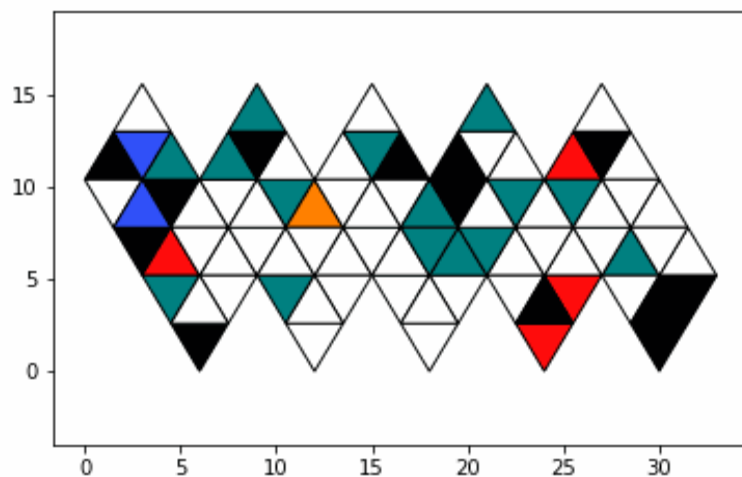


NOVEL CHEMICAL STRUCTURE INPUT METHOD

ELLA M GALE: ELLA.GALE@BRISTOL.AC.UK



Overall Motivation

We want to apply machine learning (artificial intelligence) to chemistry

But,

The big advances in ML have been focussed on specific problem domains, like image recognition and text parsing

Chemical problems are a very different domain set,

Therefore we cannot blindly apply standard ML and expect results

We need to look at how ML has dealt with issues it has faced and apply those lessons in a domain appropriate manner

e.g. we don't use the data augmentation of standard ML as it is not suitable, but we do use the approach of data augmentation, and we must develop chemically relevant data augmentation.

We need to use our chemical knowledge to formulate the problems in the easiest way for the algorithm to solve.

PROBLEMS UNIQUE TO APPLYING AI/ML TO CHEMISTRY

IMAGE/TEXT DATA*

Many labeled examples

Complete data

Can use data augmentation

New data easy to find

Data 'free' on internet
(actually 'stolen' from flickr)

* Types of data ML (esp. neural networks) has had huge success with

CHEMICAL DATA

Few labeled examples

Incomplete data

Can't use data augmentation

New data involves lab work

Some data often behind paywall, in private databases or in old textbooks

MOTIVATION 1: 3D SHAPE IS IMPORTANT

Problem	Importance of chirality / 3D shape to solution	Types of algorithms used for solution
Suggesting target molecules	Critical	(quantum mechanics, molecule mechanics)
De novo drug design	Critical	Neural networks, clustering, search, generative
Forward reaction prediction	Medium (needs to be tracked)	Neural networks, general ML, search
Retrosynthesis	High (needs to be tracked)	Search, neural networks
Condition recommendation	Low	Search, neural networks
Structure assignment	Critical	General ML, neural networks,

MOTIVATION 2: CHEMICAL DATA IS SMALL

**Small compared to
chemical space**





Chemical Space

10⁶⁰

Potential Pharmacological molecules possible

Image source: © Alexandre Borrel, PhD

The ChemMaps tool allows the user to visualise a structure as a star within a galaxy of other compounds whose closeness reflects their structural similarity see ChemMaps.com



Chemical Space

<0.0000000001%

Of Possible Molecules Have Been Synthesised

Source: © Alexandre Borrel, PhD

The ChemMaps tool allows the user to visualise a structure as a star within a galaxy of other compounds whose closeness reflects their structural similarity

