



AccMLBio @ ICML'24

# MiniMol: A Parameter-Efficient Foundation Model for Molecular Learning

Kerstin Klaser\*, Błażej Banaszewski\*,  
Callum McLean, Andrew Fitzgibbon  
Graphcore

Samuel Maddrell-Mander  
Dayhoff Labs

Ali Parviz  
NJIT, Mila

Luis Müller  
RWTH Aachen University

Shenyang (Andy) Huang  
Mila

# MOTIVATION

## Foundation models

Properties	Large Multimodal Models	Molecular Models
Scaling	Large scale	Transformers are data hungry; GNNs are difficult to scale
Representation	Sequence models	No consensus on inductive bias for biological tasks
Data modality	Text and images	Synergies not clear
Data quantity	Organically generated content via social interactions	Biological data is scarce and expensive
Training objective	Self-supervised; Masked modelling	Only limited generalizability of mask modelling training demonstrated

# MOTIVATION

## Foundation models

Properties	Large Multimodal Models	Molecular Models
Scaling	Large scale	Transformers are data hungry; GNNs are difficult to scale
Representation	Sequence models	No consensus on inductive bias for biological tasks
Data modality	Text and images	Synergies not clear
Data quantity	Organically generated content via social interactions	Biological data is scarce and expensive
Training objective	Self-supervised; Masked modelling	Only limited generalizability of mask modelling training demonstrated

# CONTRIBUTIONS

## Foundation Molecular Models

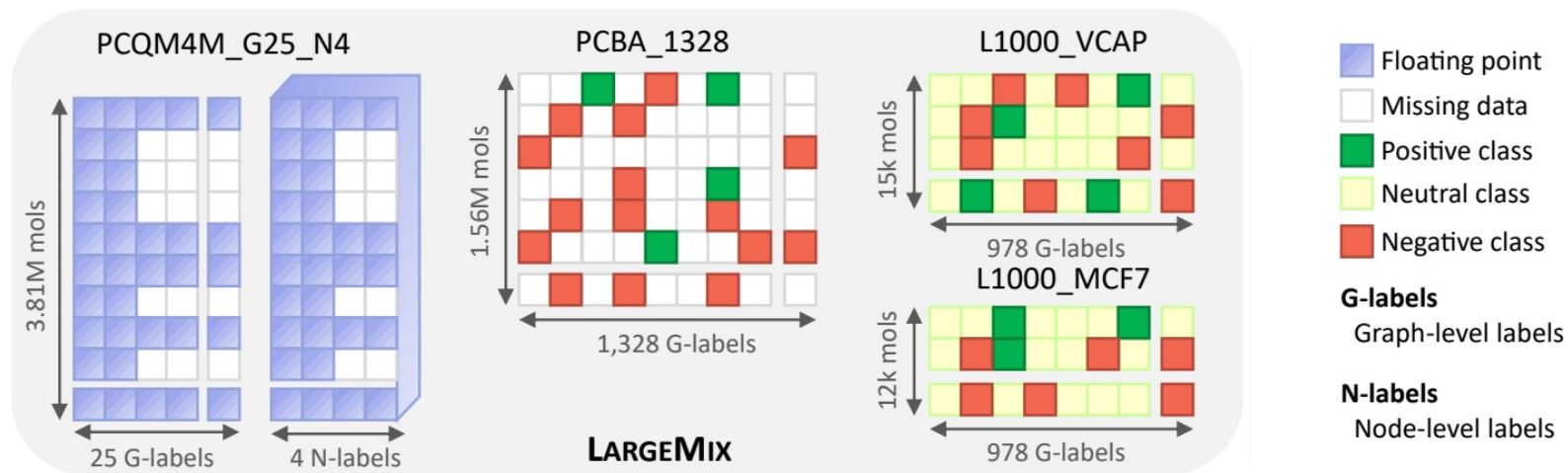
Problems	Solutions	
	Graphium	MiniMol
Transformers are data hungry; GNNs are difficult to scale	Software infrastructure for pre-training GNNs	Parameter-efficiency
No consensus on inductive bias for biological tasks	Molecules as graphs	Molecular fingerprinting for downstream transfer
Synergies not clear	Biological and quantum	Dataset correlation analysis
Biological data is scarce and expensive	Introduction of curated datasets for molecular machine learning	/
Only limited generalizability of mask modelling training demonstrated	Multi-task, multi-level supervised learning	/

