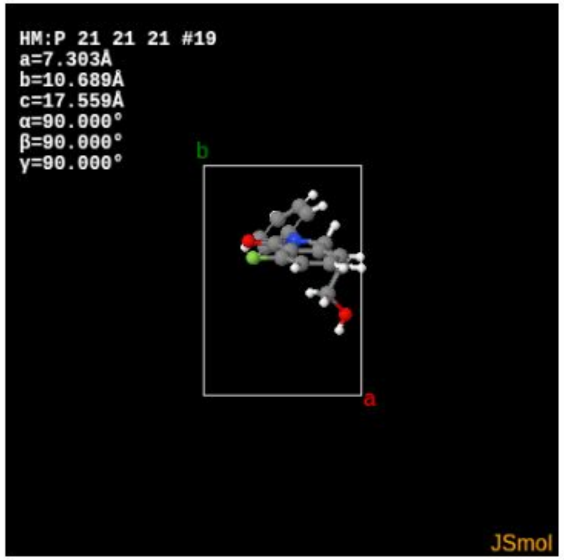
ML-UFF powered validation of the COD

Justinas Šlepavičius, Andrius Merkys, Antanas Vaitkus, Saulius Gražulis, Linas Vilčiauskas

COD

7250815 << **7250816** >> 7700000

Preview



Coordinates	7250816.cif
Original paper (by DOI)	HTML

▾ Structure parameters

Formula	C16 H14 F N O2
Calculated formula	C16 H14 F N O2
Title of publication	Tetrabutylammonium decatungstate-catalyzed hydroalkylation/alkoxylation of 3-methyleneisoindolin-1-ones with alcohols/ethers through hydrogen atom transfer process
Authors of publication	Fu, Han; Wang, Feixia; Wu, Ziyan; Xie, Xiaoyu; Zhang, Yicheng; Liu, Jie; Wang, Lei
Journal of publication	RSC Advances
Year of publication	2025
Journal volume	15
Journal issue	37
Pages of publication	30285 - 30289

Crystallography Open Database (COD) is an open-access collection of crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding biopolymers.

It contains more than half a million structures

It would be impossible to validate all the structures using DFT

ML-UFFs

Together with the AI boom of the pandemic, Machine Learning Universal Force Fields (ML-UFFs) were introduced

ML-UFFs aim to be a 'learned' Universal ForceField trained on DFT data

The number and accuracy of MLP-UFFs introduced is increasing

Matbench is a popular ML-UFF ranking website, where ML-UFFs are tested in a systematic way

