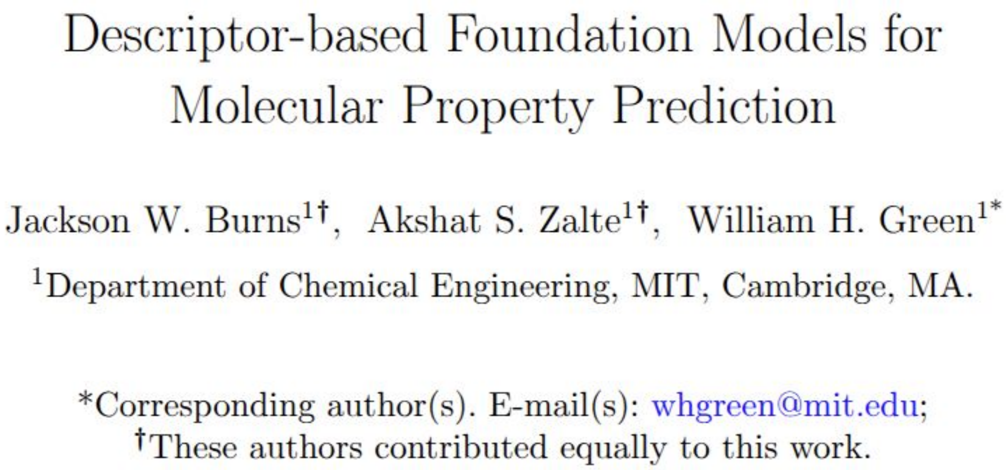
CheMeleon



Descriptor-based Foundation Models for Molecular Property Prediction

Jackson W. Burns[1†], Akshat S. Zalte[1†], William H. Green[1*]

[1] Department of Chemical Engineering, MIT, Cambridge, MA.

*Corresponding author(s). E-mail(s): whgreen@mit.edu;

†These authors contributed equally to this work.

https://www.alphaxiv.org/abs/2506.15792
https://github.com/JacksonBurns/chemeleon

CheMeleon

License MIT DOI 10.5281/zenodo.16821655

As you may know…

Predictive models

- Random Forest ⇒ Input: Fingerprint / Descriptors
- Fast Prop ⇒ Input: Topological / Phycochemical features (moldred)
- ChemProp ⇒ Input: Molecular graph

ChemProp learns molecular representation!

Limitation of Learned Representation method

Studies have shown that models leveraging learned representation (LR) tend to out-perform those using fixed molecular representations. However, in the limit of small datasets ($\lesssim O(1,000)$ samples), Chemprop and other LR models struggle often outperformed by classical methods like Random Forest.

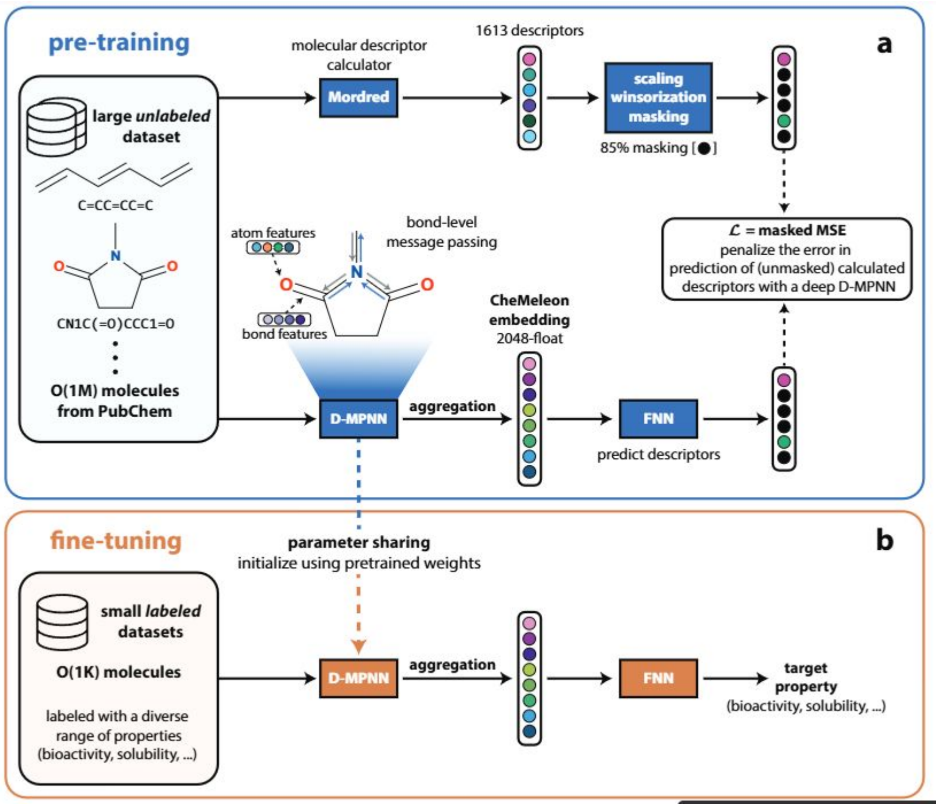
This forces the model to simultaneously learn both a suitable representation and the target property mapping, which is challenging in low-data regimes and leads to poor generalization and overfitting.