# CONTRIBUTIONS

## Foundation Molecular Models

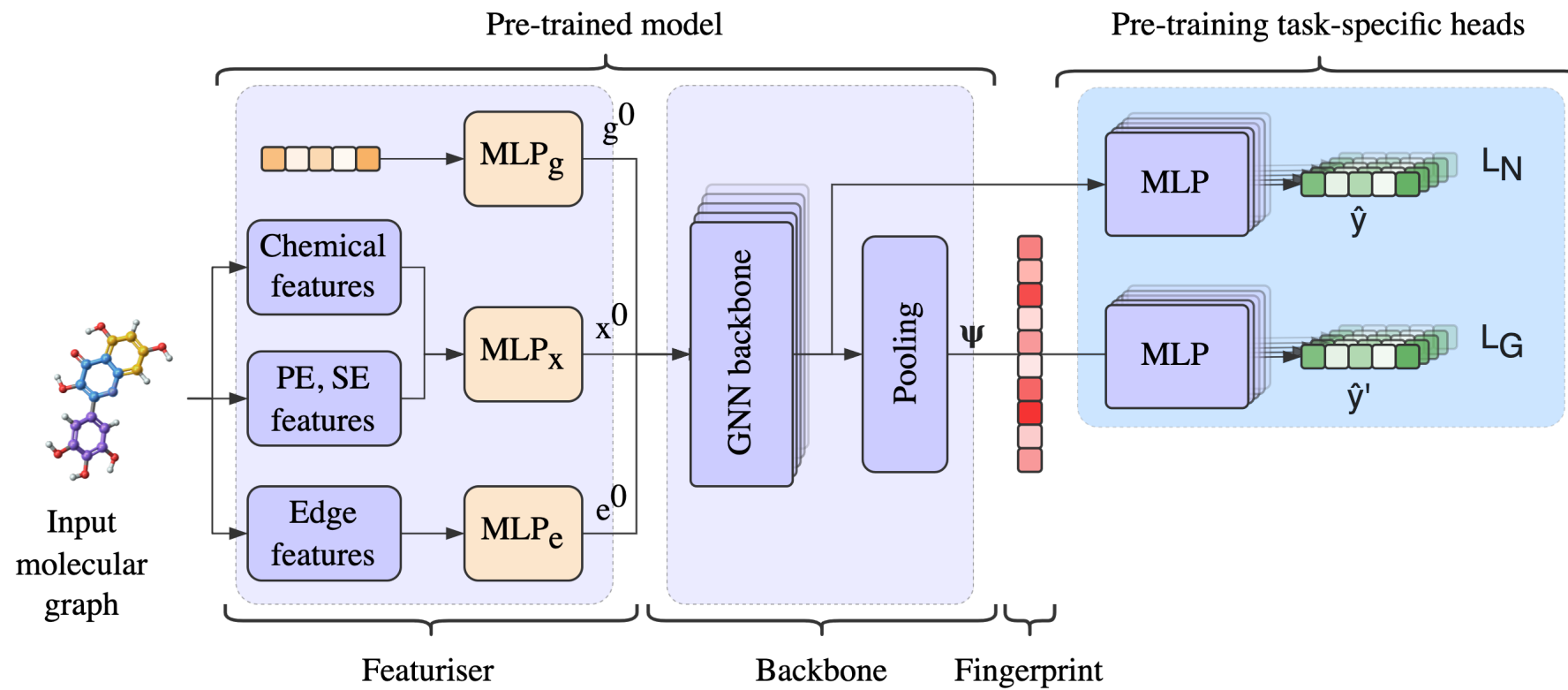
Problems	Solutions	
	Graphium	MiniMol
Transformers are data hungry; GNNs are difficult to scale	Software infrastructure for pre-training GNNs	Parameter-efficiency
No consensus on inductive bias for biological tasks	Molecules as graphs	Molecular fingerprinting for downstream transfer
Synergies not clear	Biological and quantum	Dataset correlation analysis
Biological data is scarce and expensive	Introduction of curated datasets for molecular machine learning	/
Only limited generalizability of mask modelling training demonstrated	Multi-task, multi-level supervised learning	/