MOTIVATION 2: CHEMICAL DATA IS SMALL

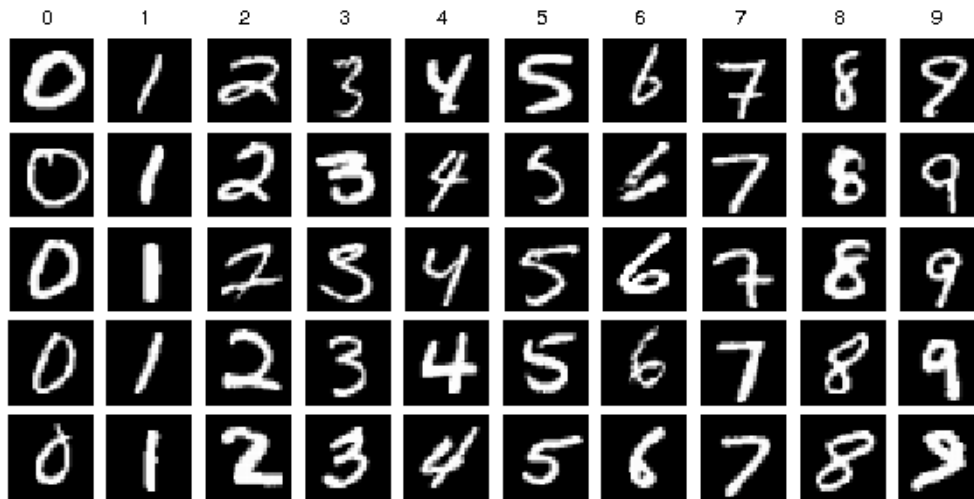
Small compared to chemical space

Small compared to datasets used successful in ML



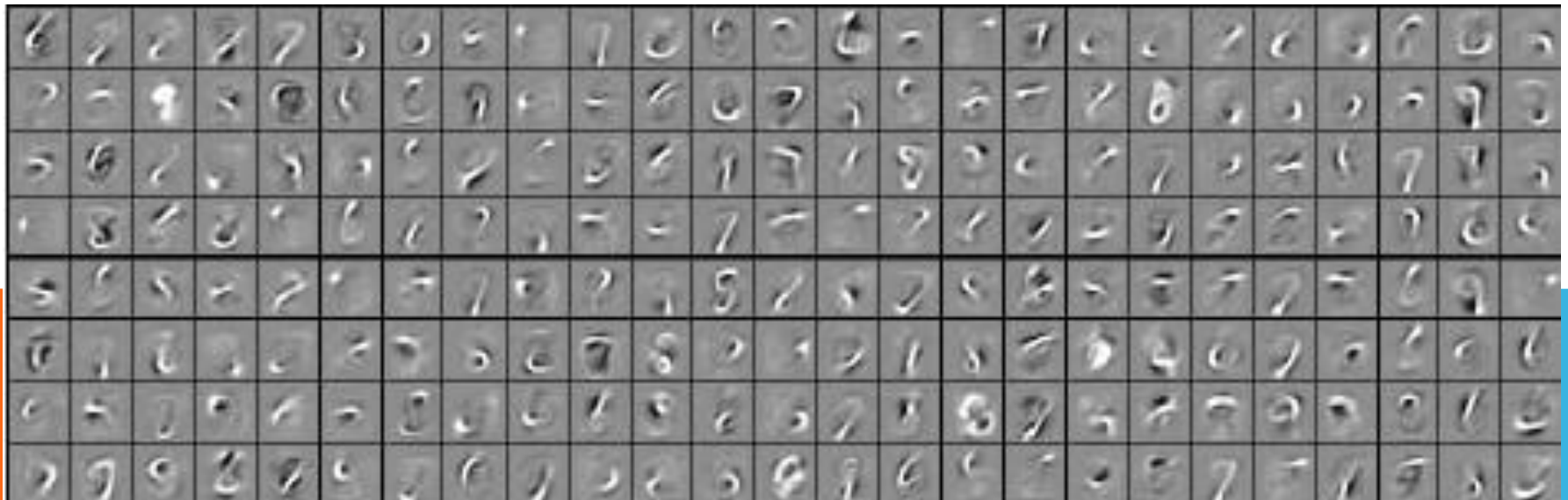
THE MNIST TASK IS 'UNDERSTANDABLE'

MNIST:



Learned features

Fried, Ohad, and Rebecca Fiebrink. "Cross-modal Sound Mapping Using Deep Learning." *NIME*. 2013.



