	0.937±0.019	1.602±0.245	1.103±0.151	1.073±0.087	0.762±0.057
TRFM_LR	1.636±0.004	1.109±0.001	1.389±0.000	0.999±0.001	1.060±0.003	0.810±0.001
RNN_MLP	0.957±0.013	0.768±0.010	1.167±0.010	0.890±0.003	0.848±0.004	0.662±0.008
TRFM_MLP	0.939±0.034	0.730±0.025	1.013±0.026	0.728±0.021	0.878±0.051	0.657±0.031
RNN_RF	0.788±0.004	0.617±0.004	1.001±0.001	0.747±0.001	0.717±0.003	0.554±0.002
TRFM_RF	0.855±0.001	0.672±0.001	1.011±0.002	0.777±0.002	0.740±0.001	0.568±0.001
CHEM-BERT	0.876±0.018	0.706±0.012	0.909±0.057	0.682±0.056	0.734±0.038	0.544±0.027
ImageMol	0.776±0.012	0.620±0.014	0.780±0.017	0.578±0.022	0.711±0.012	0.554±0.009
VideoMol	**0.708±0.017**	**0.547±0.015**	**0.775±0.017**	**0.577±0.009**	**0.655±0.007**	**0.496±0.006**
	4. AA2AR		**5. AA3R**		**6. CNR2**	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
MoCLR$_{GIN}$	0.748±0.012	0.588±0.008	0.840±0.014	0.692±0.010	0.926±0.047	0.758±0.036
MoCLR$_{GCN}$	0.819±0.011	0.651±0.008	0.855±0.010	0.700±0.011	0.978±0.023	0.803±0.021
RNN_LR	1.801±0.600	1.193±0.335	2.295±0.463	1.190±0.155	5.505±0.093	1.611±0.032
TRFM_LR	1.130±0.000	0.906±0.000	1.155±0.001	0.919±0.001	1.700±0.001	1.213±0.000
RNN_MLP	0.967±0.002	0.773±0.005	0.883±0.010	0.707±0.012	1.091±0.015	0.881±0.013
TRFM_MLP	0.948±0.013	0.744±0.005	0.945±0.010	0.749±0.014	1.144±0.055	0.903±0.038
RNN_RF	0.887±0.002	0.692±0.001	0.796±0.009	0.624±0.007	0.965±0.002	0.766±0.001
TRFM_RF	0.926±0.003	0.735±0.004	0.856±0.001	0.701±0.002	0.965±0.002	0.800±0.002
CHEM-BERT	0.862±0.071	0.674±0.058	0.861±0.058	0.684±0.047	0.925±0.051	0.727±0.041
ImageMol	0.734±0.015	0.573±0.009	0.793±0.008	0.634±0.001	0.905±0.004	0.717±0.015
VideoMol	**0.712±0.011**	**0.543±0.005**	**0.786±0.006**	**0.617±0.004**	**0.864±0.005**	**0.679±0.010**
	7. DRD2		**8. DRD3**		**9. HRH3**	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
MoCLR$_{GIN}$	0.814±0.009	0.591±0.007	0.858±0.017	0.673±0.022	0.734±0.006	0.581±0.004
MoCLR$_{GCN}$	0.855±0.022	0.634±0.017	0.914±0.024	0.725±0.025	0.740±0.016	0.576±0.006
RNN_LR	1.142±0.077	0.839±0.038	1.316±0.011	0.942±0.005	1.616±0.236	0.943±0.070
TRFM_LR	1.000±0.000	0.719±0.000	1.219±0.000	0.914±0.000	1.169±0.002	0.911±0.002
RNN_MLP	0.895±0.006	0.694±0.005	1.021±0.007	0.819±0.007	0.871±0.015	0.702±0.011
TRFM_MLP	0.919±0.016	0.686±0.016	1.012±0.041	0.790±0.023	0.863±0.011	0.676±0.009
RNN_RF	0.837±0.001	0.612±0.001	0.861±0.001	0.685±0.001	0.771±0.002	0.613±0.002
TRFM_RF	0.864±0.001	0.636±0.002	0.904±0.001	0.717±0.001	0.770±0.002	0.602±0.001
CHEM-BERT	0.816±0.011	0.587±0.013	0.803±0.029	0.631±0.026	0.770±0.033	0.594±0.022
ImageMol	0.772±0.014	0.573±0.009	0.735±0.018	0.576±0.014	0.710±0.006	0.561±0.006
VideoMol	**0.742±0.004**	**0.556±0.005**	**0.715±0.014**	**0.554±0.012**	**0.668±0.008**	**0.506±0.002**
	10. OPRM					
	RMSE	**MAE**				
MoCLR$_{GIN}$	0.856±0.008	0.664±0.016				
MoCLR$_{GCN}$	0.853±0.009	0.653±0.020				
RNN_LR	2.649±1.024	1.744±0.614				
TRFM_LR	1.694±0.001	1.282±0.000				
RNN_MLP	1.022±0.014	0.781±0.008				
TRFM_MLP	1.084±0.009	0.849±0.007				
RNN_RF	0.876±0.010	0.671±0.009				
TRFM_RF	0.852±0.002	0.660±0.003				
CHEM-BERT	0.893±0.024	0.672±0.019				
ImageMol	0.849±0.018	0.645±0.015				
VideoMol	**0.795±0.015**	**0.579±0.011**				

Table S5: The ROC-AUC performance (%) of different methods on 6 molecular property prediction benchmarks with scaffold split. All experiments are run 10 times using random seeds from 0 to 9. GraphMVP-C, Mole-BERT, Uni-Mol and ImageMol are reproduced from their source code and other results from Mole-BERT.