Focused on learning meaningful, general representations using molecular foundation model

Prior art

- ChemBERTa-2: pretrained on >10M SMILE strings
- MolFormer: pretrained on >10M SMILE strings
- MolCLR: pretrained based on graph
- GROVER: pretrained with combination of graph and transformer
- MolE: self-supervised approach of molecular graph + multitask learning
- ….

Proposed method: Chemelon

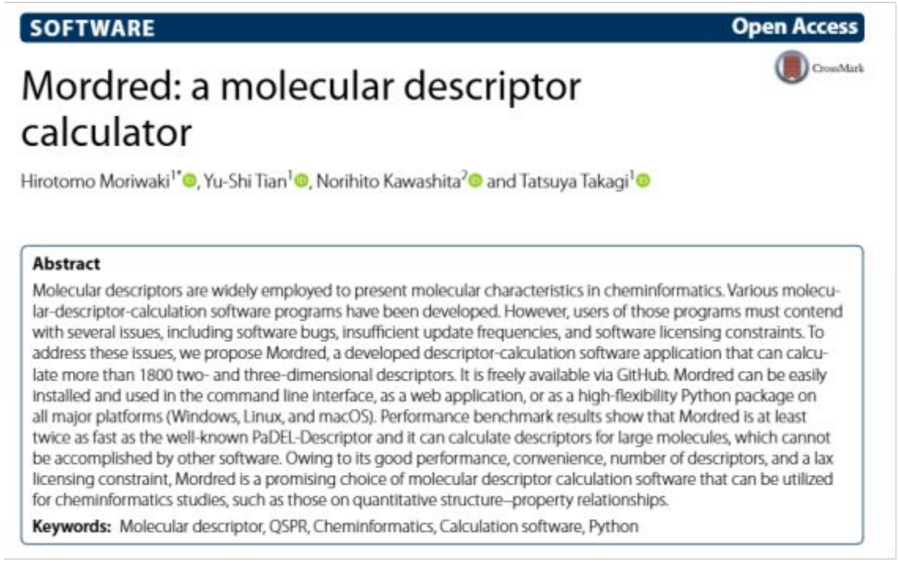

pre-training
a
molecular descriptor calculator
1613 descriptors
large *unlabeled* dataset
Mordred
scaling winsorization masking
85% masking [●]
C=CC=CC=C
CN1C(=O)CCC1=O
O(1M) molecules from PubChem
atom features
bond-level message passing
bond features
N
O
O
$\mathcal{L}$ = masked MSE
penalize the error in prediction of (unmasked) calculated descriptors with a deep D-MPNN
CheMeleon embedding 2048-float
D-MPNN
aggregation
FNN
predict descriptors
fine-tuning
b
parameter sharing
initialize using pretrained weights
small *labeled* datasets
O(1K) molecules
labeled with a diverse range of properties (bioactivity, solubility, ...)
D-MPNN
aggregation
FNN
target property (bioactivity, solubility, ...)

D-MPNN foundation model

pre-trained to predict Mordred precomputed descriptors

- When incorporating experimental values, the sparsity of the data and the presence of experimental errors become problematic.
- Using data such as QM for training introduces method-specific biases and narrows the chemical space.

