The Journey to Transferable, Accurate, Anisotropic Molecular Machine Learning

The DCM Approach

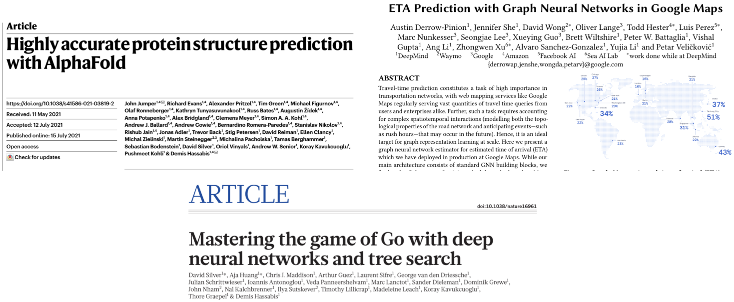
Eric D. Boittier

University of Basel

November 13, 2025

Outline

Ordinary World



Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2
Received: 11 May 2021
Accepted: 12 July 2021
Published online: 15 July 2021
Open access
Check for updates

John Jumper[1,4], Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]

ETA Prediction with Graph Neural Networks in Google Maps

Austin Derrow-Pinion[1], Jennifer She[1], David Wong[2*], Oliver Lange[3], Todd Hester[4*], Luis Perez[5*], Marc Nunkesser[3], Seongjae Lee[3], Xueying Guo[3], Brett Wiltshire[1], Peter W. Battaglia[1], Vishal Gupta[1], Ang Li[1], Zhongwen Xu[6*], Alvaro Sanchez-Gonzalez[1], Yujia Li[1] and Petar Veličković[1]
[1]DeepMind [2]Waymo [3]Google [4]Amazon [5]Facebook AI [6]Sea AI Lab *work done while at DeepMind
{derrowap,jenshe,wongda,petarv}@google.com

ABSTRACT

Travel-time prediction constitutes a task of high importance in transportation networks, with web mapping services like Google Maps regularly serving vast quantities of travel time queries from users and enterprises alike. Further, such a task requires accounting for complex spatiotemporal interactions (modelling both the topological properties of the road network and anticipating events—such as rush hours—that may occur in the future). Hence, it is an ideal target for graph representation learning at scale. Here we present a graph neural network estimator for estimated time of arrival (ETA) which we have deployed in production at Google Maps. While our main architecture consists of standard GNN building blocks, we

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silver[1*], Aja Huang[1*], Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

Figure: AlphaFold, Google Maps, AlphaGo - Graph Neural Networks

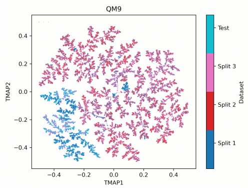
Call to Adventure

- Markus's 2019 talk on Hemoglobin sparks curiosity
- Mentions of a new, graph neural network model called PhysNet
- The rest is history...

Refusal of the Call

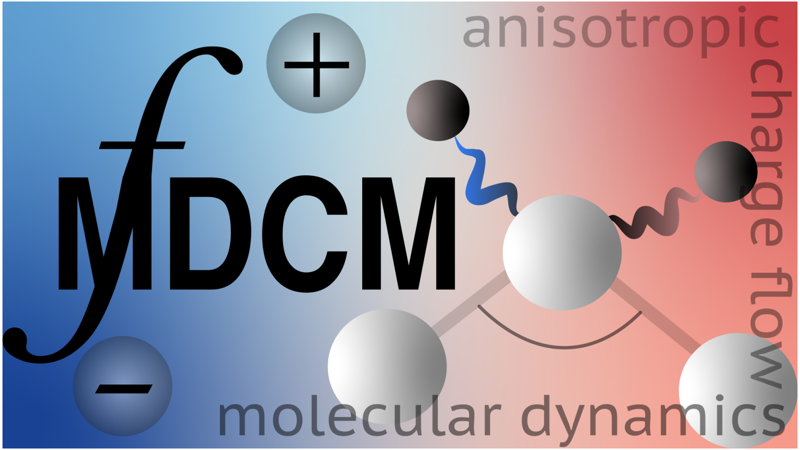
In atomistic machine learning, two paradigms emerge:

- **Chemical Space** (combinatorial explosion of possible molecules)
- **Conformational space** (many possible conformations for a given molecule)

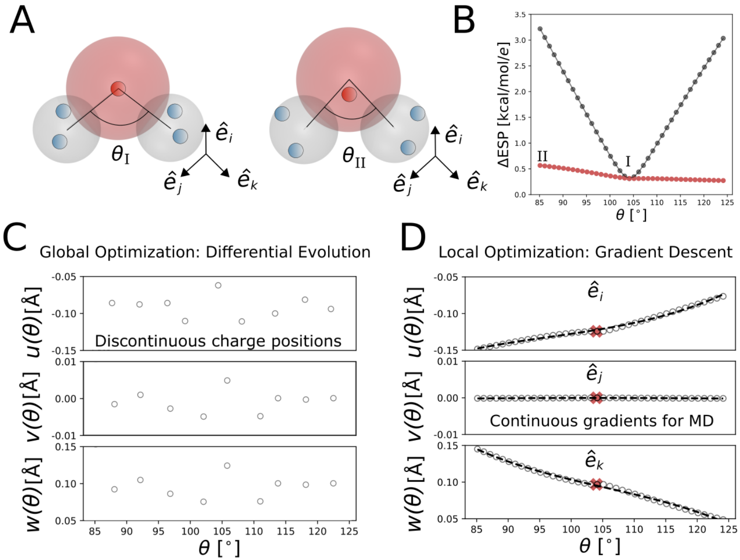


Crossing the First Threshold

Fluctuating MDCM (f-MDCM)