	Tox21	ToxCast	Sider	HIV	BBBP	BACE
#Molecules	7831	8576	1427	41127	2039	1513
#Task	12	617	27	1	1	1
InfoGraph	73.3 (0.6)	61.8 (0.4)	58.7 (0.6)	74.2 (0.9)	68.7 (0.6)	74.3 (2.6)
GPT-GNN	74.9 (0.3)	62.5 (0.4)	58.1 (0.3)	65.2 (2.1)	64.5 (1.4)	77.9 (3.2)
ContextPred	73.6 (0.3)	62.6 (0.6)	59.7 (1.8)	75.6 (1.0)	70.6 (1.5)	78.8 (1.2)
GraphLoG	75.0 (0.6)	63.4 (0.6)	59.6 (1.9)	76.1 (0.8)	68.7 (1.6)	78.6 (1.0)
G-Contextual	75.0 (0.6)	62.8 (0.7)	58.7 (1.0)	76.3 (1.5)	69.9 (2.1)	79.3 (1.1)
G-Motif	73.6 (0.7)	62.3 (0.6)	61.0 (1.5)	73.8 (1.2)	66.9 (3.1)	73.0 (3.3)
AD-GCL	74.9 (0.4)	63.4 (0.7)	61.5 (0.9)	76.7 (1.2)	70.7 (0.3)	76.6 (1.5)
JOAO	74.8 (0.6)	62.8 (0.7)	60.4 (1.5)	76.9 (0.7)	66.4 (1.0)	73.2 (1.6)
SimGRACE	74.4 (0.3)	62.6 (0.7)	60.2 (0.9)	75.0 (0.6)	**71.2 (1.1)**	74.9 (2.0)
GraphCL	75.1 (0.7)	63.0 (0.4)	59.8 (1.3)	75.1 (0.7)	67.8 (2.4)	74.6 (2.1)
GraphMAE	75.2 (0.9)	63.6 (0.3)	60.5 (1.2)	76.8 (0.6)	**71.2 (1.0)**	78.2 (1.5)
3D InfoMax	74.5 (0.7)	63.5 (0.8)	56.8 (2.1)	76.1 (1.3)	69.1 (1.2)	78.6 (1.9)
MGSSL	75.2 (0.6)	63.3 (0.5)	61.6 (1.0)	75.8 (0.4)	68.8 (0.6)	78.8 (0.9)
AttrMask	75.1 (0.9)	63.3 (0.6)	60.5 (0.9)	75.3 (1.5)	65.2 (1.4)	77.8 (1.8)
MolCLR	75.5 (0.5)	63.9 (0.5)	60.3 (1.3)	74.4 (1.3)	66.8 (3.4)	75.3 (2.9)
GraphMVP-C	74.6(0.4)	63.4 (0.6)	60.6 (1.3)	77.1 (2.1)	69.9 (1.4)	79.6 (1.7)
ImageMol	75.5 (1.0)	65.6 (0.9)	64.9 (1.3)	76.8 (1.3)	70.5 (1.3)	78.1 (3.5)
Uni-Mol (1 conf)	78.3 (0.4)	68.7 (0.5)	63.7 (1.3)	79.2 (1.0)	69.6 (2.0)	81.0 (3.9)
Uni-Mol (10 conf)	**78.8 (0.7)**	**69.0 (0.5)**	63.6 (1.4)	79.2 (0.9)	69.9 (2.7)	81.7 (3.4)
Mole-BERT	77.0 (0.3)	64.4 (0.2)	63.2 (0.7)	77.7 (0.7)	65.7 (2.3)	80.2 (0.9)
VideoMol	**78.8 (0.5)**	66.7 (0.5)	**66.3 (0.9)**	**79.4 (0.5)**	70.7 (2.2)	**82.4 (0.9)**
Rank	1	2	1	1	3	1

Table S6: The RMSE or MAE performance of different methods on 6 molecular property prediction benchmarks with scaffold split. All experiments are run 10 times using random seeds from 0 to 9. We report RMSE for FreeSolv, ESOL and Lipo datasets and MAE for QM7 and QM8, QM9 datasets, respectively. We reproduced all comparison methods using the same settings. We use GIN backbone for MolCLR because it achieves the best results..

RMSE	FreeSolv	ESOL	Lipo	QM7	QM8	QM9
GraphMVP	2.559±0.158	1.322±0.062	0.773±0.016	120.344±6.237	0.02049±0.00032	**0.00891±0.00010**
EdgePred	2.843±0.091	1.367±0.041	0.778±0.013	104.387±3.292	0.02058±0.00061	0.00929±0.00009
GraphMVP-C	2.766±0.199	1.333±0.055	0.768±0.013	121.022±5.699	0.02022±0.00047	0.00896±0.00011
MolCLR	3.112±0.638	1.462±0.068	0.799±0.018	144.426±6.591	0.03598±0.00085	0.01488±0.00020
ImageMol	2.113±0.235	0.964±0.067	**0.702±0.060**	116.384±8.445	0.02419±0.00033	0.02061±0.00019
Mole-BERT	2.988±0.155	1.115±0.017	0.727±0.006	101.922±2.331	0.02073±0.00033	0.00910±0.00010
VideoMol	**1.728±0.053**	**0.866±0.017**	0.743±0.009	**76.736±1.561**	**0.01890±0.00020**	0.00896±0.00003

Table S7: The ROC-AUC performance of different methods on 11 SARS-CoV-2 datasets with balanced scaffold split. All compared results are obtained from ImageMol.

	3CL	ACE2	hCYTOX	MERS-PPE_cs	MERS-PPE	CoV1-PPE_cs
REDIAL-2020	0.713	0.753	0.710	0.703	0.696	0.661
ImageMol	**0.762±0.007**	0.720±0.001	0.727±0.009	0.771±0.009	0.773±0.011	0.775±0.005
VideoMol	0.709±0.006	**0.759±0.025**	**0.765±0.003**	**0.828±0.027**	**0.814±0.004**	**0.836±0.029**
===== continue =====						
	CoV1-PPE	CPE	Cytotox	AlphaLISA	TruHit	Mean
REDIAL-2020	0.665	0.651	0.688	0.79	0.734	0.706
ImageMol	0.703±0.008	0.669±0.011	0.728±0.001	0.793±0.007	0.806±0.006	0.748
VideoMol	**0.737±0.007**	**0.747±0.013**	**0.761±0.002**	**0.841±0.004**	**0.862±0.002**	**0.787**

Furthermore, we calculated the uncertainty intervals with 95% confidence intervals (CI) of ImageMol and VideoMol on 10 compound-kinase interaction datasets and 11 SARS-CoV-2 viral activity prediction datasets. In details, we used the popular bias-corrected and accelerated (BCa) bootstrap intervals [1][2] to calculate 95% uncertainty intervals, which corrects for both

bias and skewness of the bootstrap parameter estimates by incorporating a bias-correction factor and an acceleration factor. The results of the uncertainty interval are reported in the Extended Table 1 (the Revised Supplemental Table 8) and Extended Table 2 (the Revised Supplemental Table 9) below, which shows the effectiveness of VideoMol with an average improvement ranging from 5.44% to 10.07%.

In summary, these new experiments highlight the robustness of our VideoMol models. We have added these new experiments and more detailed explanations in the revised manuscript.