Mordred



SOFTWARE **Open Access**

Mordred: a molecular descriptor calculator

Hirotomo Moriwaki[1*], Yu-Shi Tian[1], Norihito Kawashita[2] and Tatsuya Takagi[1]

Abstract

Molecular descriptors are widely employed to present molecular characteristics in cheminformatics. Various molecular-descriptor-calculation software programs have been developed. However, users of those programs must contend with several issues, including software bugs, insufficient update frequencies, and software licensing constraints. To address these issues, we propose Mordred, a developed descriptor-calculation software application that can calculate more than 1800 two- and three-dimensional descriptors. It is freely available via GitHub. Mordred can be easily installed and used in the command line interface, as a web application, or as a high-flexibility Python package on all major platforms (Windows, Linux, and macOS). Performance benchmark results show that Mordred is at least twice as fast as the well-known PaDEL-Descriptor and it can calculate descriptors for large molecules, which cannot be accomplished by other software. Owing to its good performance, convenience, number of descriptors, and a lax licensing constraint, Mordred is a promising choice of molecular descriptor calculation software that can be utilized for cheminformatics studies, such as those on quantitative structure–property relationships.

Keywords: Molecular descriptor, QSPR, Cheminformatics, Calculation software, Python

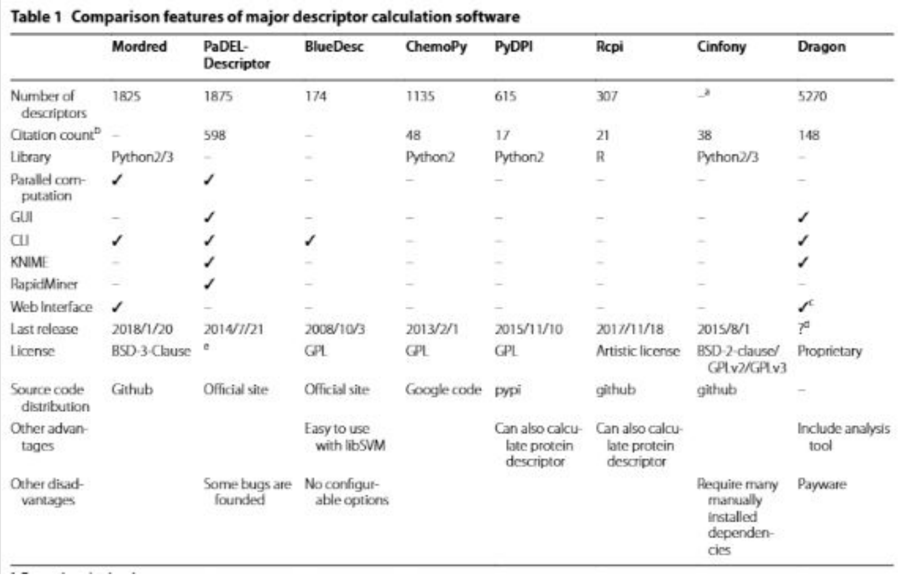
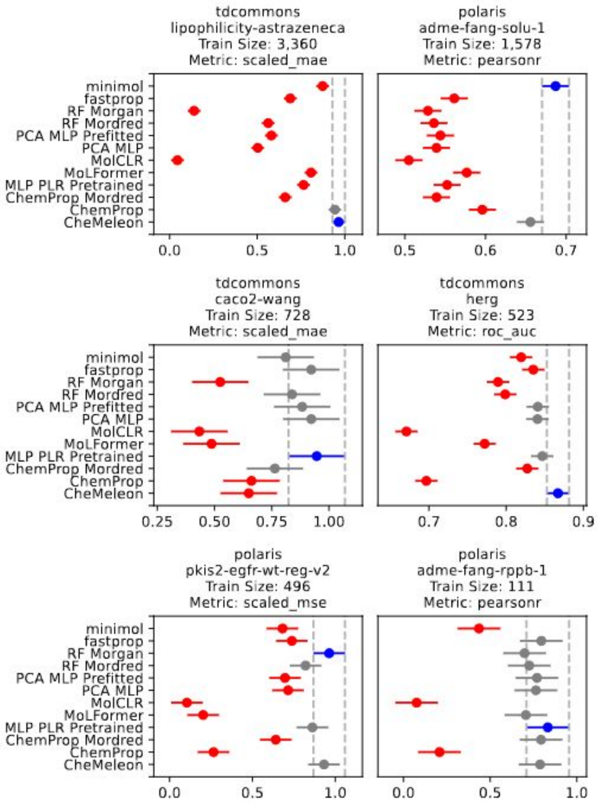