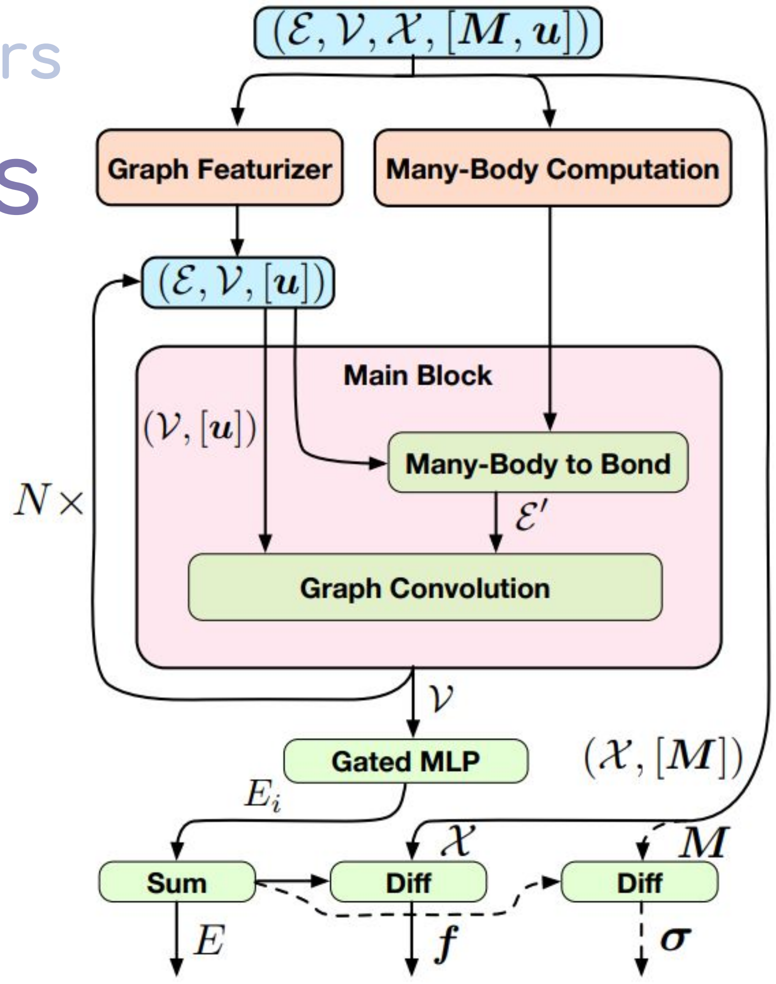
Model	RMSD ↓	Training Set	Params	Targets	Date Added
eSEN-30M-OAM	0.061	6.6M (113M) OMat24+MPtrj+sAlex	30.2M	EFS_G	2025-03-17
ORB v3	0.075	6.47M (133M) MPtrj+Alex+OMat24	25.5M	EFS_G	2025-04-05
SevenNet-MF-ompa	0.064	6.6M (113M) OMat24+sAlex+MPtrj	25.7M	EFS_G	2025-03-13
GRACE-2L-OAM	0.067	6.6M (113M) OMat24+sAlex+MPtrj	12.6M	EFS_G	2025-02-06
DPA-3.1-3M-FT	0.069	163M OpenLAM	3.27M	EFS_G	2025-06-05
eSEN-30M-MP	0.075	146k (1.58M) MPtrj	30.1M	EFS_G	2025-03-17
MACE-MPA-0	0.073	3.37M (12M) MPtrj+sAlex	9.06M	EFS_G	2024-12-09
AlphaNet-v1-OMA	0.079	6.6M (113M) OMat24+sAlex+MPtrj	4.65M	EFS_G	2025-05-12
MatterSim v1 5M	0.073	17M MatterSim	4.55M	EFS_G	2024-12-16
GRACE-1L-OAM	0.072	6.6M (113M) OMat24+sAlex+MPtrj	3.45M	EFS_G	2025-02-06
Eqnorm MPtrj	0.084	146k (1.58M) MPtrj	1.31M	EFS_G	2025-05-26
Nequix MP	0.085	146k (1.58M) MPtrj	708k	EFS_G	2025-08-17
DPA-3.1-MPtrj	0.080	146k (1.58M) MPtrj	4.81M	EFS_G	2025-06-05
SevenNet-l3i5	0.085	146k (1.58M) MPtrj	1.17M	EFS_G	2024-12-10
HIENet	0.080	146k (1.58M) MPtrj	7.51M	EFS_G	2025-07-01
MatRIS v0.5.0 MPtrj	0.077	146k (1.58M) MPtrj	5.83M	EFS_GM	2025-03-13
GRACE-2L-MPtrj	0.090	146k (1.58M) MPtrj	15.3M	EFS_G	2024-11-21
MACE-MP-0	0.091	146k (1.58M) MPtrj	4.69M	EFS_G	2023-07-14
eqV2 M	0.069	3.37M (102M) OMat24+MPtrj	86.6M	EFS_D	2024-10-18
ORB v2	0.097	3.25M (32.1M) MPtrj+Alex	25.2M	EFS_D	2024-10-11
eqV2 S DeNS	0.076	146k (1.58M) MPtrj	31.2M	EFS_D	2024-10-18
ORB v2 MPtrj	0.101	146k (1.58M) MPtrj	25.2M	EFS_D	2024-10-14
M3GNet	0.112	62.8k (188k) MPF	228k	EFS_G	2022-09-20
CHGNet	0.095	146k (1.58M) MPtrj	413k	EFS_GM	2023-03-03

