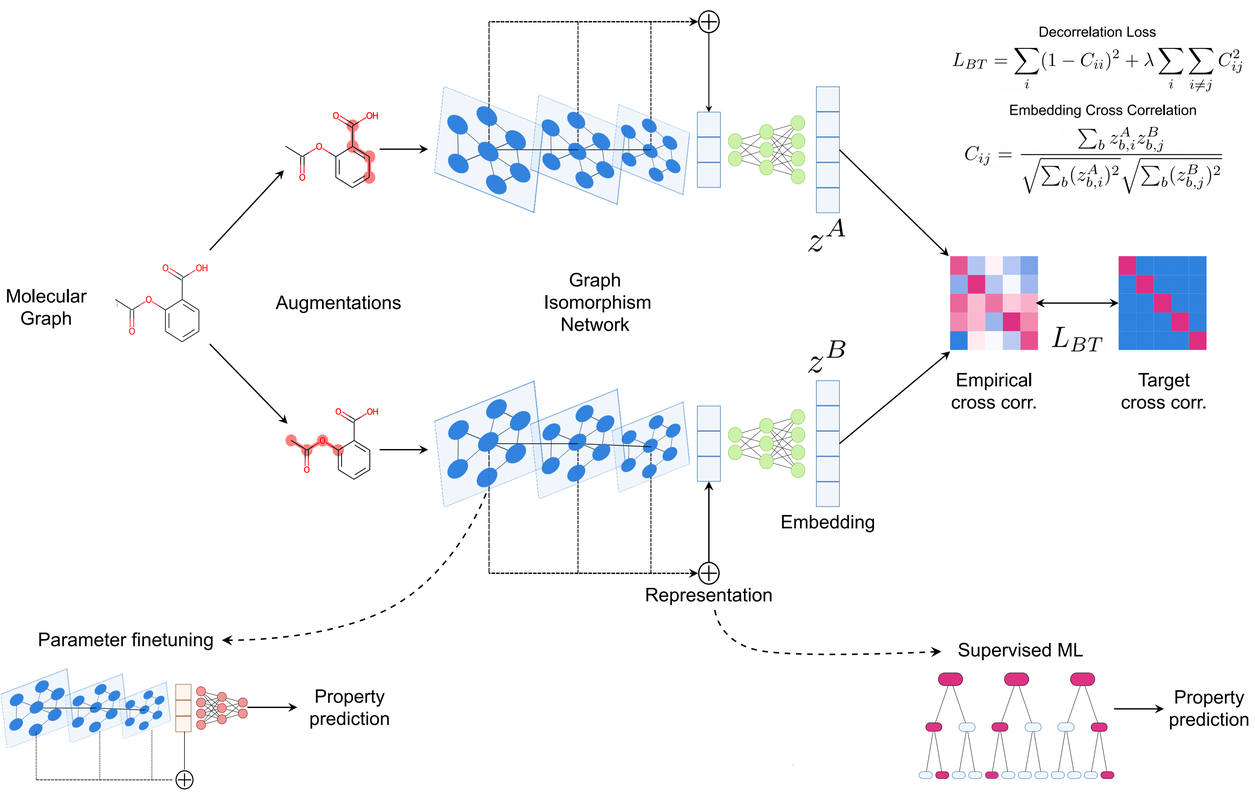
MolE: pre-trained molecular representations can aid antimicrobial discovery

Roberto Olayo-Alarcon[1,2], Martin K. Amstalden[3], Annamaria Zannoni[4], Cynthia Sharma[4], Ana Rita Brochado[3], Mina Rezaei[1], Christian L. Müller[1,2,5]

[1]Department of Statistics, Ludwig-Maximilians-Universität München, Munich, [2]Institute of Computational Biology, Helmholtz Munich, Munich, [3]Department of Molecular Infection Biology II, Institute of Molecular Infection Biology (IMIB), Julius-Maximilians-Universität Würzburg, Würzburg
[4]Department of Microbiology, Julius-Maximilians-Universität Würzburg, Biocenter, Würzburg, [5]Center for Computational Mathematics, Flatiron Institute, New York

The MolE pre-training framework

MolE (**Mol**ecular representation learned through **E**mbedding decorrelation) is a non-contrastive, self-supervised deep learning framework that aims at learning task-independent molecular representations by pre-training on unlabelled chemical structures.



MolE's representation can enhance the performance of supervised machine learning schemes such as XGBoost on various molecular property prediction tasks.
At the same time, fine-tuning the pre-trained parameters can improve the performance of GNN-based prediction schemes.