Extended Table 1 (the Revised Supplemental Table 8). The uncertainty intervals with 95% confidence intervals of ImageMol and VideoMol on 10 compound-kinase interaction datasets. UI(·) represents the uncertainty intervals and "Improvement" represents the relative performance improvement of VideoMol compared to ImageMol.

Dataset	UI(RMSE)			UI(MAE)		
	ImageMol	VideoMol	Improvement	ImageMol	VideoMol	Improvement
5HT1A	0.782±0.057	**0.719±0.059**	8.06%	0.629±0.048	**0.550±0.046**	12.56%
5HT2A	0.816±0.109	**0.810±0.102**	0.74%	0.587±0.059	**0.583±0.059**	0.68%
AA1R	0.718±0.062	**0.662±0.068**	7.80%	0.559±0.045	**0.499±0.046**	10.73%
AA2AR	0.739±0.055	**0.714±0.056**	3.38%	0.575±0.045	**0.544±0.045**	5.39%
AA3R	0.796±0.056	**0.795±0.065**	0.13%	0.632±0.051	**0.622±0.053**	1.58%
CNR2	0.916±0.073	**0.878±0.072**	4.15%	0.722±0.060	**0.686±0.064**	4.99%
DRD2	0.779±0.060	**0.749±0.053**	3.85%	0.574±0.041	**0.559±0.040**	2.61%
DRD3	0.738±0.053	**0.704±0.054**	4.61%	0.580±0.044	**0.548±0.042**	5.52%
HRH3	0.747±0.067	**0.669±0.061**	10.44%	0.582±0.050	**0.507±0.047**	12.89%
OPRM	0.898±0.089	**0.797±0.075**	11.25%	0.667±0.065	**0.584±0.062**	12.44%

Extended Table 2 (the Revised Supplemental Table 9). The uncertainty intervals with 95% confidence intervals of ImageMol and VideoMol on 11 SARS-CoV-2 viral activity prediction datasets. UI(·) represents the uncertainty intervals and "Improvement" represents the relative performance improvement of VideoMol compared to ImageMol.

Dataset	UI(AUC)		
	ImageMol	VideoMol	Improvement
3CL	0.685±0.117	**0.710±0.110**	3.65%
ACE2	0.658±0.133	**0.763±0.112**	15.96%
hCYTOX	0.736±0.087	**0.760±0.087**	3.26%
MERS-PPE_cs	0.727±0.119	**0.817±0.103**	12.38%
MERS-PPE	0.720±0.082	**0.799±0.074**	10.97%
CoV1-PPE_cs	0.688±0.115	**0.832±0.098**	20.93%
CoV1-PPE	0.701±0.063	**0.736±0.060**	4.99%
CPE	0.646±0.084	**0.736±0.077**	13.93%
Cytotox	0.729±0.051	**0.760±0.054**	4.25%

	AlphaLISA	0.762±0.062	**0.836±0.049**	9.71%
	TruHit	0.772±0.059	**0.855±0.046**	10.75%

References
[1] Efron B. Better bootstrap confidence intervals[J]. *Journal of the American statistical Association*, 1987, 82(397): 171-185.
[2] Efron B, Tibshirani R J. An introduction to the bootstrap[M]. Chapman and Hall/CRC, 1994.

Excerpt from Revised Manuscript

For 10 compound-kinase interaction datasets, VideoMol achieves better AUC performance than other methods across BTK (AUC=0.861±0.023), CDK4-cyclinD3 (AUC=0.972±0.039), EGFR (AUC=0.905±0.017), FGFR1 (AUC=0.848±0.027), FGFR2 (AUC=0.988±0.017), FGFR3 (AUC=0.896±0.039), FGFR4 (AUC=0.852±0.080), FLT3 (AUC=0.981±0.026), KPCD3 (AUC=0.867±0.036) and MET (AUC=0.963±0.026) with an average performance improvement of 5.9% ranging from 1.8% to 20.3% (**Fig. 2a** and Supplementary Table 3). In particular, VideoMol outperforms the state-of-the-art methods of ImageMol and MolCLR with average improvements of 6.7% and 20.6%.

In classification task, using the area under the receiver operating characteristic (ROC) curve (AUC), VideoMol achieves elevated performance across BBBP (AUC=70.7%±1.5), Tox21 (AUC=78.8%±0.4), HIV (AUC=79.4%±0.5), BACE (AUC=82.4%±0.9), SIDER (AUC=66.3%±0.9), ToxCast (AUC=66.7%±0.5), outperforming other methods (**Fig. 2c** and Supplementary Table 5). In regression task, VideoMol achieves low error values across FreeSolv (RMSE=1.728±0.053), ESOL (RMSE=0.866±0.017), Lipo (RMSE=0.743±0.009), QM7 (MAE=76.736±1.561), QM8 (MAE=0.01890±0.0020) and QM9 (MAE=0.00896±0.00003), outperforming other methods (**Fig. 2d** and Supplementary Table 6).

We found that VideoMol achieved elevated ROC-AUC performance (3CL=0.709±0.006, ACE2=0.759±0.020, hCYTOX=0.765±0.003, MERS-PPE_cs=0.828±0.027, MERS-PPE=0.814±0.004, CPE=0.747±0.013, CoV1-PPE_cs=0.836±0.029, CoV1-PPE=0.737±0.007, Cytotox=0.761±0.002, AlphaLISA=0.841±0.004, TruHit=0.862±0.002) with an average 3.9% improvement ranging from 3.3% to 7.8% compared with ImageMol and an average 8.1% improvement ranging from 0.6% to 17.5% compared with REDIAL-2020 (**Fig. 2e** and Supplementary Table 7).

Furthermore, we calculated the uncertainty intervals with 95% confidence intervals (CI) of ImageMol and VideoMol using 10 compound-kinase interaction datasets and 11 SARS-CoV-2 viral activity prediction datasets. In details, we used the popular bias-corrected and accelerated (BCa) bootstrap intervals[41,42] to calculate the uncertainty intervals with 95% confidence intervals (CI), which corrects for both bias and skewness of the bootstrap parameter estimated by incorporating a bias-correction factor and an acceleration factor. The results of the uncertainty interval show the effectiveness of VideoMol with an average improvement ranging from 5.44%

	to 10.07% (Supplementary Tables 8-9). References [41] Efron, B. Better bootstrap confidence intervals. Journal of the American statistical Association 82, 171-185 (1987). [42] Efron, B. & Tibshirani, R.J. An introduction to the bootstrap. (Chapman and Hall/CRC, 1994).