A material graph neural network

M3G-Net

188k Training Structures

211k Parameters

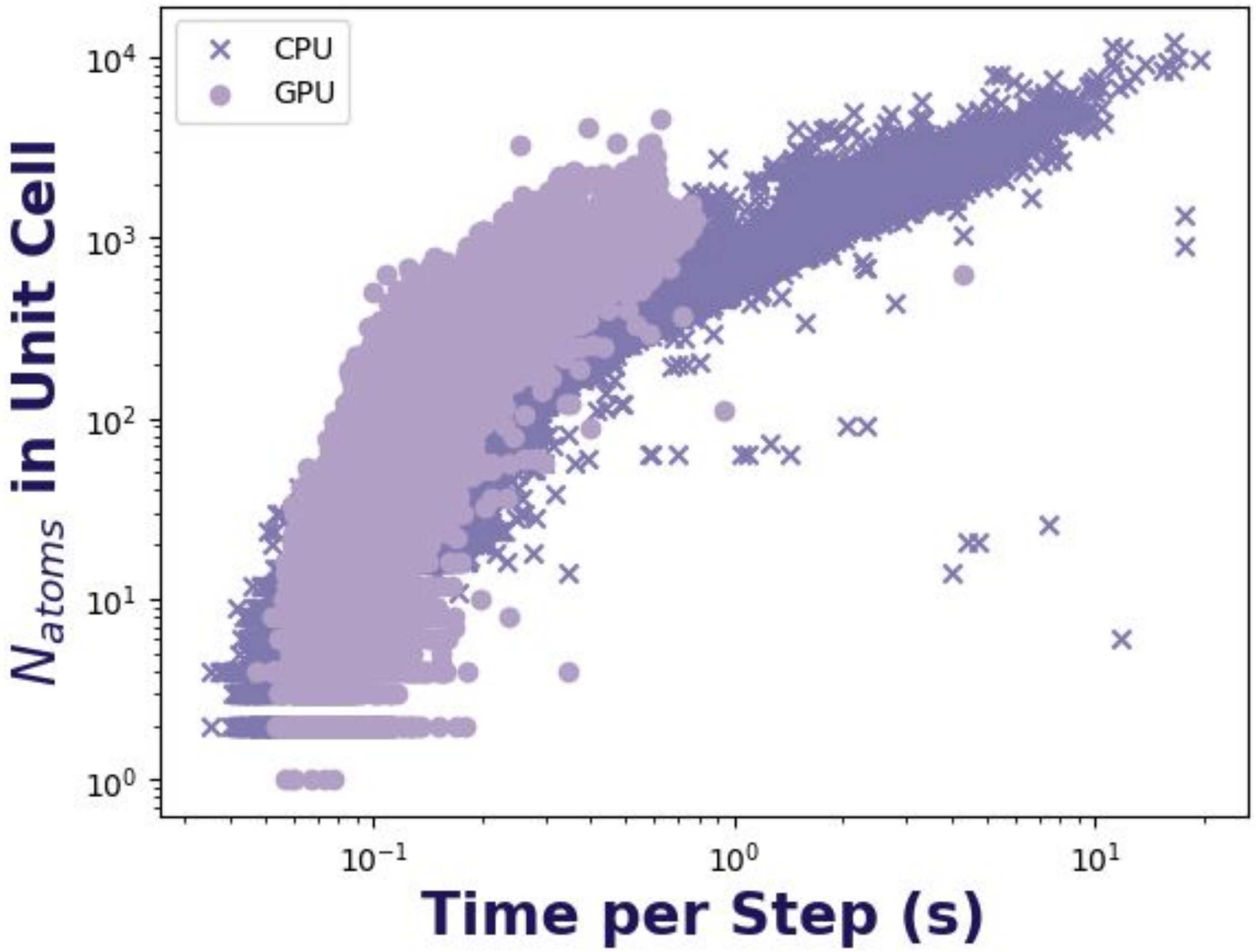


Incorporates 3 body interactions