Performance on classification tasks
(Higher is better)

	BBBP	Tox21	SIDER	ClinTox	BACE	HIV
Supervised Learning						
GCN	71.8 ± 0.9	70.9 ± 2.6	53.6 ± 3.2	62.5 ± 2.8	71.6 ± 2	74 ± 3
GIN	65.8 ± 4.5	74 ± 0.8	57.3 ± 1.6	58 ± 4.4	70.1 ± 5.4	75.3 ± 1.9
SchNet	84.8 ± 2.2	**77.2 ± 2.3**	53.9 ± 3.7	71.5 ± 3.7	76.6 ± 1.1	70.2 ± 3.4
MGCN	**85 ± 6.4**	70.7 ± 1.6	55.2 ± 1.8	63.4 ± 4.2	73.4 ± 3	73.8 ± 1.6
D-MPNN	71.2 ± 3.8	68.9 ± 1.3	**63.2 ± 2.3**	**90.5 ± 5.3**	**85.3 ± 5.3**	**75 ± 2.1**
Static Features + XGBoost						
ECFP4	68.51 ± 1	70.76 ± 1.4	62.58 ± 1.8	84.52 ± 0.1	**84.33 ± 1.1**	75.77 ± 1.6
N-Gram	74 ± 0.1	73.51 ± 0.5	63.68 ± 1.5	**89.96 ± 1.2**	81.87 ± 0.1	**78.89 ± 0.4**
MolCLR	65.74 ± 0.6	72.05 ± 1.3	62.68 ± 2.4	76.04 ± 3.1	68.78 ± 0.1	74.29 ± 1
MolE (ours)	**75.98 ± 0.3**	**76.15 ± 0.6**	**64.48 ± 2.6**	84.9 ± 0.1	**84.17 ± 0.7**	**78.52 ± 0.2**
Finetune Pre-trained						
GROVER	70 ± 0.7	73.5 ± 0.1	64.8 ± 0.1	81.2 ± 3	82.6 ± 0.7	68.2 ± 0.1
MolCLR	73.12 ± 2.1	74.61 ± 1.6	63.23 ± 3.1	87.79 ± 5.3	81.96 ± 1.1	**77.61 ± 0.9**
MolE (ours)	**75.34 ± 0.9**	**75.21 ± 1.7**	**65 ± 4.4**	**92.85 ± 1.3**	**84.17 ± 0.9**	**77.37 ± 1.2**

Average test ROC AUC metrics shown

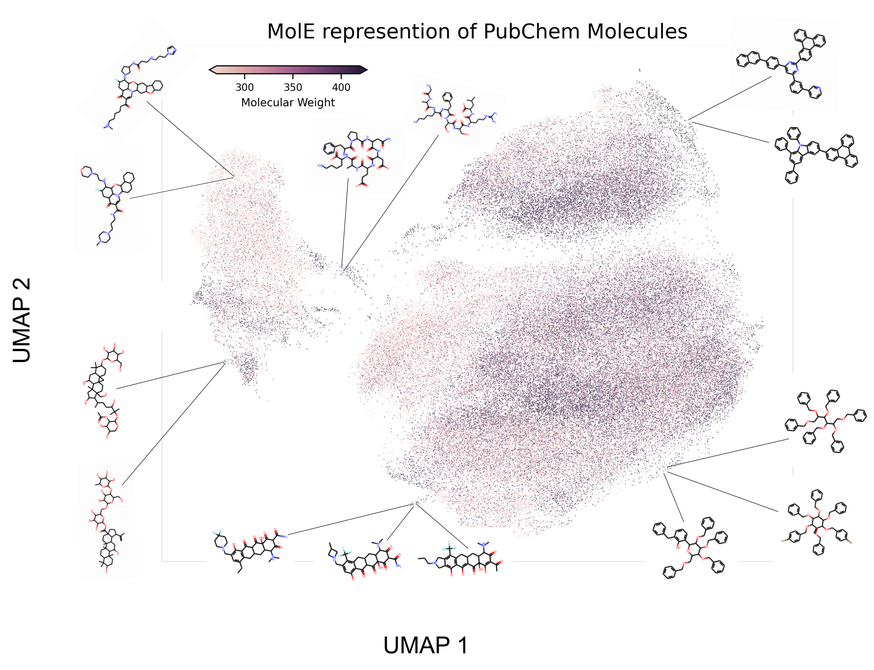
Performance on regression tasks
(Lower is better)