# PRE-TRAINING DATA

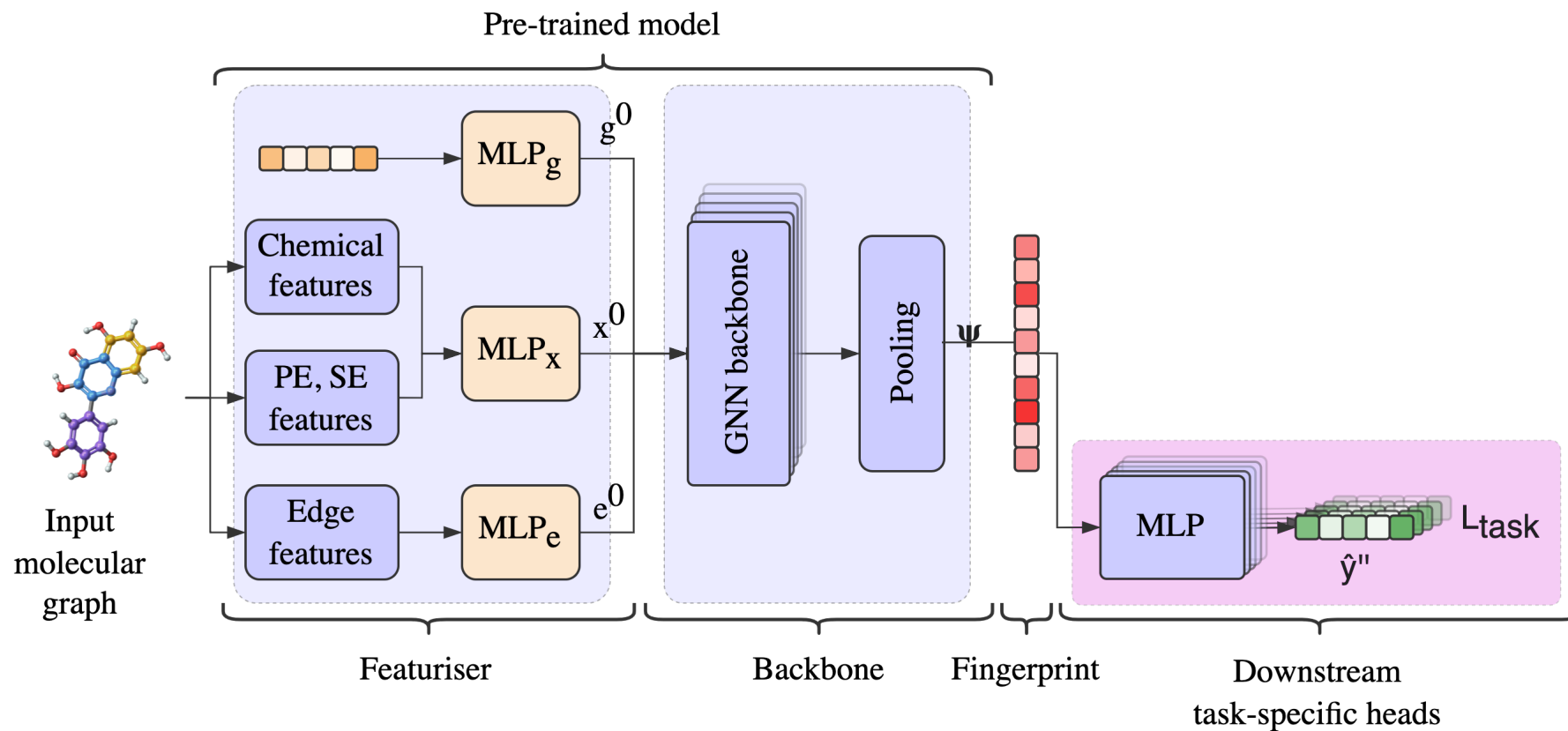


Dataset	# Molecules	# Labels	# Data Points	% of All Data Points
PCQM4M_G25	3.81M	25 (G)	93M	17%
PCQM4M_N4	3.81M	4 (N)	197.7M	37%
PCBA_132B	1.56M	1328 (G)	224.4M	41%
L1000_VCAP	15K	978 (G)	15M	3%
L1000_MCF7	12K	978 (G)	11M	2%

# MINIMOL PRE-TRAINING



# MINIMOL FOR DOWNSTREAM TASKS





# MINIMOL ON TDC ADMET

Table 2. Results on downstream evaluation of MiniMol (GINE) with max pooling (see Appendix A.4 for pooling experiments) on TDC ADMET benchmarks, and comparison to the TDC leaderboard and MoE. The rank is determined for each dataset individually, on a set of 7 scores, which include the test results from the TOP5 leaderboard, MoE and MiniMol. The best result is shown in green and the top-3 results are highlighted in purple.

	TDC Dataset			Leaderboard	MoE		MiniMol (GINE)	
	Name	Size	Metric	SOTA Result	Result	Rank	Result	Rank
ABSORPTION	Caco2 Wang	906	MAE (↓)	<b>0.276 ± .005</b>	0.310 ± .010	6	0.324 ± .012	7
	Bioavailability Ma	640	AUROC (↑)	<b>0.748 ± .033</b>	0.654 ± .028	7	0.699 ± .008	6
	Lipophilicity AZ	4,200	MAE (↓)	<b>0.467 ± .006</b>	<b>0.469 ± .009</b>	3	<b>0.455 ± .001</b>	1
	Solubility AqSolDB	9,982	MAE (↓)	<b>0.761 ± .025</b>	0.792 ± .005	5	<b>0.750 ± .012</b>	1
	HIA Hou	578	AUROC (↑)	<b>0.989 ± .001</b>	0.963 ± .019	7	<b>0.994 ± .003</b>	1
	Pgp Broccatelli	1,212	AUROC (↑)	<b>0.938 ± .002</b>	0.915 ± .005	7	<b>0.994 ± .002</b>	1