Ref 1.4 – "Intuition on why videos make more sense and contribute to learning better features" –

Reviewer Comment	It will help if the authors include some intuition on why videos make more sense and contribute to learning better features. It is not so easy to understands for readers that are less experienced with conformers. Some running/motivating examples could help the presentation.
Author Response	We thank the reviewer for this great point and we have added more intuition and examples on why videos make more sense and contribute to learning better features in the revised Results and Discussion.
Excerpt from Revised Manuscript	Results: **Framework of VideoMol** Molecules exist in nature and are constantly conformational dynamics, making video the most direct representation method. The molecular 3D information can be directly observed from the video without the help of manual feature extraction, such as the distance between pairs of atoms and the angle formed between multiple atoms and so on. In addition, we evaluated the advantages of different representations in feature extraction capabilities and found that our proposed video representation has obvious advantages over existing representations with a 66% improvement rate on 8 basic attributes (Supplementary Methods and Supplementary Table 1). Therefore, these significant differences motivate us to develop VideoMol for accurately predicting the targets and properties of molecules in the form of videos derived from molecules.

Table S1: The RMSE results of different molecular representations on 8 basic attributes (molecular weight, MolLogP, MolMR, BalabanJ, NumHAcceptors, NumHDonors, NumValenceElectrons).

	modality	model	use conformer?	**prop**
graph-based	2D graph	GCN	×	62.304
		GIN	×	62.980
		EGNN	×	17.418
	3D graph	EGNN	√	16.684
image-based	image from imagemol	ResNet18	×	12.469
	video-1frame	ResNet18	√	11.237
	video-5frame	ResNet18	√	8.088
	video-60frame	ViT	√	**7.511**

C.2 Results of different representations on 8 basic attributes

To fairly compare the effects of different representations, we evaluated the

	representation without using any self-supervised tasks. It is well known that the development of drug discovery depends on accurately capturing chemical and biological representations of molecules. Here, we used several commonly used representative methods (such as GCN, GIN, EGNN, and the representation used by ImageMol) to inspect the model's ability to understand the 8 basic attributes of molecules, including molecular weight, MolLogP, MolMR, BalabanJ, NumHAcceptors, NumHDonors, NumValenceElectrons and TPSA. We randomly collected 10,000 molecules from the pre-training dataset and used exactly the same experimental setup for fair comparison. In detail, we split the training set, validation set, and test set using a ratio of 8:1:1 and reported the results on the test set based on the best validation set score. As shown in Supplementary Table 1, we found that VideoMol using only one frame outperformed that of the 2D graph-based methods, the 3D-based graph method and the 2D image-based method, revealing the advantage of 3D representation. Specifically, compared with the second-place ImageMol without pre-training, the performance of video-1frame improved by 11%. When we utilized all video frames (video-60frame), the performance is further significantly improved from 12.47 to 7.55 with a 66% improvement rate. In summary, the proposed 3D representation (whether based on a single frame image or a 60-frame video) has advantages compared to existing molecular representation approaches. We will further improve our VideoMol framework by inceasing the number of 3D frames and integrating other types of 3D representation (such as AlphaFold3[11]) in the near future. [11] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024: 1-3.

Ref 1.5 – "Sensitivity of VideoMol for video generation source" –

Reviewer Comment	How sensitive is the framework to the source for video generation, in this case RDKit. I assume it is quite dependent on this platform, and thus, VideoMol probably does not allow for mixing videos from different sources. Does this pose any kind of limitation in real-world applications?
Author Response	We thank the reviewer for this critique. To evaluate the sensitivity of VideoMol to video generation sources, we utilized two additional methods to generate molecular videos as below: 1. OpenBabel[1]: It is a chemical toolbox designed to code many languages of chemical data, which generates 3D conformer by four steps: (1) Use the OBBuilder to create a 3D structure using rules and fragment templates; (2) Use 250 steps of a steepest descent geometry optimization with the MMFF94 forcefield; (3) Use 200 iterations of a Weighted Rotor conformational search (optimizing each conformer with 25 steps of a steepest descent); (4) Use 250 steps of a conjugate gradient geometry optimization. 2. DeepChem[2]: It aims to provide a high quality open-source toolchain that

democratizes the use of deep-learning in drug discovery, materials science, quantum chemistry, and biology. It uses three steps to generate molecular conformer: (1) Generate a pool of conformers using UFF force field; (2) Minimize conformers; (3) Prune conformers using an RMSD threshold.

As shown in the revised **Extended Table 1** (the Revised Supplemental Table 27), we found that the video generation source has no significant impact on VideoMol with an average performance of 0.755±0.068 (Openbabel), 0.755±0.072 (DeepChem), 0.742±0.064 (RDKit) in RMSE metric and 0.581±0.057 (Openbabel), 0.576±0.060 (DeepChem), 0.565±0.053 (RDKit) performance in MAE metric. Therefore, VideoMol has low sensitivity to video generation sources.

Extended Table 1 (the Revised Supplemental Table 27). The performance of different video generation source on 10 kinases datasets with balanced scaffold split.

	1. 5HT1A		2. 5HT2A		3. AA1R	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Openbabel	0.733±0.009	0.562±0.006	0.796±0.004	0.597±0.003	0.666±0.014	0.510±0.011
DeepChem	0.718±0.006	0.557±0.009	0.790±0.016	0.595±0.019	0.677±0.014	0.504±0.010
RDKit	**0.708±0.017**	**0.547±0.015**	**0.775±0.017**	**0.577±0.009**	**0.655±0.007**	**0.496±0.006**
—— continue ——						
	4. AA2AR		5. AA3R		6. CNR2	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Openbabel	0.705±0.004	0.548±0.008	**0.779±0.015**	0.618±0.009	0.903±0.018	0.714±0.012
DeepChem	**0.703±0.005**	0.558±0.004	0.798±0.019	0.620±0.015	0.896±0.013	0.708±0.012
RDKit	0.712±0.011	**0.543±0.005**	0.786±0.006	**0.617±0.004**	**0.864±0.005**	**0.679±0.010**
—— continue ——						
	7. DRD2		8. DRD3		9. HRH3	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Openbabel	0.763±0.018	0.562±0.009	0.745±0.005	0.580±0.003	0.679±0.004	0.524±0.004
DeepChem	0.745±0.005	**0.551±0.002**	0.728±0.014	0.556±0.008	**0.666±0.005**	**0.505±0.002**
RDKit	**0.742±0.004**	0.556±0.005	**0.715±0.014**	**0.554±0.012**	0.668±0.008	0.506±0.002
—— continue ——						
	10. OPRM		Mean			
	RMSE	MAE	RMSE	MAE		
Openbabel	**0.776±0.008**	0.590±0.001	0.755±0.068	0.581±0.057		
DeepChem	0.825±0.008	0.607±0.009	0.755±0.072	0.576±0.060		
RDKit	0.795±0.015	**0.579±0.011**	**0.742±0.064**	**0.565±0.053**		