Tests, Allies, and Enemies



Tests, Allies, and Enemies

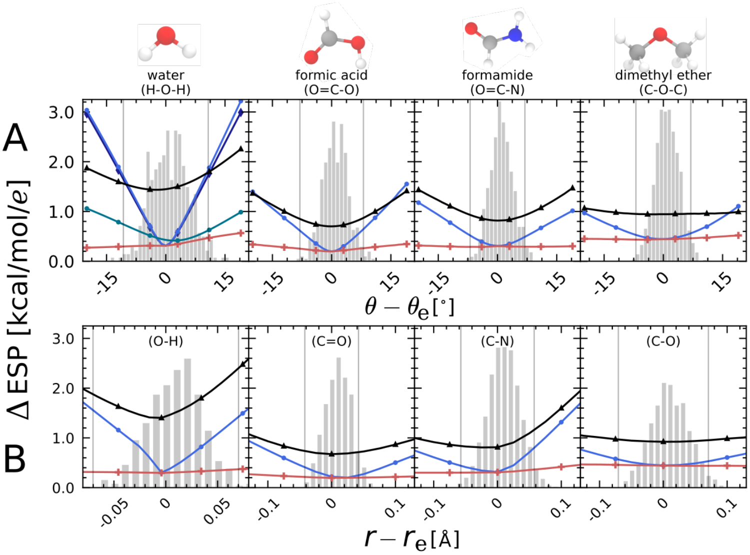


Figure: Early ideas: f-MDCM

Tests, Allies, and Enemies

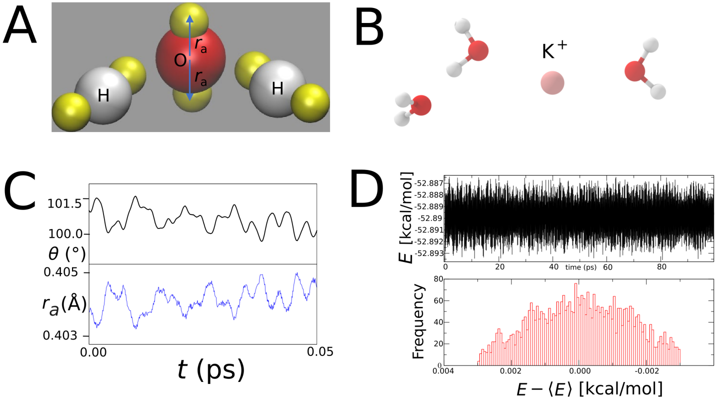


Figure: Energy conservation and tests

Approach to the Innermost Cave

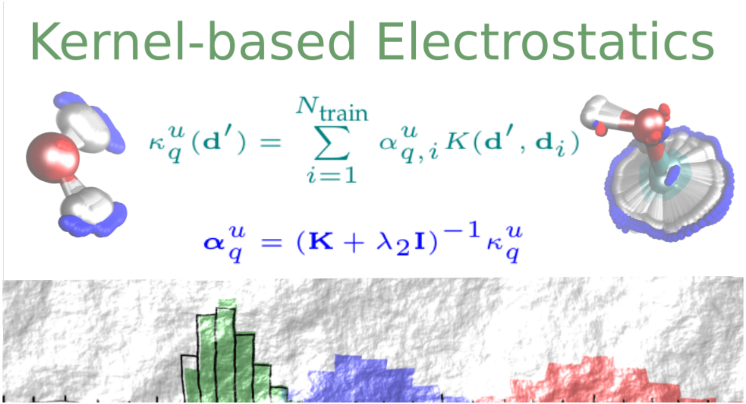


Figure: The kMDCM approach

Approach to the Innermost Cave

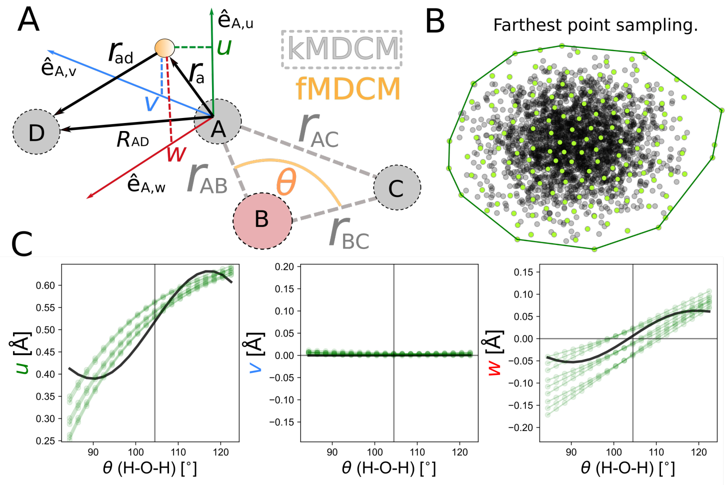


Figure: The kMDCM approach (continued)

Approach to the Innermost Cave

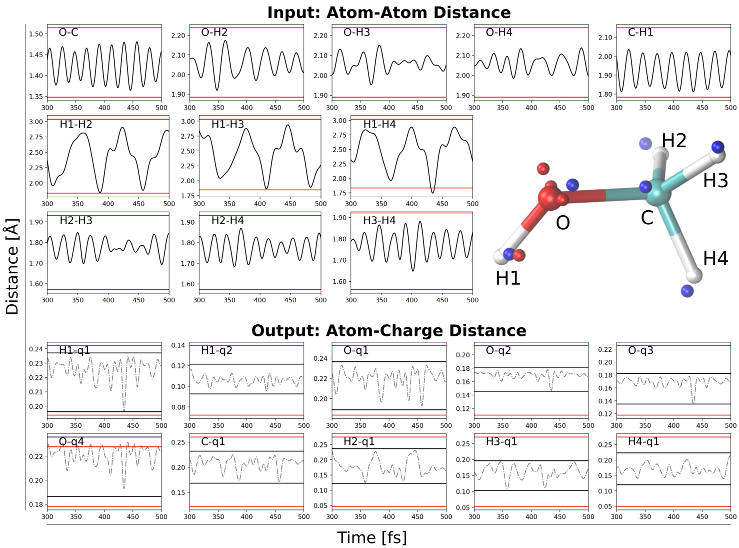


Figure: The kMDCM approach (continued)