DATA USED IN VISUAL CLASSIFICATION TASKS



IMAGENET:

Input data: colour
244x244 pixel
image

No. Classes =
1000

No. Examples:,
1,300,000

No. examples per
class: 1,300

MOLECULENET: CHEMICAL BENCHMARKS

Table 1: Dataset Details: number of compounds and tasks, recommended splits and metrics

Category	Dataset	Data Type	# Tasks	Task Type	# Compounds	Rec - Split	Rec - Metric
Quantum Mechanics	QM7	SMILES, 3D coordinates	1	Regression	7160	Stratified	MAE
	QM7b	3D coordinates	14	Regression	7210	Random	MAE
	QM8	SMILES, 3D coordinates	12	Regression	21786	Random	MAE
	QM9	SMILES, 3D coordinates	12	Regression	133885	Random	MAE
Physical Chemistry	ESOL	SMILES	1	Regression	1128	Random	RMSE
	FreeSolv	SMILES	1	Regression	642	Random	RMSE
	Lipophilicity	SMILES	1	Regression	4200	Random	RMSE
Biophysics	PCBA	SMILES	128	Classification	437929	Random	PRC-AUC
	MUV	SMILES	17	Classification	93087	Random	PRC-AUC
	HIV	SMILES	1	Classification	41127	Scaffold	ROC-AUC
	PDBbind	SMILES, 3D coordinates	1	Regression	11908	Time	RMSE
	BACE	SMILES	1	Classification	1513	Scaffold	ROC-AUC
	BBBP	SMILES	1	Classification	2039	Scaffold	ROC-AUC
Physiology	Tox21	SMILES	12	Classification	7831	Random	ROC-AUC
	ToxCast	SMILES	617	Classification	8575	Random	ROC-AUC
	SIDER	SMILES	27	Classification	1427	Random	ROC-AUC
	ClinTox	SMILES	2	Classification	1478	Random	ROC-AUC

Size of Chemical Benchmark Datasets:

Min: 642

Max: 437,929

Average: 7210 \Pm 26,071.9

Size of visual benchmark datasets

MNIST: 60,000

cats versus dogs:25,000

Size of IMAGENET: 1,300,000

Taken from: Wu, Zhenqin, et al. "MoleculeNet: a benchmark for molecular machine learning." *Chemical science* **9.2** (2018): 513-530.

SOLUTION TO SMALL DATASETS: DATA AUGMENTATION

Visual data augmentation

- **Crop**
- **Rotation**
- **Reflection!**

Original



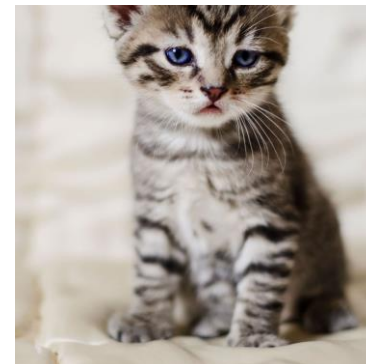
Reflection



Rotation



Crop



SOLUTION TO SMALL DATASETS:



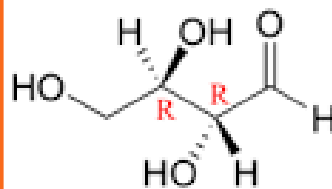
- Crop
- Rotation
- Reflection

Rotation

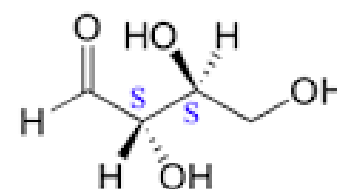


Crop

But how would this work with molecules???