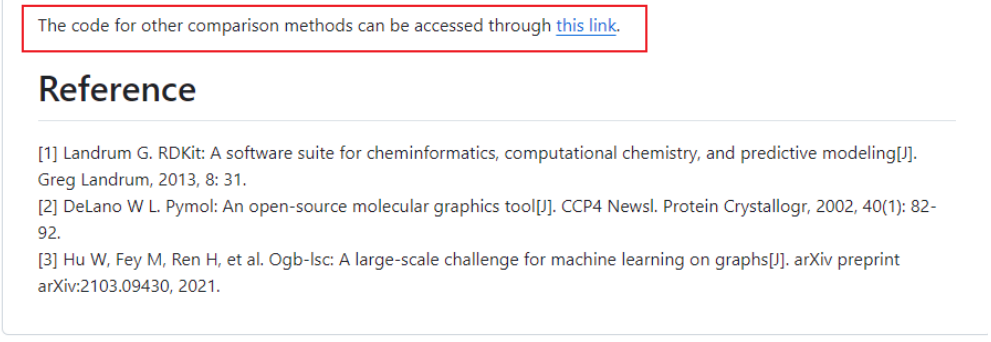
References

[1] O'Boyle N M, Banck M, James C A, et al. Open Babel: An open chemical toolbox[J]. *Journal of Cheminformatics*, 2011, 3: 1-14. URL: https://github.com/openbabel/openbabel

[2] Altae-Tran H, Ramsundar B, Pappu A S, et al. Low data drug discovery

	with one-shot learning[J]. *ACS Central Science*, 2017, 3(4): 283-293. URL: https://github.com/deepchem/deepchem
Excerpt from Revised Manuscript	Ablation study: *Sensitivity of VideoMol for video generation source.* To verify the sensitivity of VideoMol to video generation sources, we used two additional platforms to generate molecular videos, which are OpenBabel[58] and DeepChem[59]. We found that the video generation source of different platforms has no significant impact on VideoMol with an average performance of 0.755±0.068 (Openbabel), 0.755±0.072 (DeepChem), 0.750±0.065 (RDKit) in RMSE metric and 0.581±0.057 (Openbabel), 0.576±0.060 (DeepChem), 0.572±0.056 (RDKit) in MAE metric (Supplementary Table 27). Therefore, VideoMol has low sensitivity to video generation sources from different platforms. References [58] O'Boyle, N.M. et al. Open Babel: An open chemical toolbox. 3, 1-14 (2011). [59] 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).

Comments on code availability **– "**The provided code is **clean and well organized" –**

Reviewer Comment	The provided code is clean and well organized. The authors also included a docker image for setting up the environment. The code for training the model as well as reproducing the results is included. I believe only the code for VideoMol is provided, and not for the baseline methods. Not necessary, but if it is easy to include the related methods, the community might appreciate a full testbed.
Author Response	We thank the reviewer for checking our codes. We have provided codes for the baseline methods and related methods/models as well in this link (*https://1drv.ms/f/s!Atau0ecyBQNTgTd736-8RPWEXSVt?e=DkOyw2*). Of course, you can also access it via our github repository (*https://github.com/ChengF-Lab/VideoMol*), as shown below:  The code for other comparison methods can be accessed through this link. Reference [1] Landrum G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling[J]. Greg Landrum, 2013, 8: 31. [2] DeLano W L. Pymol: An open-source molecular graphics tool[J]. CCP4 Newsl. Protein Crystallogr, 2002, 40(1): 82-92. [3] Hu W, Fey M, Ren H, et al. Ogb-lsc: A large-scale challenge for machine learning on graphs[J]. arXiv preprint arXiv:2103.09430, 2021.

Responses to the Reviewer #3

Overall Summary – "The work is **valuable** to the field in multiple areas**" –**

Reviewer Comment	The manuscript presents a video-based pretrained model that can be used to make downstream task predictions across multiple tasks with finetuning. The results are an improvement upon authors' previous work, ImageMol. The improvement is tied to use of video, which can be considered as an augmentation to static images, as well as the use of more comprehensive fingerprints that provide chemical, pharmacological and physicochemistry information. The authors show, through extensive testing, that the new model outperforms the previous and is at least as good or as better as some SOTA models for different tasks. Impact to field: The work is **valuable to the field in multiple areas**: it is a demonstration of technology transfer from video representation learning. It shows new self supervision tasks that are meaningful for molecule structure videos.
Author Response	We thank the Reviewer for great summary and his/her support on the important value of our proposed VideoMol in multiple drug discovery tasks.

Ref 2.1 – "More explanations about dynamics" –

Reviewer Comment	Unlike stated, the model does not capture a dynamic conformation of the molecule. The videos are not generated to represent any physical dynamics, or conformer change, or changes to torsion angles etc. They are movies with standardized rotations around given axis. As such, they are only augmentations to enrich the model input about the 3D structure of the molecule. Authors should consider another wording than dynamics to prevent misleading the reader.
Author Response	We thank the reviewer for these critiques. We agreed with the reviewer that the current ViodeMol framework cannot capture the physical dynamics or conformational changes of ligand-receptor dynamics. Integrating physical dynamics or conformational changes from 3D ligand-receptor structures or models (i.e., alphaFold3 [1]) may improve performance of ViodeMol in the future. We have changed our original claims and added more explanations in the revised manuscript. Reference [1] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024: 1-3.
Excerpt from Revised Manuscript	Results: **Framework of VideoMol** Molecules exist in nature and are constantly conformational dynamics,

making video the most direct representation method. The molecular 3D information can be directly observed from the video without the help of manual feature extraction, such as the distance between pairs of atoms and the angle formed between multiple atoms and so on. In addition, we evaluated the advantages of different representations in feature extraction capabilities and found that our proposed video representation has obvious advantages over existing representations with a 66% improvement rate on 8 basic attributes (Supplementary Section C.2 and Supplementary Table 1). Therefore, these significant differences motivate us to develop VideoMol for accurately predicting the targets and properties of molecules in the form of videos derived from molecules.