	FreeSolv	ESOL	Lipo	QM7	QM8
Supervised Learning					
GCN	2.87 ± 0.14	1.43 ± 0.05	0.85 ± 0.08	122.7 ± 2.2	0.036 ± 0.001
GIN	2.76 ± 0.18	1.45 ± 0.02	0.85 ± 0.07	124.8 ± 0.7	0.037 ± 0.001
SchNet	3.22 ± 0.76	1.05 ± 0.06	0.91 ± 0.1	**74.2 ± 6**	0.02 ± 0.002
MGCN	3.35 ± 0.01	1.27 ± 0.15	1.11 ± 0.04	77.6 ± 4.7	0.022 ± 0.002
D-MPNN	**2.18 ± 0.91**	**0.98 ± 0.26**	**0.65 ± 0.05**	105.8 ± 13.2	**0.014 ± 0.002**
Static Features + XGBoost					
ECFP4	4.08 ± 0.12	1.58 ± 0.02	0.92 ± 0.01	169 ± 2.81	0.024 ± 0.001
N-Gram	2.99 ± 0.01	**0.9 ± 0.04**	0.78 ± 0.01	102 ± 2.45	0.027 ± 0.001
MolCLR	3.65 ± 0.06	1.43 ± 0.02	0.9 ± 0.02	203 ± 0.01	0.03 ± 0.001
MolE	**2.78 ± 0.09**	0.95 ± 0.04	**0.75 ± 0.01**	**94 ± 6.77**	**0.021 ± 0.001**
Finetune Pre-trained					
GROVER	**2.1 ± 0.05**	**0.89 ± 0.01**	0.81 ± 0.01	94.5 ± 0.9	0.021 ± 0.001
MolCLR	3.3 ± 0.12	1.3 ± 0.04	0.75 ± 0.02	69 ± 2	**0.018 ± 0.004**
MolE	2.3 ± 0.17	0.9 ± 0.05	**0.71 ± 0.02**	**58.19 ± 1.56**	**0.018 ± 0.001**

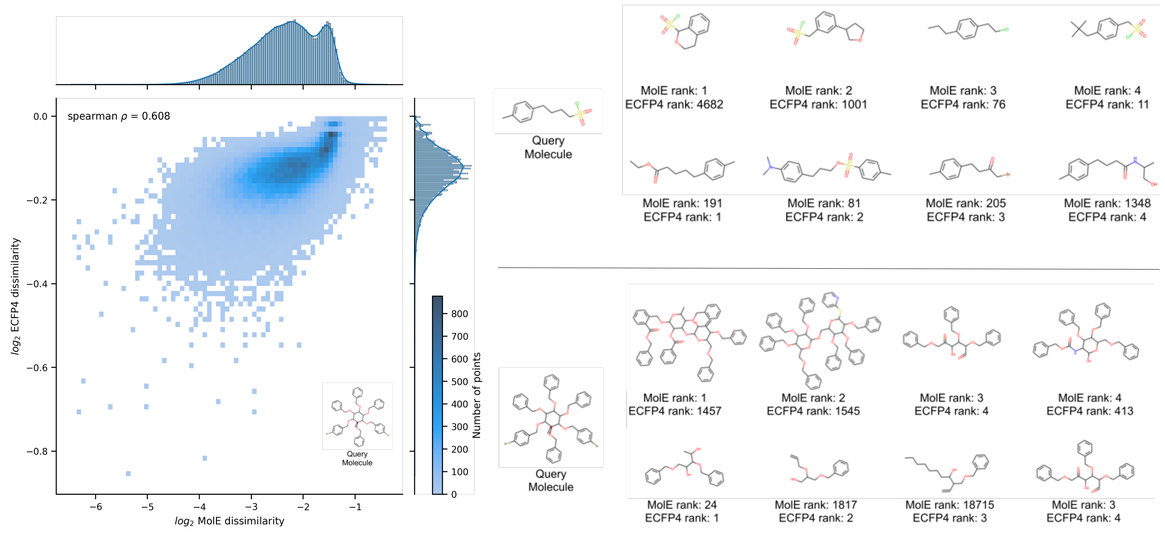
Average RMSE shown for FreeSolv, ESOL, and Lipo, MAE is shown for QM7 and QM8

MolE learns a new and informative chemical representation

Through pre-training MolE is able to learn a similar representation for molecules with shared functional groups and/or topological features

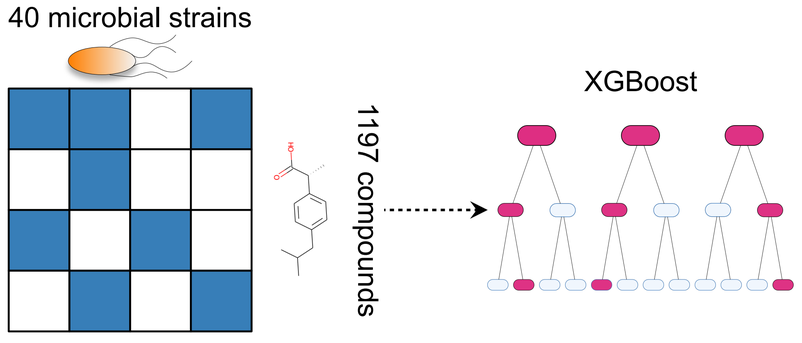


The information captured by MolE is distinct from that provided by ECFP4. This leads to differences when ranking molecules by similarity to a query.