D-erythrose

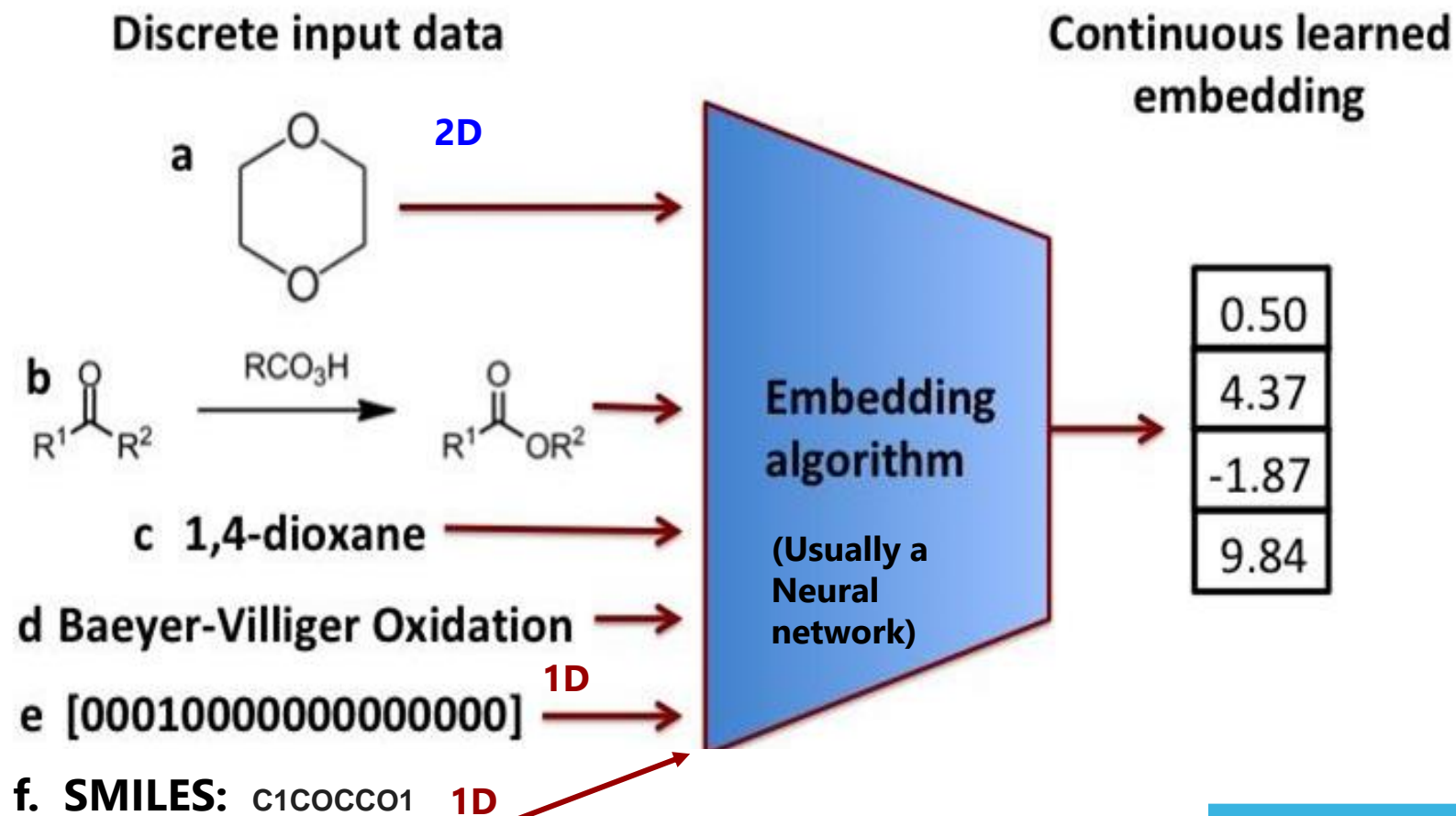


L-erythrose

Premise:

- Encode the 3D structure and symmetry properties
- Use deep-neural networks (NNs)
- Add in data augmentation
- We need a simpler benchmark tasks!