C.2 Results of different representations on 8 basic attributes

To fairly compare the effects of different representations, we evaluated the representation without using any self-supervised tasks. It is well known that the development of drug discovery depends on accurately capturing chemical and biological representations of molecules. Here, we used several commonly used representative methods (such as GCN, GIN, EGNN, and the representation used by ImageMol) to inspect the model's ability to understand the 8 basic attributes of molecules, including molecular weight, MolLogP, MolMR, BalabanJ, NumHAcceptors, NumHDonors, NumValenceElectrons and TPSA.
We randomly collected 10,000 molecules from the pre-training dataset and used exactly the same experimental setup for fair comparison. In detail, we split the training set, validation set, and test set using a ratio of 8:1:1 and reported the results on the test set based on the best validation set score. As shown in Supplementary Table 1, we found that VideoMol using only one frame outperformed that of the 2D graph-based methods, the 3D-based graph method and the 2D image-based method, revealing the advantage of 3D representation. Specifically, compared with the second-place ImageMol without pre-training, the performance of video-1frame improved by 11%. When we utilized all video frames (video-60frame), the performance is further significantly improved from 12.47 to 7.55 with a 66% improvement rate.
In summary, the proposed 3D representation (whether based on a single frame image or a 60-frame video) has advantages compared to existing molecular representation approaches. We will further improve our VideoMol framework by inceasing the number of 3D frames and integrating other types of 3D representation (such as AlphaFold311) in the near future.

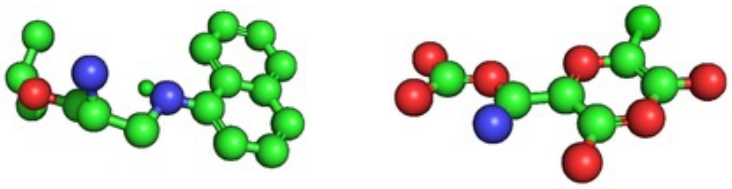
[11] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024: 1-3.

Discussion
Using a simple extension to VideoMol, we can allow the model to learn the correlations and variances between different conformations in the same molecule from videos of dynamic changes, thereby further playing an important role in molecular dynamics scenarios.
We believe that it is promising to represent molecules and perform inferences through videos as molecular imaging techniques continue to

	advance. In summary, the introduction of VideoMol on the one hand enriches the form of molecular representation in the field of computational drug discovery, and on the other hand inspires people to learn and understand the molecules from different perspectives.

Ref 2.2 – "Equivariant graph neural networks or transformers can be used to eliminate the need to perform augmentation for different rotations" –

Reviewer Comment	If the manuscript's main aim was to inject more information about the 3d nature of the molecules, they could have considered an equivariant graph neural network or transformer. An equivariant neural network would remove the need to perform augmentation for different rotations.
Author Response	We thank the reviewer for this valuable point. We agreed with reviewer that equivariant graph neural networks or transformers can indeed inject more information about the 3D nature of molecules by learning on well-characterized rotation-independent features. However, video and graph are complementary and they each have unique advantages in data presentation and analysis. There are several essential differences between VideoMol and equivariant graph neural network in obtaining the 3D of molecules as reflected below: **(1) Modality representation**. Obviously, video and graph are completely different in representation. There are several advantages for choosing video for molecular representation. First, video is a more intuitive representation method and information related to 3D nature can be directly observed from the videos, which allows the model to learn the information of bioactive molecules directly from the video without the help of any manual feature extraction. Secondly, VideoMol attempts the feasibility of learning representations from molecular videos and can be extended to learn a video related to potential ligand-receptor information by integrating with alphaFold3 or other available tools in the future. **(2) Feature extraction**. VideoMol extracts dense pixel-level features, while the graph-based model extracts relaxed node-level features. VideoMol allows the model to perceive 3D information in molecules by learning local textures in videos, such as the distance between pairs of atoms and the angle formed between three atoms and so on (as shown in the below **Extended Figure 1**). In contrast, 3D graph-based methods require the intervention of explicit knowledge to guide the model to learn this information. These significant differences motivate us to develop a deep learning method based on video representations and we can also see significant advantages of VideoMol as we demonstrated in multiple drug discovery tasks. We have added these new explanations in the revised manuscript (Section Motivation for using video representations).

	 **Extended Figure 1.** Visualization of a certain frame in a video.

Ref 2.3 – "Design experiment to understand why video information does not lead to sensitivity to conformer" –

Reviewer Comment	This is where it gets interesting: if the success of the model was truly due to better representation of 3d structure, we would expect the model to be sensitive to different conformers, especially on tasks that provide a binding affinity proxy. While, in multiple places in the manuscript the opposite is claimed, that model is robust to molecule conformer choice. Perhaps authors can devise an experiment to understand why video information does not lead to sensitivity to conformer.
Author Response	We agreed with the reviewer that the experimental section of "*Robustness of VideoMol"* is confused because we evaluated the performance of different models on different conformers, which does not reflect the robustness of a single model to different conformations. Therefore, we designed new experiments and evaluated the sensitivity of the same VideoMol model to different conformations. Specifically, we directly used pre-trained VideoMol to extract features of molecules with different conformers from 10 kinases datasets of binding affinity profiles and compared the similarities between different videos with different conformers. Since the similarity between conformers is related to their RMSD (Root-Mean-Square Deviation) distance, we also calculated the similarity of features in different RMSD intervals. As shown in the **Extended Table 1** (the Revised Supplemental Table 28) below, we found that VideoMol was discriminative for videos from different conformers. Further, when the RMSD between two conformations is larger, the feature similarity extracted by VideoMol shows a decreasing trend. Especially in the 90-100 percentile range, the feature similarity extracted by VideoMol is always the lowest. Therefore, VideoMol is sensitive to different conformers. We have added these new results and more detailed explanations in the revised manuscript.

Extended Table 1 (the Revised Supplemental Table 28). The ability of VideoMol to distinguish different conformers. The percentile interval refers to sorting all RMSD values from small to large and selecting the value corresponding to the percentile interval.