Approach to the Innermost Cave

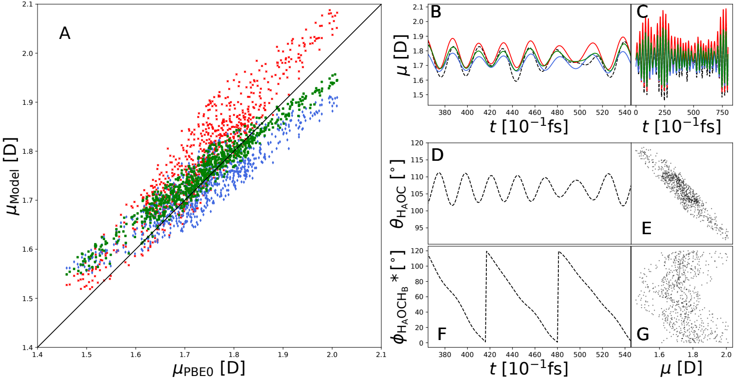


Figure: The kMDCM approach (continued)

The Ordeal

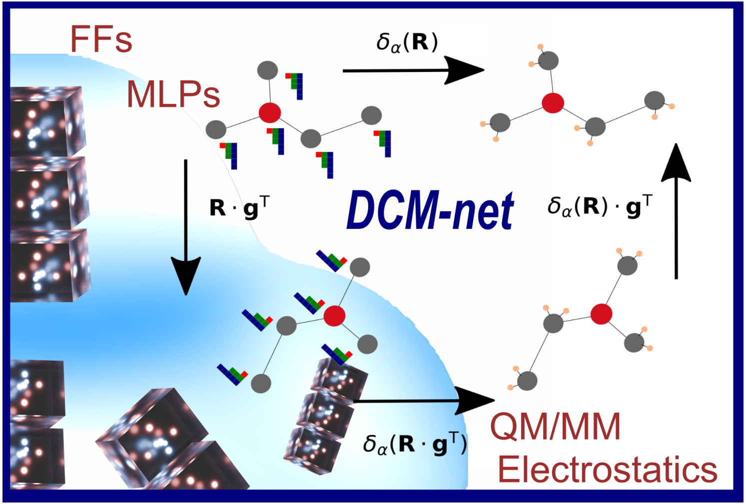
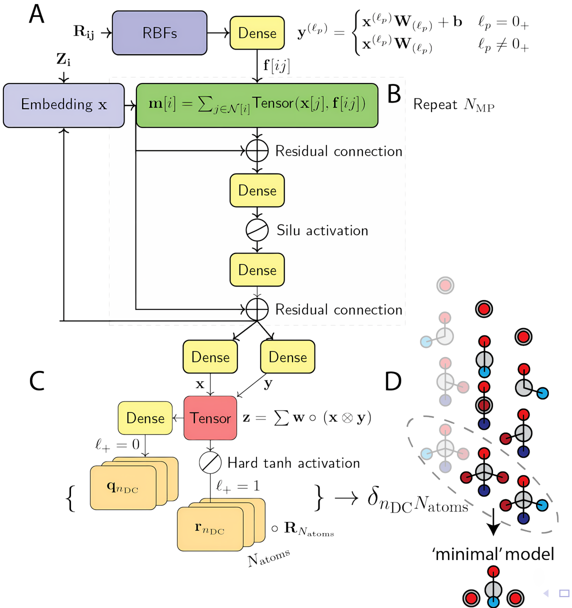


Figure: DCM-Net

The Ordeal


A
$\mathbf{R_{ij}}$
RBFs
Dense
$\mathbf{y}^{(\ell_p)} = \begin{cases} \mathbf{x}^{(\ell_p)}\mathbf{W}_{(\ell_p)} + \mathbf{b} & \ell_p = 0_+ \\ \mathbf{x}^{(\ell_p)}\mathbf{W}_{(\ell_p)} & \ell_p \neq 0_+ \end{cases}$
$\mathbf{Z_i}$
$\mathbf{f}[ij]$
Embedding $\mathbf{x}$
$\mathbf{m}[i] = \sum_{j \in \mathcal{N}[i]} \text{Tensor}(\mathbf{x}[j], \mathbf{f}[ij])$
B
Repeat N_{MP}
Residual connection
Dense
Silu activation
Dense
Residual connection
Dense
Dense
C
$\mathbf{x}$
$\mathbf{y}$
Dense
Tensor
$\mathbf{z} = \sum \mathbf{w} \circ (\mathbf{x} \otimes \mathbf{y})$
$\ell_+ = 0$
Hard tanh activation
$\ell_+ = 1$
$\{ \mathbf{q}_{n_{\text{DC}}} \quad \mathbf{r}_{n_{\text{DC}}} \circ \mathbf{R}_{N_{\text{atoms}}} \} \rightarrow \delta_{n_{\text{DC}} N_{\text{atoms}}}$
N_{atoms}
D
'minimal' model