DISTRIB.	BBB Martins	1,975	AUROC (↑)	<b>0.916 ± .001</b>	0.903 ± .005	7	<b>0.923 ± .002</b>	1
	PPBR AZ	1,797	MAE (↓)	<b>7.526 ± .106</b>	8.073 ± .335	6	7.807 ± .188	4
	VDss Lombardo	1,130	Spearman (↑)	<b>0.713 ± .007</b>	<b>0.654 ± .031</b>	3	0.570 ± .015	7
METABOLISM	CYP2C9 Veith	12,092	AUPRC (↑)	<b>0.859 ± .001</b>	0.801 ± .003	5	0.819 ± .001	4
	CYP2D6 Veith	13,130	AUPRC (↑)	<b>0.790 ± .001</b>	0.682 ± .008	6	0.718 ± .003	5
	CYP3A4 Veith	12,328	AUPRC (↑)	<b>0.916 ± .000</b>	0.867 ± .003	7	0.878 ± .001	5
	CYP2C9 Substrate	666	AUPRC (↑)	<b>0.441 ± .033</b>	<b>0.446 ± .062</b>	2	<b>0.481 ± .013</b>	1
	CYP2D6 Substrate	664	AUPRC (↑)	<b>0.736 ± .024</b>	0.699 ± .018	7	<b>0.726 ± .006</b>	2
	CYP3A4 Substrate	667	AUROC (↑)	<b>0.662 ± .031</b>	<b>0.670 ± .018</b>	1	0.644 ± .006	6
EXCRET.	Half Life Obach	667	Spearman (↑)	<b>0.562 ± .008</b>	<b>0.549 ± .024</b>	4	0.493 ± .002	7
	Clearance Hepatocyte	1,102	Spearman (↑)	<b>0.498 ± .009</b>	0.381 ± .038	7	<b>0.448 ± .006</b>	4
	Clearance Microsome	1,020	Spearman (↑)	<b>0.630 ± .010</b>	0.607 ± .027	6	<b>0.652 ± .007</b>	1
TOXICITY	LD50 Zhu	7,385	MAE (↓)	<b>0.552 ± .009</b>	0.823 ± .019	7	<b>0.588 ± .010</b>	3
	hERG	648	AUROC (↑)	<b>0.880 ± .002</b>	0.813 ± .009	7	0.849 ± .007	6
	Ames	7,255	AUROC (↑)	<b>0.871 ± .002</b>	<b>0.883 ± .005</b>	1	0.856 ± .001	5
	DILI	475	AUROC (↑)	<b>0.925 ± .005</b>	0.577 ± .021	7	<b>0.944 ± .007</b>	1
TDC Leaderboard Mean Rank:					5.2		<b>3.4</b>	