Table 1 Comparison features of major descriptor calculation software

	Mordred	PaDEL-Descriptor	BlueDesc	ChemoPy	PyDPI	Rcpi	Cinfony	Dragon
Number of descriptors	1825	1875	174	1135	615	307	–[a]	5270
Citation count[b]	–	598	–	48	17	21	38	148
Library	Python2/3	–	–	Python2	Python2	R	Python2/3	–
Parallel computation	✓	✓	–	–	–	–	–	–
GUI	–	✓	–	–	–	–	–	✓
CLI	✓	✓	✓	–	–	–	–	✓
KNIME	–	✓	–	–	–	–	–	✓
RapidMiner	–	✓	–	–	–	–	–	–
Web Interface	✓	–	–	–	–	–	–	✓[c]
Last release	2018/1/20	2014/7/21	2008/10/3	2013/2/1	2015/11/10	2017/11/18	2015/8/1	?[d]
License	BSD-3-Clause	[e]	GPL	GPL	GPL	Artistic license	BSD-2-clause/ GPLv2/GPLv3	Proprietary
Source code distribution	Github	Official site	Official site	Google code	pypi	github	github	–
Other advantages			Easy to use with libSVM		Can also calculate protein descriptor	Can also calculate protein descriptor		Include analysis tool
Other disadvantages		Some bugs are founded	No configurable options				Require many manually installed dependencies	Payware

https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0258-y

Results



- Blue are the absolute highest performers on the given benchmark,
- Red are practically worse performers and are considered to have ”lost” on the indicated benchmark.

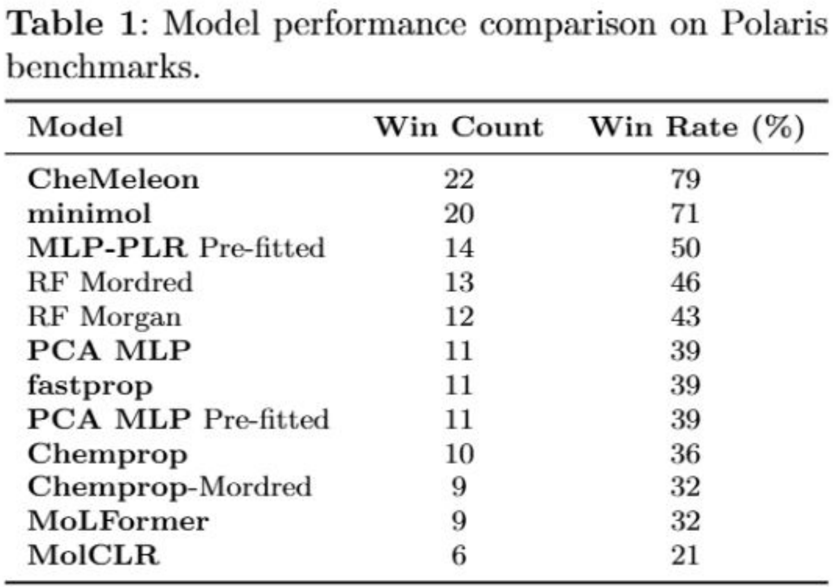


Table 1: Model performance comparison on Polaris benchmarks.

Model	Win Count	Win Rate (%)
CheMeleon	22	79
minimol	20	71
MLP-PLR Pre-fitted	14	50
RF Mordred	13	46
RF Morgan	12	43
PCA MLP	11	39
fastprop	11	39
PCA MLP Pre-fitted	11	39
Chemprop	10	36
Chemprop-Mordred	9	32
MoLFormer	9	32
MolCLR	6	21