The Ordeal

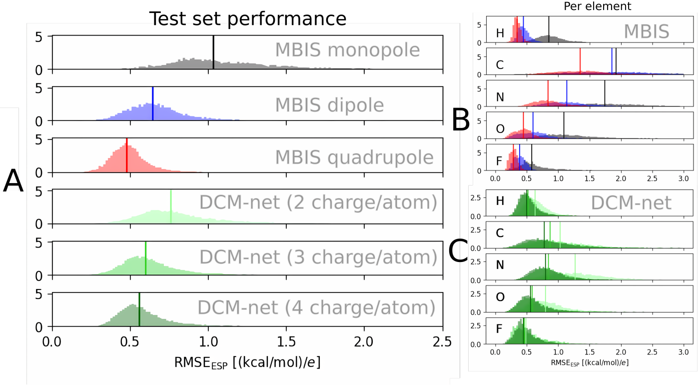


Figure: DCM-Net

The Ordeal

Training

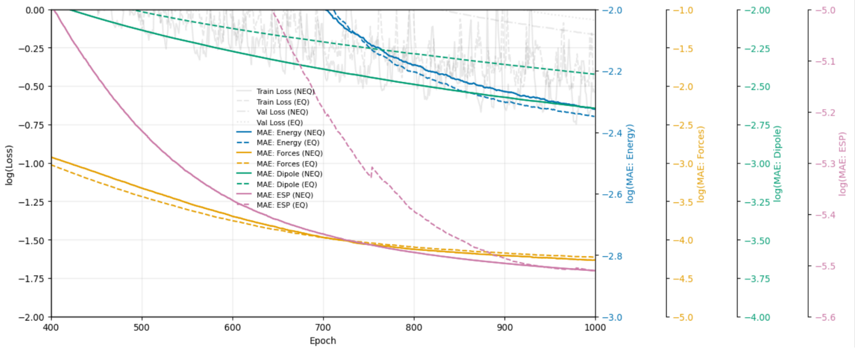


Figure: Equivariance test: Training

The Ordeal

Validation

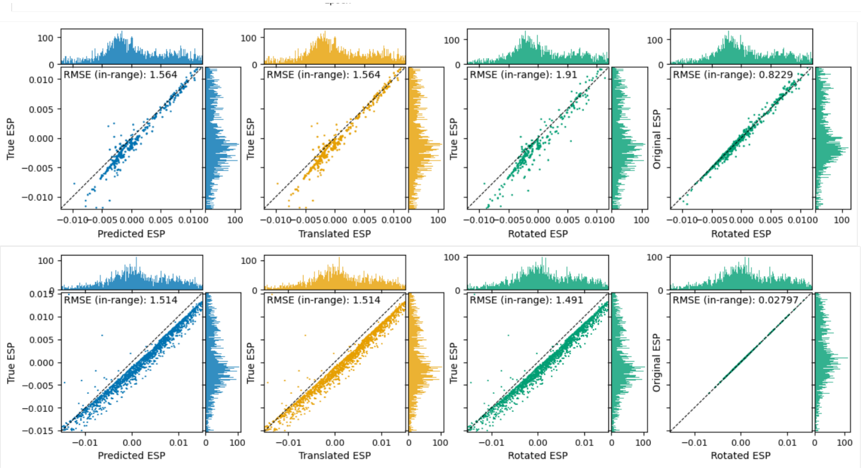


Figure: Equivariance test: Validation

The Ordeal

Test

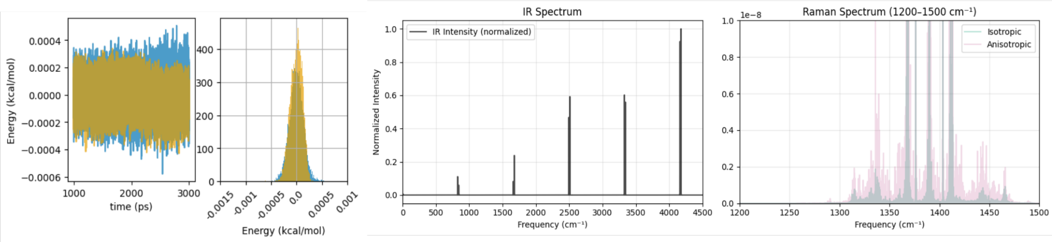
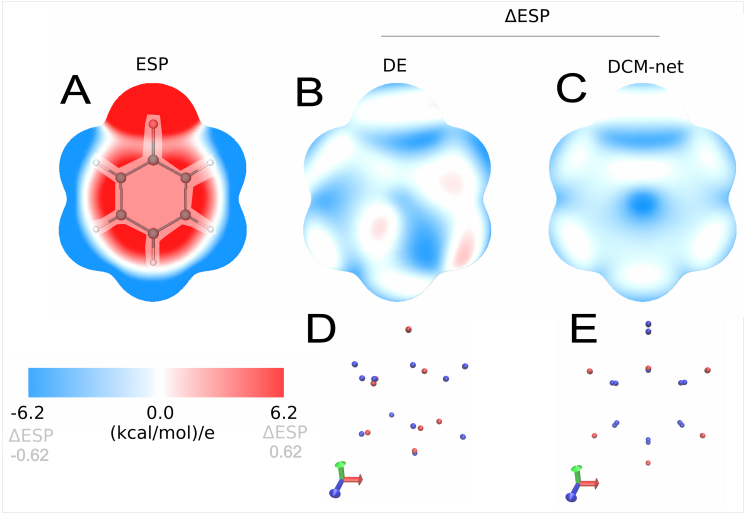


Figure: Equivariance test: Test

Reward

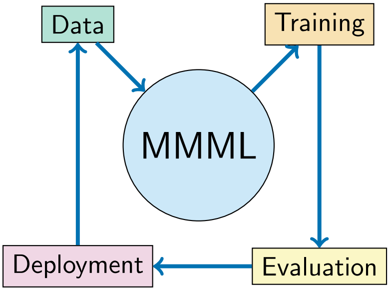
Rapid, Accurate, Anisotropic Molecular Machine Learning



The Road Back

Modern Molecular ML Framework

- Production-ready CLI tools
- PhysNet + DCMNet architectures
- Efficient data handling
- Comprehensive validation
- Multi-state predictions



Dataset Splitting

Create train/validation/test splits

```
# Split dataset (80% train, 10% val, 10% test)
python -m mmml.cli.split_dataset glycol_cleaned.npz \
    -o splits/ \
    --train-size 0.8 \
    --valid-size 0.1 \
    --test-size 0.1 \
    --seed 42
```

Features:

- Reproducible splits (seed-based)
- Filters out non-per-structure fields
- Saves split indices for reproducibility
- Stratified splitting available

Output: `data_train.npz`, `data_valid.npz`, `data_test.npz`

Basic Training

Train a PhysNet model for energies, forces, and dipoles

```
# Simple training with auto-detection
python -m mmml.cli.make_training \
    --data splits/data_train.npz \
    --ckpt_dir checkpoints/glycol_run1 \
    --n_train 4000 \
    --n_valid 500 \
    --num_epochs 100 \
    --batch_size 16
```

Auto-detected:

- Number of atoms (from dataset)
- Padding removal (if needed)
- Checkpoint path (made absolute)

Acknowledgments

Thank you for using MMML!

MMML Development Team

Built with:

- JAX & Flax (neural networks)
- e3x (equivariant operations)
- ASE (atomic simulations)
- NumPy & SciPy (numerical computing)

Questions?

GitHub: `mmml`
Documentation: `docs/`