REM3DI: Learning 3D molecular descriptors from atomistic foundation models

Steffen Wedig[1,2] Rokas Elijošius[1] Christoph Schran[1] Lars Leon Schaaf[1]

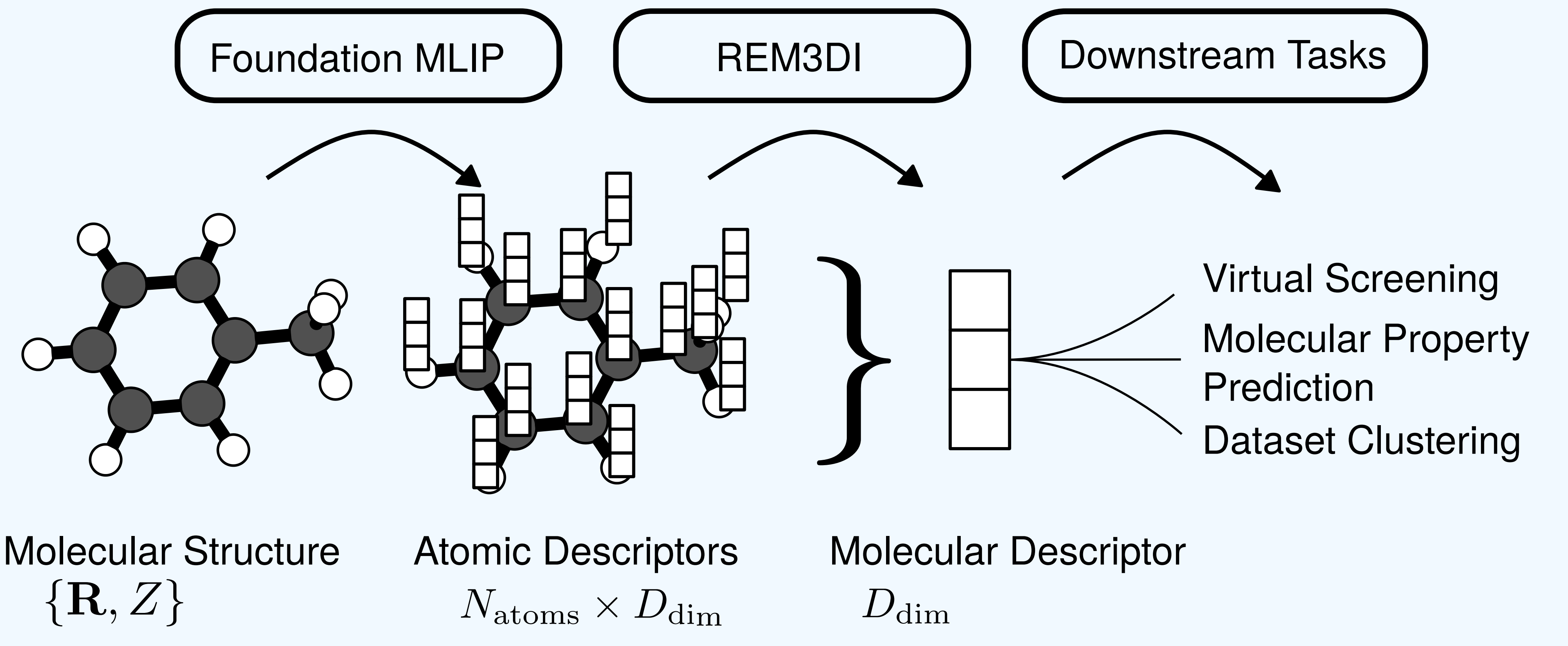
[1] University of Cambridge [2]Dunia Innovations

NEURAL INFORMATION PROCESSING SYSTEMS

NeurReps Workshop

Introduction

- Predicting molecular properties in drug design requires a numerical representation of small organic molecules.
- Conventional cheminformatics fingerprints encode selected features (e.g. presence of specific functional groups).
- Machine Learning Interatomic Potentials learn accurate representation of local atomic neighbourhoods to model potential energy surface. REM3DI aggregates descriptors into one physically-motivated and transferable representation across tasks and molecular chemistries.