Percentile interval in RMSD	5HT1A	5HT2A	AA1R	AA2AR	AA3R
0-10	0.692	0.749	0.799	0.803	0.823
10-20	0.724	0.726	0.786	0.784	0.771
20-30	0.741	0.747	0.755	0.764	0.772
30-40	0.726	0.733	0.749	0.747	0.752
40-50	0.735	0.745	0.759	0.772	0.758
50-60	0.703	0.744	0.765	0.761	0.760
60-70	0.708	0.733	0.742	0.756	0.745
70-80	0.663	0.748	0.753	0.743	0.766
80-90	0.631	0.702	0.745	0.753	0.752
90-100	0.518	0.625	0.654	0.702	0.733
0-100 (all data)	0.684	0.725	0.751	0.758	0.763
====== continue ======					
Percentile interval in RMSD	CNR2	DRD2	DRD3	HRH3	OPRM
0-10	0.769	0.729	0.701	0.740	0.730
10-20	0.747	0.733	0.740	0.721	0.740
20-30	0.736	0.750	0.779	0.690	0.708
30-40	0.735	0.748	0.795	0.703	0.716
40-50	0.723	0.760	0.775	0.709	0.712
50-60	0.710	0.743	0.785	0.706	0.725
60-70	0.696	0.748	0.777	0.698	0.727
70-80	0.712	0.715	0.775	0.700	0.716
80-90	0.709	0.707	0.723	0.669	0.680
90-100	0.680	0.615	0.631	0.585	0.658
0-100 (all data)	0.722	0.725	0.748	0.692	0.711

Excerpt from Revised Manuscript

Results:

VideoMol captures conformational differences of molecules. We used pre-trained VideoMol to extract features of molecules with different conformers from 10 compound-kinase interaction datasets and compared the cosine similarities between different videos with different conformers. Since the similarity between conformers is related to their RMSD (Root-Mean-Square Deviation) distance, we also calculated the similarity of features in different RMSD intervals. We found that VideoMol is discriminative for videos from different conformers (Supplementary Table 28). Further, when the RMSD between two conformations is larger, the feature similarity extracted by VideoMol shows a decreasing trend. Especially in the 90-100 percentile range, the feature similarity extracted by VideoMol is always the lowest. Therefore, VideoMol can effectively capture conformational differences of molecules.

Ref 2.4 – "Add more discussion about understanding from where exactly the accuracy improvement comes from" –

Reviewer Comment	Which makes me think the success of the model is not due to better 3d representation but one of the several other changes: 1-working with video frames and the new self-supervision tasks have expanded the effective size of the data and complexity the network processes each molecule (perhaps this is why the number of frames seem to change the prediction accuracy) 2-the large number of domain information that is crafted into the fingerprint may be impactful in several tasks in this work. Further understanding from where exactly the accuracy improvement comes from, can be considered for future work. In the teamtime, the claims of impact of 3D could be dialed down.
Author Response	We thank the Reviewer for these excellent points. We agree that the success achieved by VideoMol may be related to the introduction of self-supervised tasks and fingerprints. In **Supplementary Table 24**, we perform ablation experiments on different self-supervised pre-training tasks, and we find that each pre-training task can promote the improvement of VideoMol performance.

Supplementary Table 24: Effect of pre-training strategy on 6 regression datasets with balanced scaffold split. w/o pretrain means no pre-trained VideoMol. video-aware, direction-aware, and chemical-aware represent pre-training VideoMol using only video-aware strategy, direction-aware strategy, and chemical-aware strategy, respectively. & represents the combination of multiple pre-training tasks. All means and standard deviations are reported through three independent runs.

	5HT1A		AA1R		AA2AR	
strategy	RMSE	MAE	RMSE	MAE	RMSE	MAE
w/o pretrain	0.993±0.001	0.804±0.002	0.919±0.006	0.755±0.008	1.073±0.01	0.882±0.013
video-aware	0.772±0.011	0.604±0.009	0.709±0.007	0.783±0.010	0.847±0.002	0.651±0.002
direction-aware	0.871±0.002	0.691±0.004	0.821±0.01	0.652±0.013	0.875±0.008	0.701±0.012
chemical-aware	0.736±0.012	0.566±0.014	0.662±0.018	**0.494±0.015**	0.716±0.001	0.562±0.002
chemical_direction	0.718±0.002	0.559±0.004	0.706±0.005	0.516±0.008	0.724±0.008	0.568±0.003
chemical_video	0.716±0.010	0.553±0.010	0.670±0.007	0.511±0.005	0.719±0.013	0.553±0.014
direction_video	0.774±0.014	0.603±0.015	0.730±0.004	0.553±0.002	0.865±0.011	0.672±0.005
VideoMol	**0.708±0.017**	**0.547±0.015**	**0.655±0.007**	0.496±0.006	**0.712±0.011**	**0.543±0.005**
	CNR2		**DRD2**		**HRH3**	
strategy	RMSE	MAE	RMSE	MAE	RMSE	MAE
w/o pretrain	1.216±0.012	1.009±0.008	0.980±0.001	0.782±0.001	0.819±0.001	0.631±0.002
video-aware	0.978±0.025	0.782±0.010	0.818±0.009	0.602±0.009	0.732±0.008	0.556±0.009
direction-aware	1.024±0.01	0.842±0.008	0.903±0.010	0.700±0.009	0.759±0.005	0.575±0.007
chemical-aware	0.890±0.004	0.698±0.002	0.759±0.003	0.565±0.003	0.669±0.010	0.512±0.010
chemical_direction	0.899±0.008	0.701±0.005	0.773±0.017	0.579±0.007	0.683±0.001	0.514±0.001
chemical_video	0.874±0.012	0.686±0.013	0.745±0.012	**0.555±0.013**	0.686±0.003	0.526±0.004
direction_video	0.997±0.025	0.789±0.021	0.832±0.005	0.615±0.007	0.734±0.006	0.559±0.002
VideoMol	**0.864±0.005**	**0.679±0.010**	**0.742±0.004**	0.556±0.005	**0.668±0.008**	**0.506±0.002**

Furthermore, to address the reviewer's concerns about the advantages of

3D representation, we evaluated the representation advantages of 3D-based molecular videos without using any self-supervised tasks and fingerprint information. It is well known that the development of drug discovery depends on the understanding of basic information about molecules. Here, we use several of the most representative methods (such as GCN, GIN, EGNN, and the representation used by ImageMol) to inspect the model's ability to understand the 8 basic attributes of molecules, including molecular weight, MolLogP, MolMR, BalabanJ, NumHAcceptors, NumHDonors, NumValenceElectrons and TPSA.