Works for all elements up to Ac

Embeddings trivialize the calculation of outputs

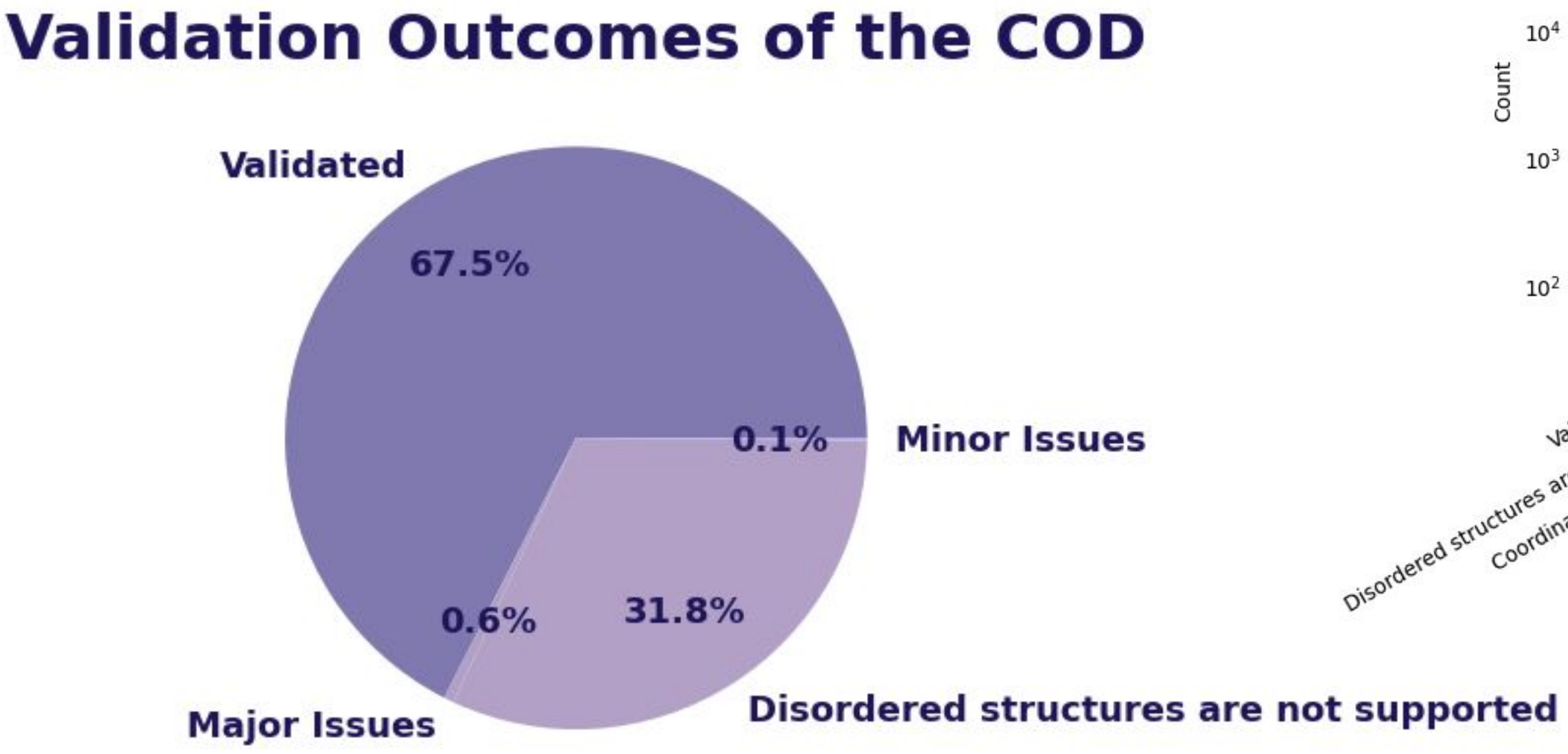
Results M3G-Net (Speed)