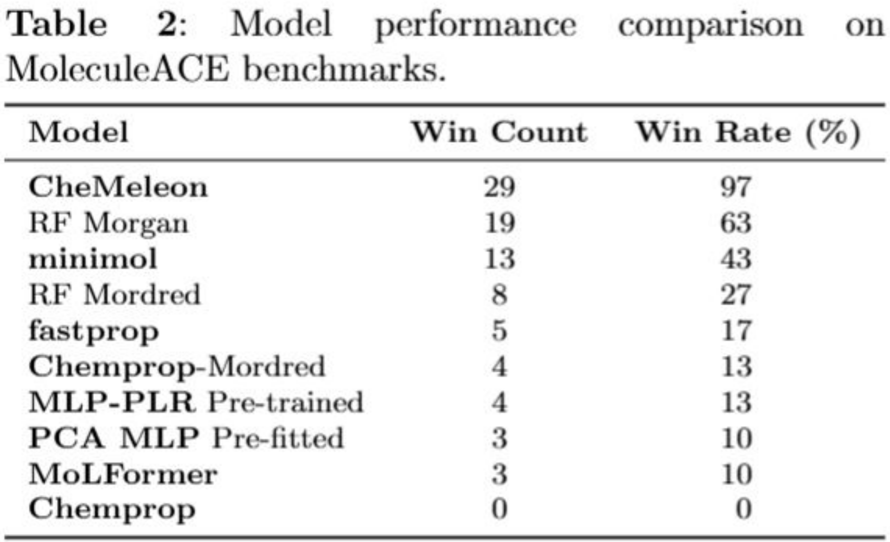
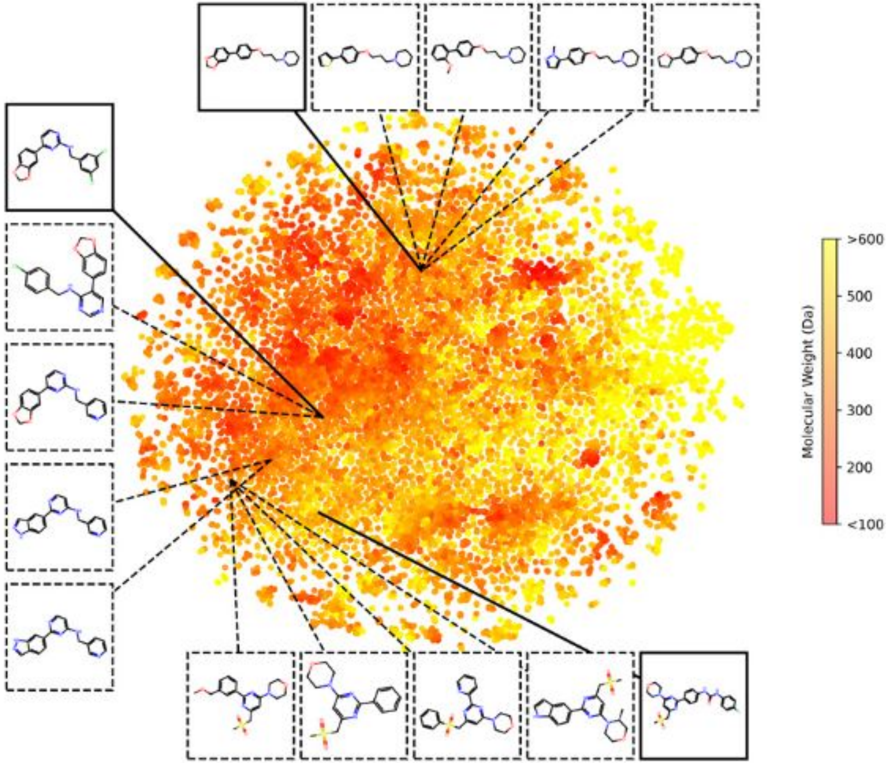


Table 2: Model performance comparison on MoleculeACE benchmarks.

Model	Win Count	Win Rate (%)
CheMeleon	29	97
RF Morgan	19	63
minimol	13	43
RF Mordred	8	27
fastprop	5	17
Chemprop-Mordred	4	13
MLP-PLR Pre-trained	4	13
PCA MLP Pre-fitted	3	10
MoLFormer	3	10
Chemprop	0	0

Foundation Fingerprint



t-SNE projection maps CheMeleon's high-dimensional embeddings into a two-dimensional space, allowing for visual inspection of how structurally related molecules cluster. Three distinct chemical series from the benchmark are highlighted, each beginning with a lead compound shown in bold.

Computational resources

Nvidia 2080 Ti x 8 for foundation model training

Nvidia Quadro RTX 4000 x 1 for finetuning