Foundation MLIP

REM3DI

Downstream Tasks

Virtual Screening

Molecular Property Prediction

Dataset Clustering

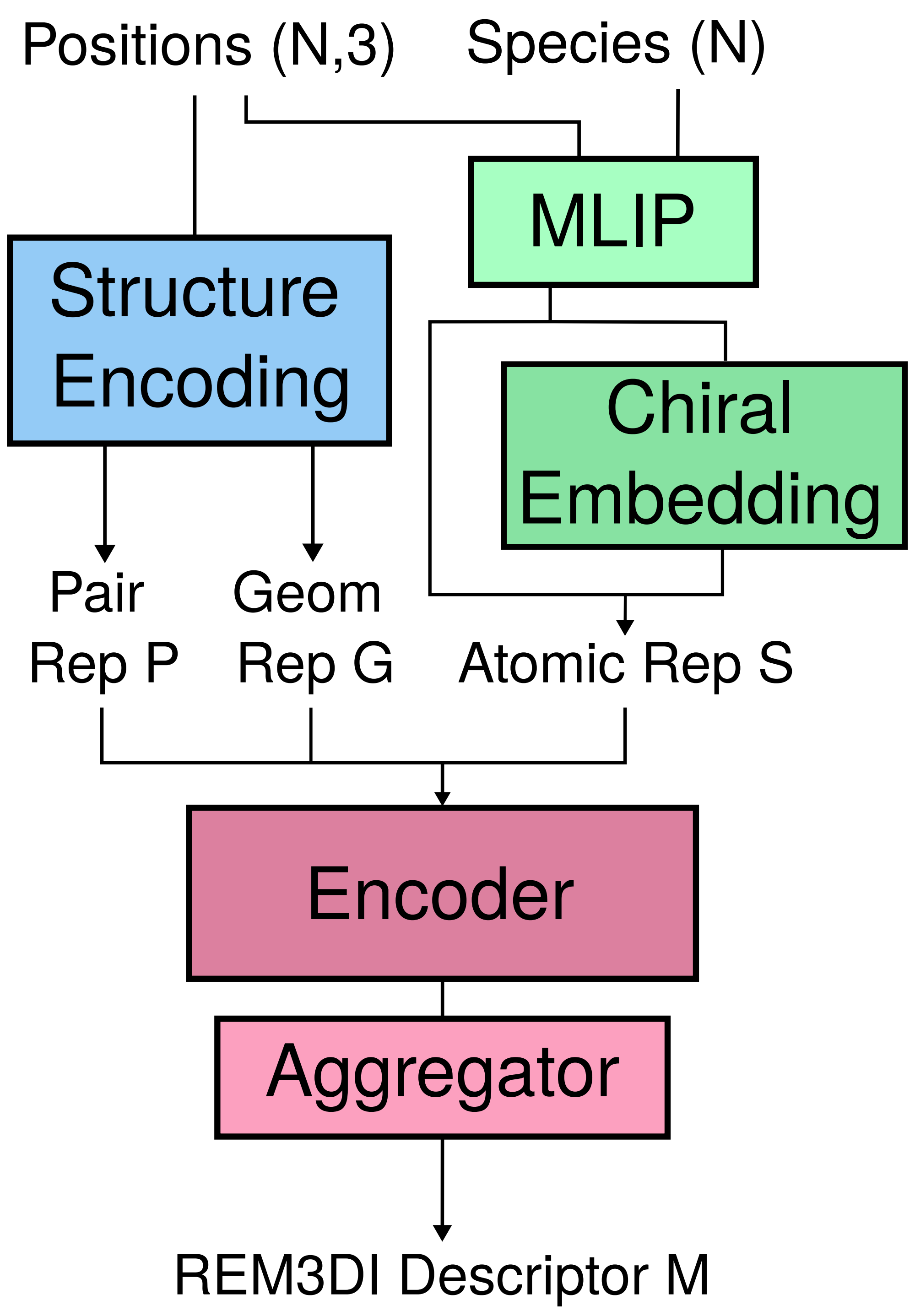
Molecular Structure $\{\mathbf{R}, Z\}$

Atomic Descriptors $N_{\text{atoms}} \times D_{\text{dim}}$

Molecular Descriptor D_{dim}

Architecture

- REM3DI learns attention based aggregation of invariant MLIP features into one fixed size vector.
- E(3)-equivariant MLIP features vary smoothly with position, and contain rich description of local atomic neighbourhood.
- Transformer encoder, biased attention from pair representation P.