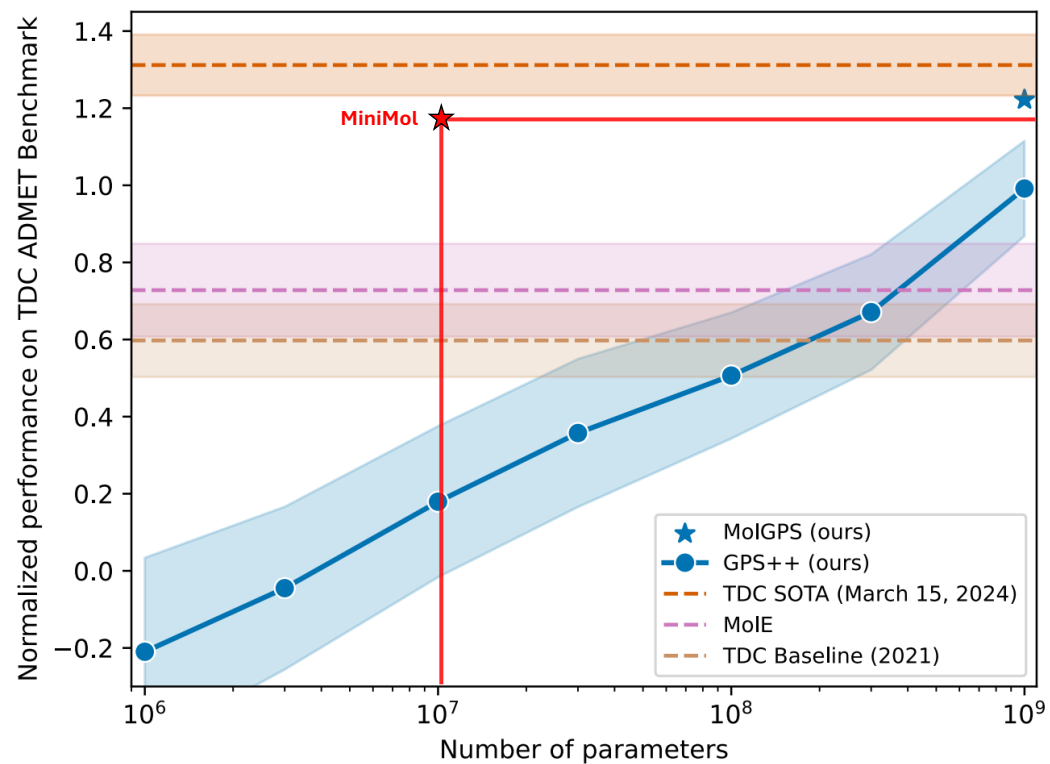
# PRE-TRAINING → DOWNSTREAM KNOWLEDGE TRANSFER

*Table 4.* Correlation analysis (Spearman’s rho) between pre-training validation and downstream performance. The green colour indicates a beneficial correlation and the red indicates a detrimental correlation. Results with a p-value over 0.1 are blank.