ENCODING CHEMICALS



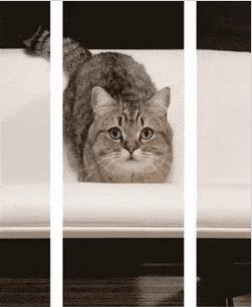
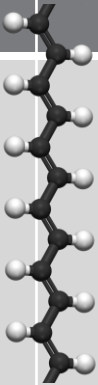

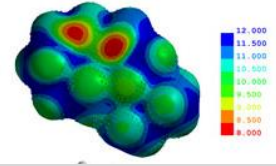
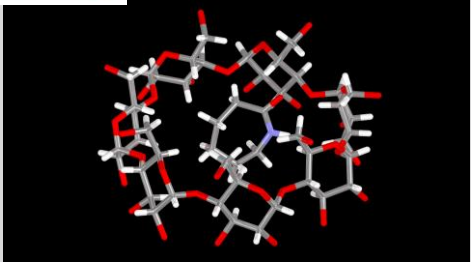


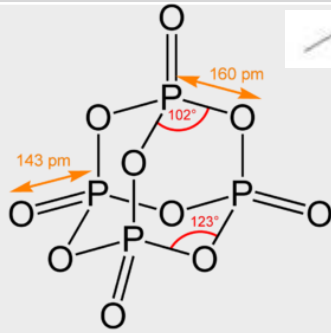
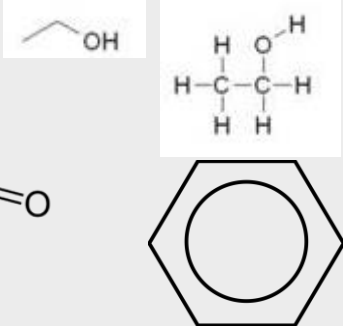


See:

[1] E. Gale & D. Durand, "Improving Reaction Prediction, *Nature Chemistry*, 12, 509-510 (2020)"

[2] G. Grethe et al., "International chemical identifier for reactions" (RInChI)", *J. Cheminformatics*, " (2018)

[3] Graphical representation standards for chemical reactions (*IUPAC Recommendations 2019*), L.S. Press, J.B. Press, K.T. Taylor

HOW DIFFICULT IS THE PROBLEM?

Data type	Human	Humanity	Chemistry
3-D	~6 months 	-infinity! 	  
2-D	~18 months 	30,000 BCE 	 
1-D and symbolic	~3-4 years 	3400 BCE 	C1COCCO1 C2H5OH Ethanol CAS 639052-78-1 [0000010100001100 ... 0]

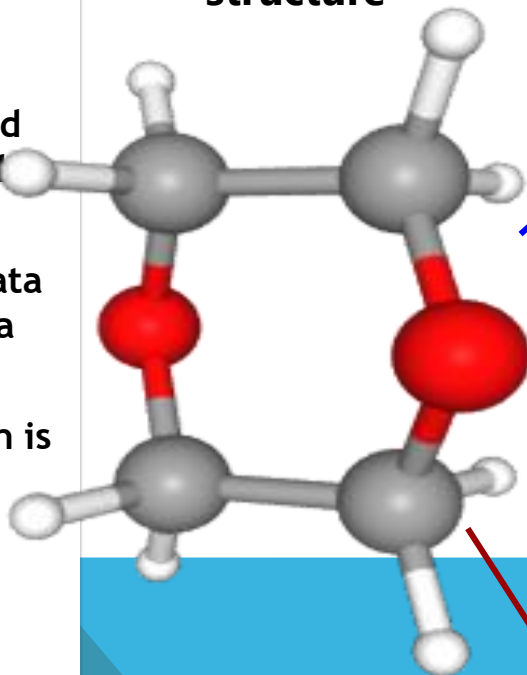
Input

ENCODING CHEMICALS (NEW)

Problems:

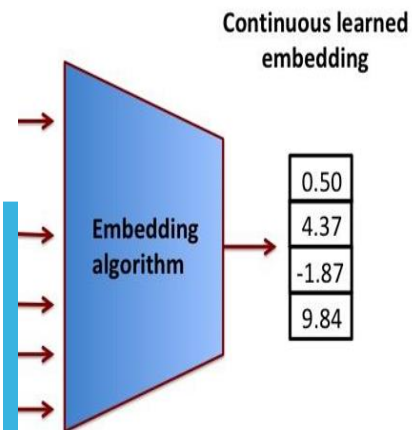
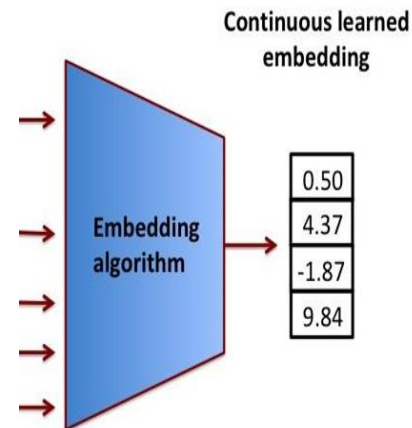
1. We want to retrain 3D structure for chirality
2. Deep-NNs have had the most success with 2D image data
3. We need lots of data for NNs to learn (data augmentation).
4. Data augmentation is usually done using reflection (not appropriate for chemistry)

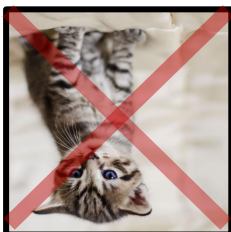
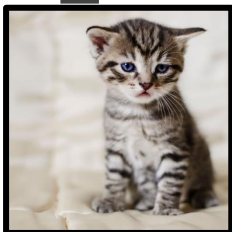
3D structure



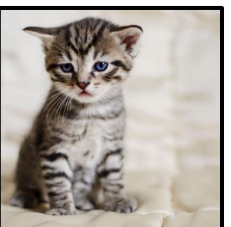
3D input can be used with spherical or 3D neural nets.

3D input can be converted to a novel 2D input that retains rotational symmetry and chirality. This will allow a speed-up compared to 3D NNs

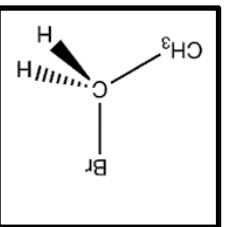
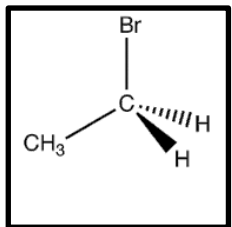




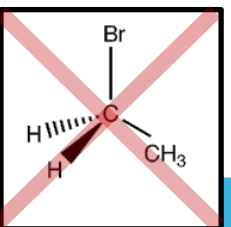
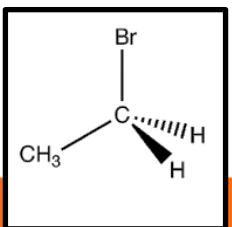
SYMMETRY, ROTATION, AUGMENTATION



- Using different rotations of a molecule is acceptable augmentation
- Using reflection operations is not



- We need to preserve:

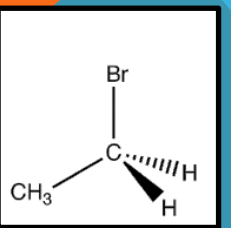
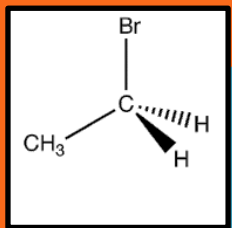