Positions (N,3)

Species (N)

MLIP

Structure Encoding

Chiral Embedding

Pair Rep P

Geom Rep G

Atomic Rep S

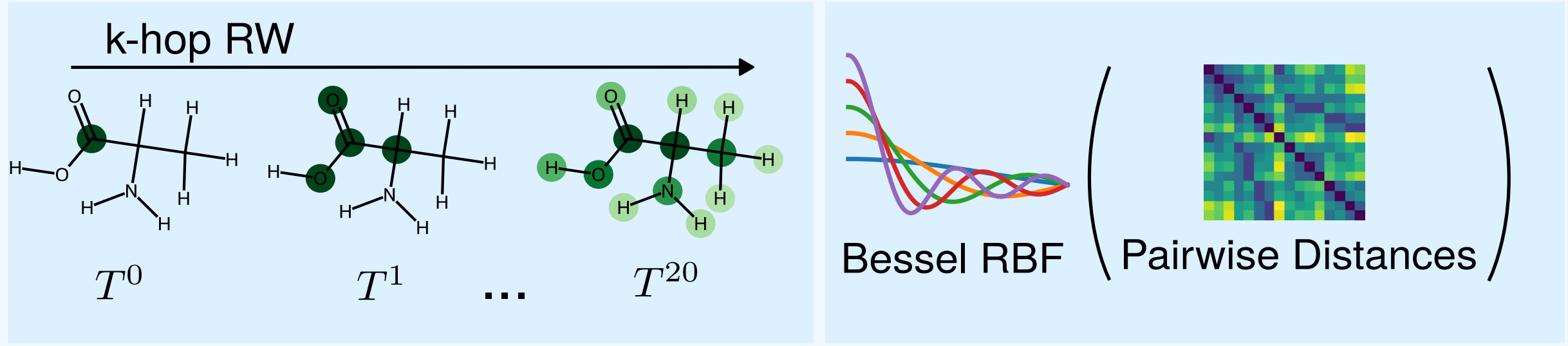
Encoder

Aggregator

REM3DI Descriptor M

Structure Encoding

- Interaction/ relevance of atoms depend on their distance. Add pair bias term to attention logits.
- Encode distances either as 3D pairwise distances, or via 2D molecular graph random walks.
- -14.96% MAE QM9 vs non-structure aware baseline