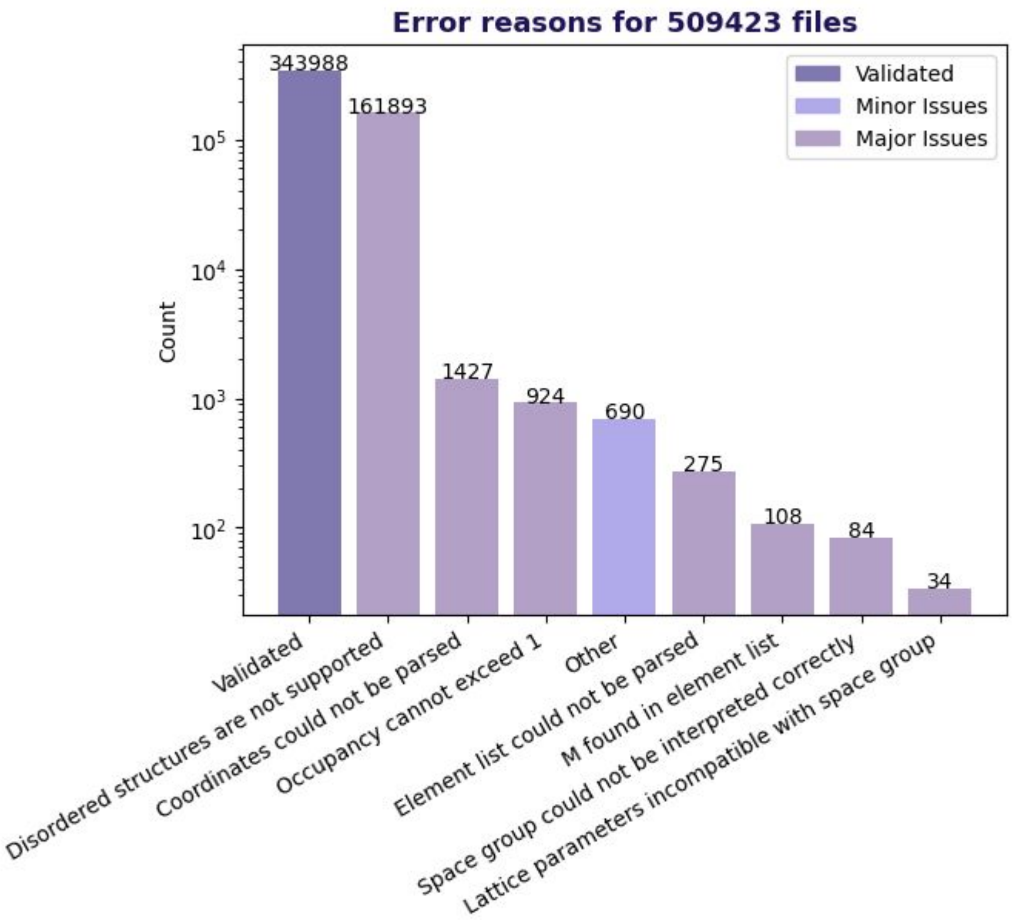


Many of the structures could be relaxed on the CPU with not a large time increase

The CPU runs were done for structures that required too much CUDA Memory

Results M3G-Net (Errors)