We randomly collected 10,000 molecules from the pre-training dataset and use exactly the same experimental setup for a fair comparison. In detail, we split the training set, validation set, and test set using a ratio of 8:1:1 and report the results on the test set based on the best validation set score. As shown in **Extended Table 1** (the Revised Supplemental Table 1) below, we found that VideoMol based on only one frame outperformed traditional 2D graph-based methods, the 3D-based graph method and the 2D image-based method, revealing the advantage of 3D representation. Specifically, compared with the second-place ImageMol without pre-training, the performance of video-1frame improved by 11%. When we utilized all video frames (video-60frame), the performance was further significantly improved from 12.469 to 7.55 with a 66% improvement rate.

Overall, the proposed 3D representation (whether based on a single frame image or a 60-frame video) has obvious advantages over existing representations. In addition, integrating physical dynamics or conformational changes from 3D ligand-receptor structures or models (i.e., alphaFold3 [1]) may improve performance of ViodeMol further. We have added these new results and more detailed explanations in the revised manuscript.

Extended Table 1 (the Revised Supplemental Table 1). The ability of VideoMol to distinguish different conformers. The percentile interval refers to sorting all RMSD values from small to large and selecting the value corresponding to the percentile interval.

	modality	model	use conformer?	**prop**
graph-based	2D graph	GCN	×	62.304
		GIN	×	62.980
		EGNN	×	17.418
	3D graph	EGNN	√	16.684
image-based	image from imagemol	ResNet18	×	12.469
	video-1frame	ResNet18	√	11.237
	video-5frame	ResNet18	√	8.088
	video-60frame	ViT	√	**7.511**

Reference

[1] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3[J]. Nature, 2024: 1-3.

Ref 2.5 – "This difference may be due to data leak in the high ROC-AUC train-test data, inflating the apparent generalizability" –

Reviewer Comment	Feedback to the methodology: the splits used in this work would not stop data leaking from train to validation sets and scaffold balancing might not be enough. Indeed, we see a hint of the issue in the COX examples where training data from ChEMBL in 8:1:1 split gave high ROC-AUC>0.9, but when the model was tested against MedChemExpress data, only less than 40% of inhibitors are successfully identified. This difference may be due to data leak in the high ROC-AUC train-test data, inflating the apparent generalizability. In general, if authors would like to claim generalizabilty, more attention to the split strategy, overlap between data points is needed according to certain similarity metric will be needed.
Author Response	We carefully examined COX-1 and COX-2 datasets from ChEMBL. We confirmed that the training, validation, and test sets did not contain any overlapping molecules, **without ant data leaking issue**. In the actual virtual screening or new drug development process (termed external validation set), the performance value will not be very high due to several common factors, such as low similarity or activity cliff between the training data and the drugs in the external validation sets to be virtually screened. We confirmed that there was no data leak issues and we have added more explanations in the revised manuscript.

Ref 2.6 – "Describe the computational requirements for pre-training and various downstream tasks" –

Reviewer Comment	Some of the tasks in the work are for high-throughout applications (e.g. virtual screening). In such cases the trade-off between accuracy and compute becomes important. The proposed method should clearly state the compute needs for pretraining and various downstream tasks. Because it works with video, compared to much smaller atomic position files, memory needs should be highlighted too.
Author Response	We have evaluated the computational requirements of the proposed VideoMol in pre-training, fine-tuning, and virtual screening stages. We found that VideoMol overall is highly cost-effective and time-effective in multiple tasks as discussed as below. We have added more detailed explanations about the computational requirements of the proposed VideoMol in pre-training, fine-tuning, and virtual screening stages in the revised manuscript.
Excerpt from Revised Manuscript	**C.3 Computational requirements of VideoMol** Here, we detail the computational requirements for the pre-training, fine-tuning, and virtual screening stages in **Supplementary Table 36**. In the pre-training phase (**Supplementary Table 36a**), VideoMol uses 256 frames in each batch for training and it requires 37G of GPU memory and takes about 9 hours to complete 1 epoch on 2 million molecular videos with 60 frames.

	Next, we used 10,000 molecules to test the impact of different batch sizes (#frame/batch) on memory and training speed in **Supplementary Table 36b**. We find that fine-tuning does not occupy a large amount of GPU memory. Finally, we evaluate the computational requirements when performing virtual screening on 1 million molecules in **Supplementary Table 36c**. We find that it only takes 9 hours when using all frames during virtual screening for 1 million molecules. **Supplementary Table 36**: The computational requirements in the pre-training, fine-tuning, and screening stages. #frame/batch represents the number of frames in a batch. #samples represent the total number of molecules. #frame/video indicate how many frames of a video to select for inference.

(a) The computational requirements in the pre-training stage.

#samples	#frame/batch	GPU memory	Training time	Server
2 million	256	~37G	~9 hours/epoch	CPU: Intel 6248R 48C@3.0GHz; GPU: A100 (40G)

(b) The computational requirements in the fine-tuning stage.

#samples	#frame/batch	GPU memory	Training time	Server
10,000	8	2.3G	~26 minutes/epoch	CPU: 13th Gen Intel® Core™ i7-13700K GPU: 4090 Ti
	16	2.6G	~15 minutes/epoch	
	32	3.2G	~12 minutes/epoch	
	64	4.3G	~12 minutes/epoch	
	128	6.5G	~12 minutes/epoch	
	256	10.7G	~12 minutes/epoch	

(c) The computational requirements in the screening stage.

#samples	#frame/video	#videos/batch	inference time	GPU memory	Server
1 million	1	480	~9 minutes	17.7 G	CPU: 13th Gen Intel® Core™ i7-13700K GPU: 4090 Ti
	5	96	~48 minutes		
	10	48	~90 minutes		
	20	24	~3 hours		
	30	16	~4.5 hours		
	60	8	~9 hours		

Ref 2.7 – "Correct Minor typos" –

Reviewer Comment	minor typos in text, highlighting one that is on figure in case it escapes proofreading: angel -> angle Fig 1b
Author Response	We have fixed this typo and further polish English of the entire manuscript.

REVIEWERS' COMMENTS

Reviewer #2 (Remarks to the Author):

I have carefully reviewed the detailed rebuttal and the revised manuscript. I am pleased to see that the authors have addressed all of the concerns I raised in my initial review. The additional experiments and clarifications provided have significantly strengthened the support for the claims. I am satisfied with the revisions and believe that the manuscript is now ready for publishing.

Reviewer #2 (Remarks on code availability):

The code is well organised, and the authors made the effort to provide guidelines for setting up an environment and running VideoMol.

Reviewer #4 (Remarks to the Author):

I have assessed the comments from the authors to Reviewers 3 concerns, and I believe the authors has responded well to the comments