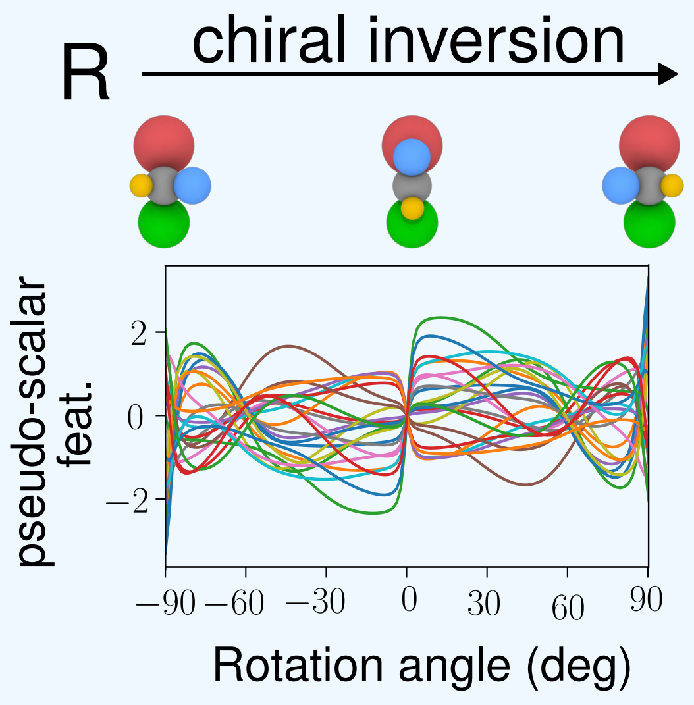


Chiral Embedding

- Construct pseudoscalar features from L=1 MLIP descriptors via triple product. Extends to L>1 features via TPs.

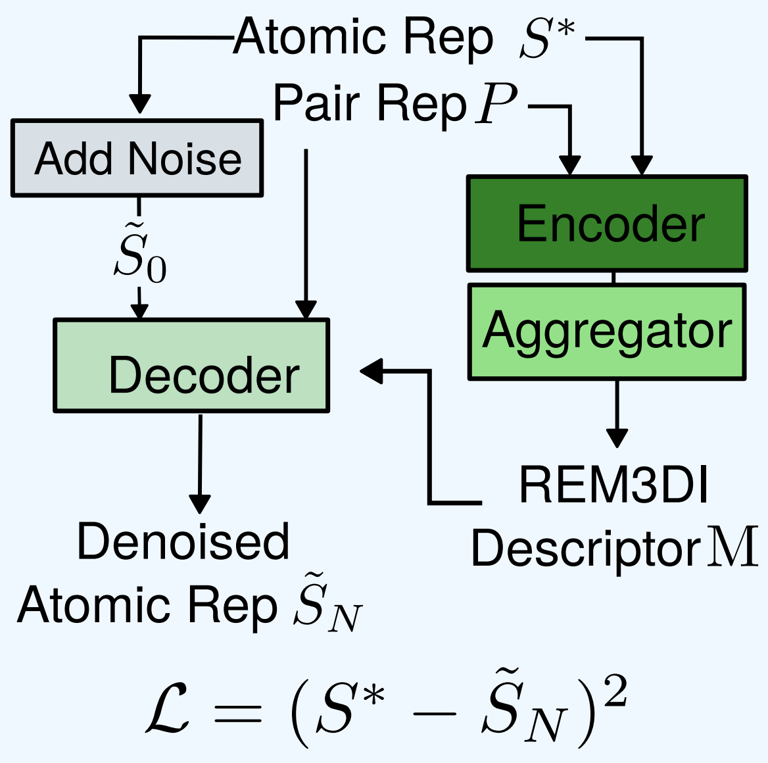
$$\mathbf{x_3} \cdot (\mathbf{x_1} \times \mathbf{x_3})$$

- Pseudoscalars are parity-odd, change sign under inversion, which enables R/S enantiomer discrimination.