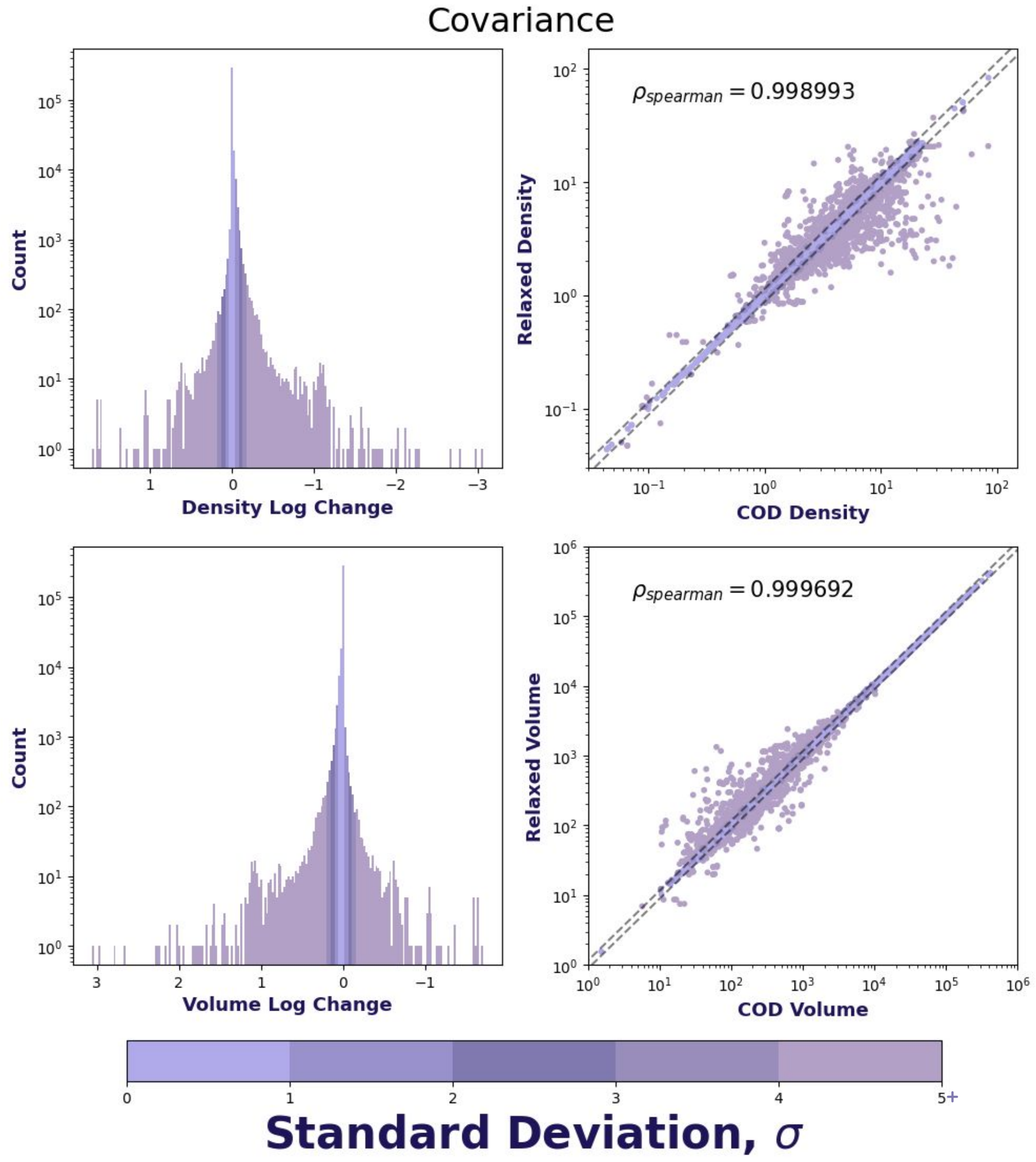




Results M3G-Net (Covariance)

$$\text{Log Change} = \ln\left(\frac{y}{x}\right)$$

Both density and volume have a very strong peak at 0 log change



$$\rho_{\text{Pearson}} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$$

The correlation coefficients of both values are very high, but there are surrounding 'clouds' of data points around the y=x line

Verification during submission

- To add an extra layer of certainty to submitted structures, M3G-Net can be used for validation during submission to the COD
- This could be achieved using a standalone compiled binary

Conclusions

- ML-UFFs can be used to verify the COD due to its low computational cost compared to quantum mechanical methods
- M3G-Net was used to validate more than 67% of the entries, with the vast majority of unvalidated entries being disordered structures
- Pearson coefficients show a very good agreement (>0.998) between the original and relaxed values of density and volume
- This method can implemented for validation during submission

Acknowledgements

This project is under grant agreement with the Research Council of Lithuania (LMTLT) (Project No. P-ITP-24-9)

Thank You!

Any Questions?