Dataset	Metric	MCF	VCAP	PCBA	G25	N4
		AUROC	AUROC	AUROC	MAE	MAE
Caco2 Wang	MAE	0.590	0.651	0.718		
Bioavailability Ma	AUROC					
Lipophilicity AZ	MAE	0.568	0.539	0.627	-0.389	
Solubility AqSolDB	MAE	0.588	0.7	0.704		
HIA Hou	AUROC	0.603	0.548	0.645	-0.337	
Pgp Broccatelli	AUROC		0.361		-0.387	
BBB Martins	AUROC	0.583	0.378	0.483	-0.492	
PPBR AZ	MAE					
VDss Lombardo	Spearman		0.343			
CYP2C9 Veith	AUPRC	0.649	0.711	0.829		0.551
CYP2D6 Veith	AUPRC	0.641	0.487	0.704		0.585
CYP3A4 Veith	AUPRC	0.649	0.713	0.818		0.608
CYP2C9 Subst.	AUPRC		-0.377	-0.445		-0.586
CYP2D6 Subst.	AUPRC					
CYP3A4 Subst.	AUROC		0.409			
Half Life Obach	Spearman		0.503	0.498		
Clearance Hepato.	Spearman					
Clearance Micro.	Spearman					
LD50 Zhu	MAE	0.543	0.522	0.617		0.342
hERG	AUROC		0.57	0.453		
AMES	AUROC	0.591	0.486	0.643	-0.628	0.528
DILI	AUROC	0.49	0.416	0.454		
Sum		6.496	7.959	7.749	-2.232	2.028