Encoder Pretraining

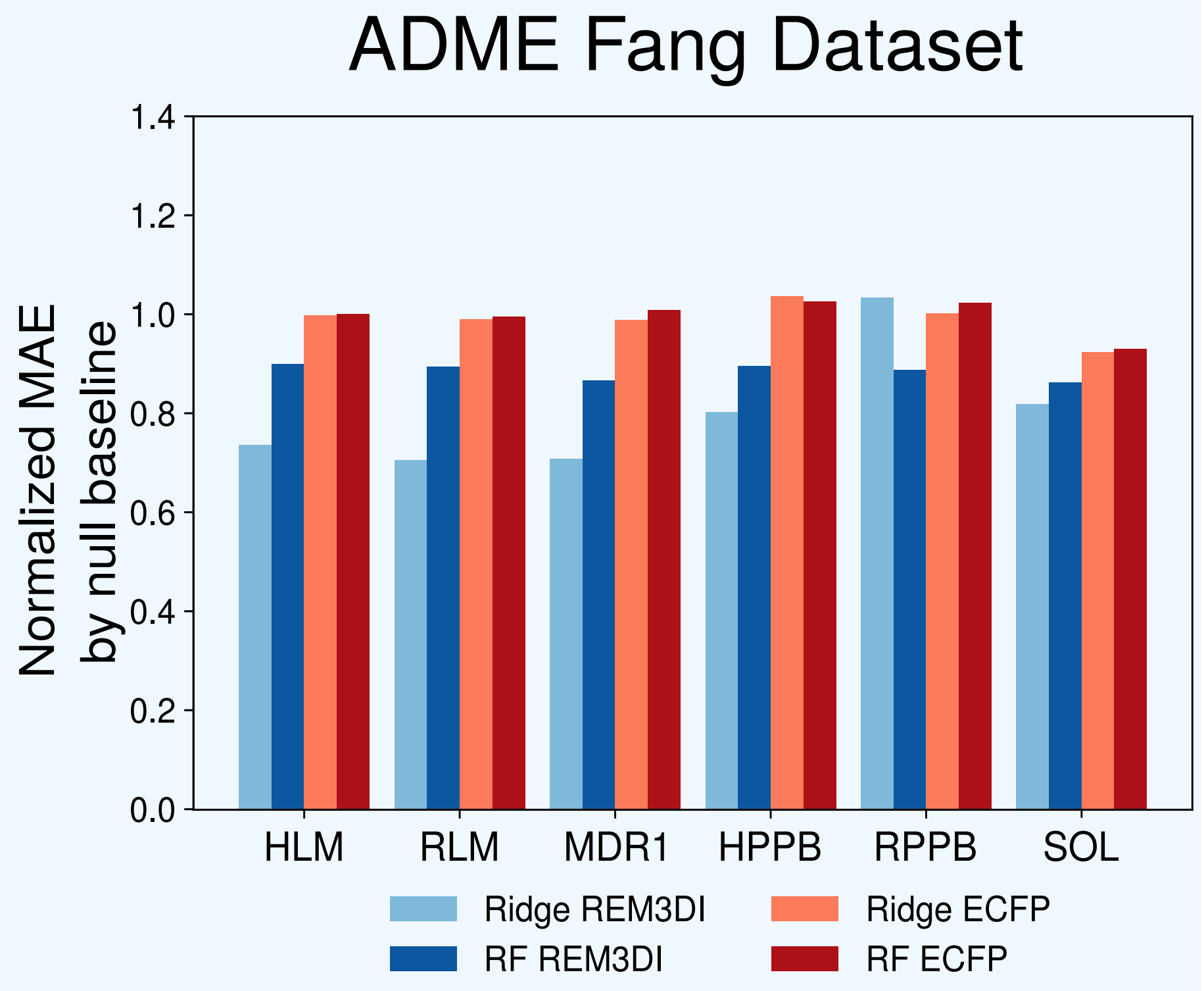
- Pretrain models on large molecular datasets via denoising.
- Improves regression performance and constructs chemically meaningful maps. (pre 400k@PCQM4M: -9.6% MAE on QM9 gap prediction)