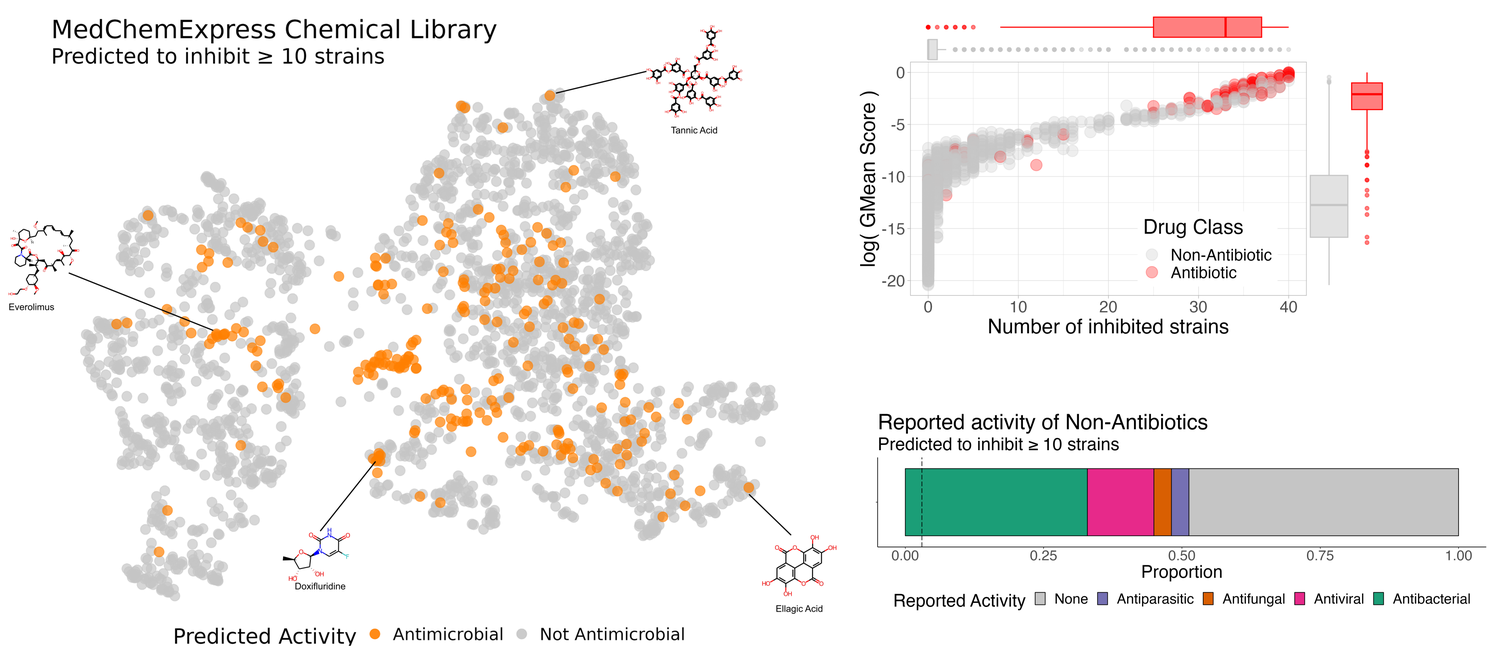
MolE can help identify compounds with antimicrobial activity

We leverage a publicly available dataset (Maier et al, 2018), to train a model to predict the antimicrobial effect of human-targeted drugs.



Chemical Features	ROC AUC	PR AUC
MolE	90.90	**77.97**
ECFP4	83.48	70.96
Chemical Descriptors	89.17	68.94
Algavi & Borenstein (New Drugs)	**91.30**	73.90

We use this model to make predictions of antimicrobial activity on a separate chemical library. We observe that known antibiotics are ranked highly by our model.Furthermore, we find that ~30% of the non-antibiotic compounds present in our positive predictions have confirmed antibacterial activity in the literature.



Email: roberto.olayo@lmu.de

Twitter: @Scietwas

Literature:
Algavi, Y. M., & Borenstein, E. (2023). A data-driven approach for predicting the impact of drugs on the human microbiome. Nature Communications, 14(1). https://doi.org/10.1038/s41467-023-39264-0
Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E. E., Brochado, A. R., Fernandez, K. C., Dose, H., Mori, H., Patil, K. R., Bork, P., & Typas, A. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. Nature, 555(7698), 623–628. https://doi.org/10.1038/nature25979
Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, & Stéphane Deny. (2021). Barlow Twins: Self-Supervised Learning via Redundancy Reduction.