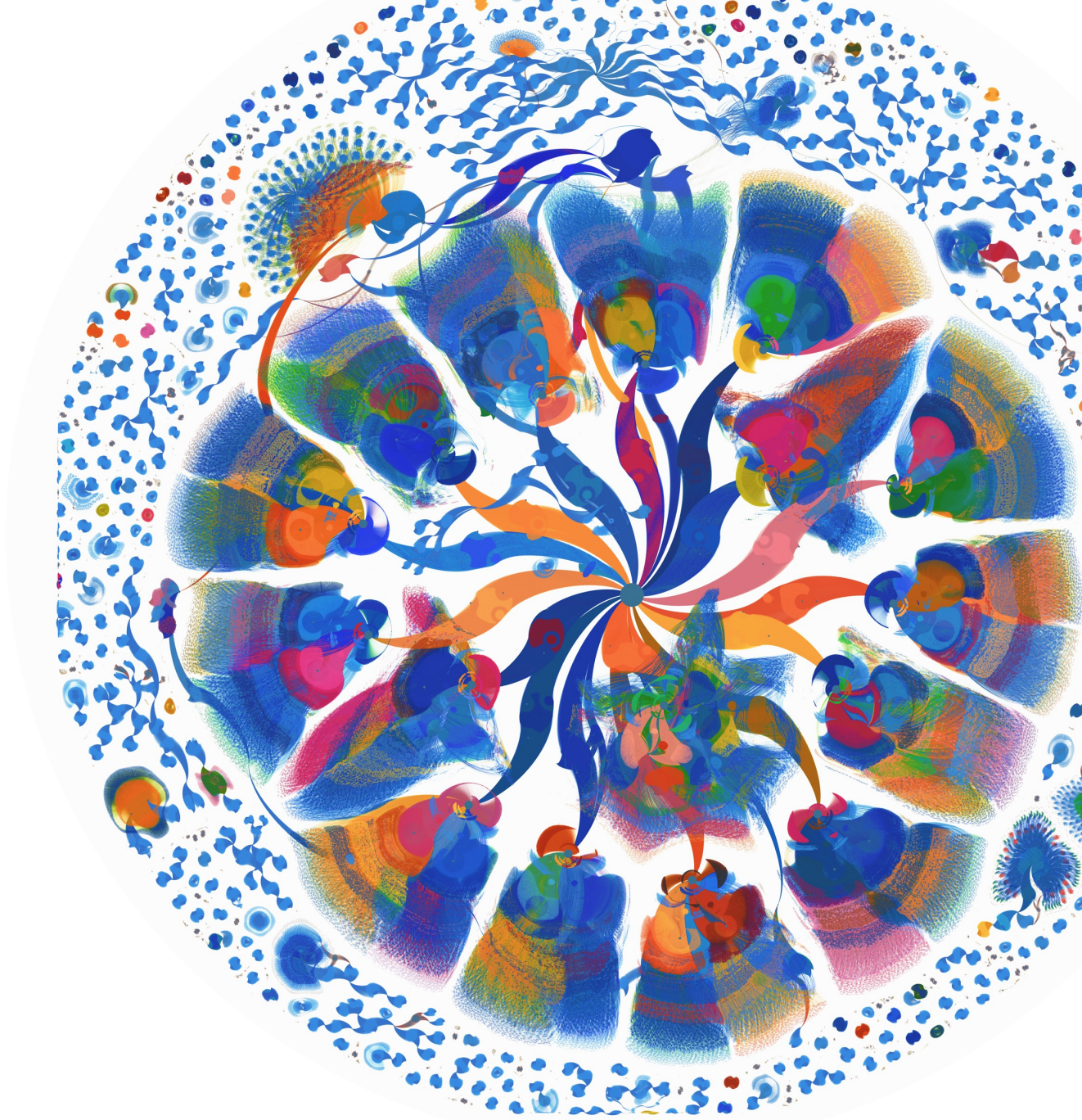
# MINIMOL COMPARISON

Model	Params(M)	Mean rank	>MoIE	TOP1	TOP3
MolGPS (Estimated/interpolated)	1000	N/A	18	12	N/A
MolE	100	5.4	=	N/A	N/A
MiniMol	10	3.4	17	8	11
AGBT (Chen et al., 2021a)	N/A	5.4	10	2	4
MolFormer (Ross et al., 2022)	N/A	5.6	7	0	5
BET (Chen et al., 2021b)	N/A	6.0	7	1	2



# CONCLUSIONS











- MiniMol is a novel parameter-efficient foundation model for molecular learning. It was pre-trained on over 3,300 biological and quantum tasks on both graph- and node-level features.
- MiniMol outperforms the previous state-of-the-art foundation model, MoE (Méndez-Lucio et al., 2022), on TDC ADMET, with only 10M parameters, 10× fewer than MoE.
- Training task-specific MLPs on MiniMol-generated fingerprints is an efficient way to transfer the knowledge.
- The correlation analysis gives insight into how to utilize pre-training datasets for downstream biological tasks.







# TRY IT YOURSELF!

 **minimol** Public

master Branches Tags  Add file Code

 blazejba	readme update	7101a87 · 3 hours ago	24 Commits
 .github/workflows	expand .gitignore		last week
 minimol	improved results on hia hou		last week
 .gitattributes	ckpts not in the install folder		last week
 .gitignore	expand .gitignore		last week
 LICENSE	license		last week
 README.md	readme update		3 hours ago
 env.yml	1.0 release		last week
 setup.cfg	Initial commit		2 weeks ago
 setup.py	improved results on hia hou		last week

 **README**  MIT license  

A parameter-efficient molecular featuriser that generalises well to biological tasks thanks to the effective pre-training on biological and quantum mechanical datasets.

The model has been introduced in the paper [MiniMol: A Parameter-Efficient Foundation Model for Molecular Learning](#), published in the ICML workshop on *Accessible and Efficient Foundation Models for Biological Discovery* in 2024.

Molecular fingerprinting with MiniMol

<https://github.com/graphcore-research/minimol/>