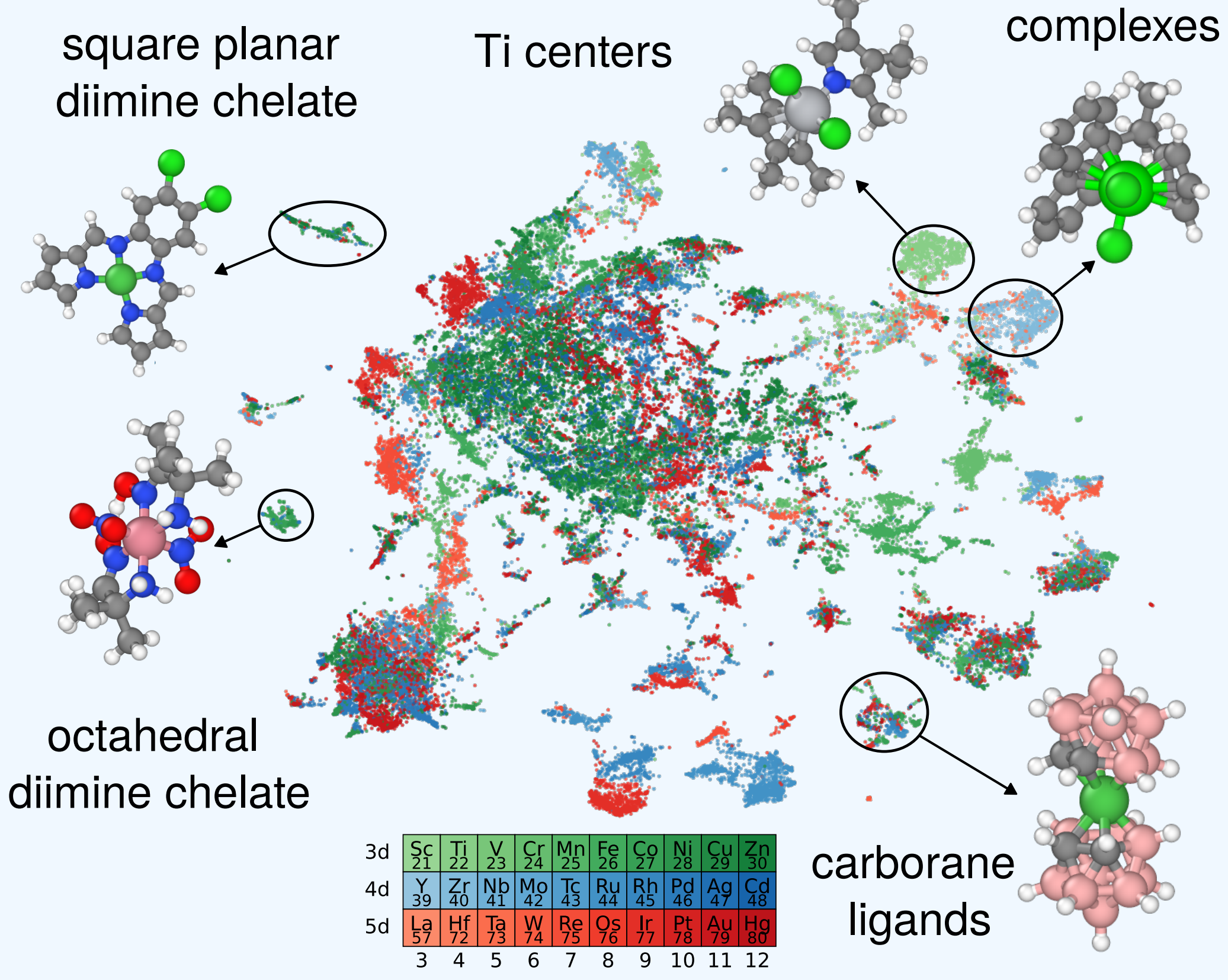
$$\mathcal{L} = (S^* - \tilde{S}_N)^2$$

Property Prediction

- Training ML regression models (ridge/ random forest) on pharmaceutical properties (ADMET/ activity) outperforms ECFP fingerprint in 12/12 properties.
- Representation transferable to wide range of tasks, and chemical systems. Does not require expensive finetuning.



Chemical Maps



- Pretraining clusters chemical space meaningfully.
- Left: Clustering of transition metal compound tmQM dataset by geometry, ligands, metal center.
- Clusters chemically and configurationally similar systems closely together.
- REM3DI is applicable to any system covered by MLIP, wider than conventional descriptors.

Summary

- REM3DI is a physics-informed and conformation-aware representation generalizing across tasks and systems from small organics to transition-metal complexes.
- We enable accurate property prediction outperforming conventional cheminformatics descriptors and practical drug discovery workflows.

Outlook

- Scale the REM3DI architecture to larger models and pretraining datasets to improve performance.
- Apply REM3DI in other atomistic domains (e.g. periodic systems) and tasks (virtual screening, MD trajectory clustering etc.)

ENCODE AI for Science · Dunia Innovations · FAST · TCM · UNIVERSITY OF CAMBRIDGE