✧ Rotational symmetry

✧ Chirality

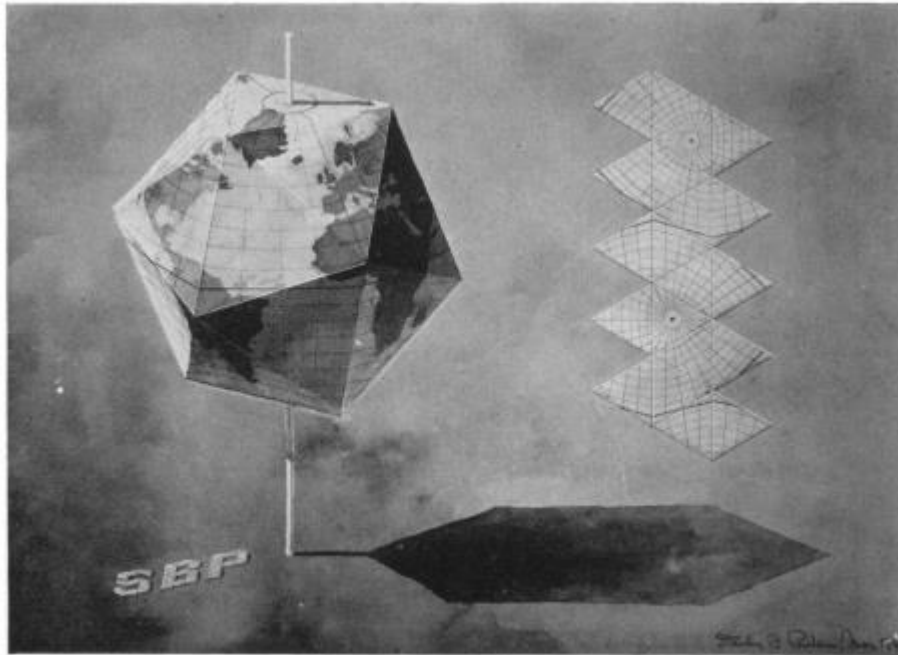
✧ Global structure

✧ Translational invariance

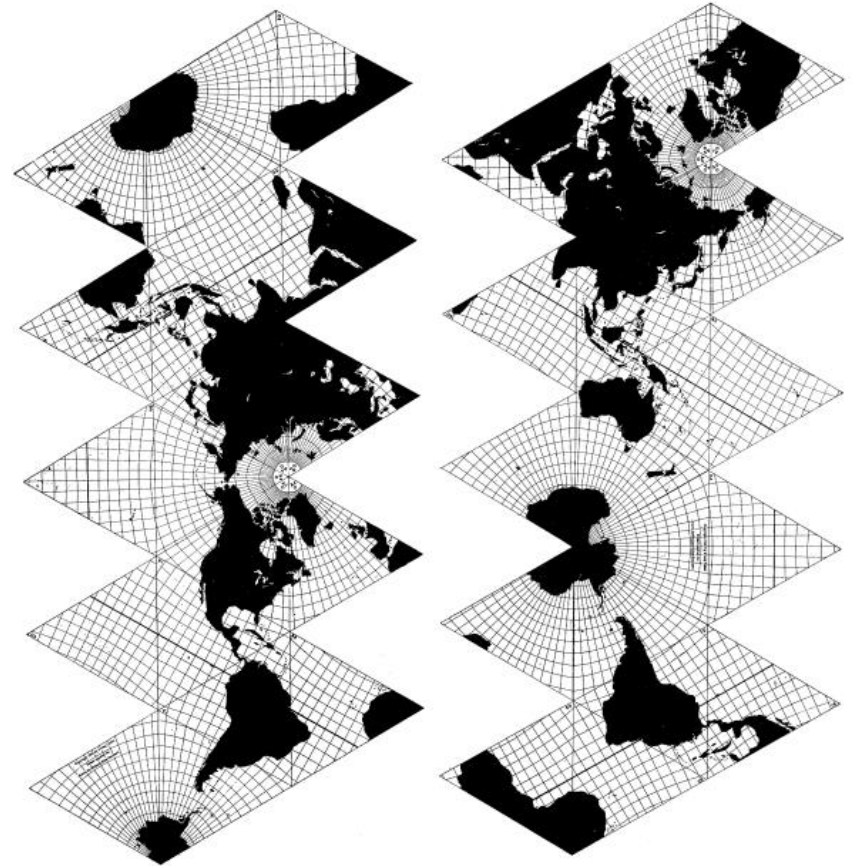


ICOSAHEDRONS OFFER THE SMALLEST DISTORTION WHEN GOING FROM 3D TO 2D

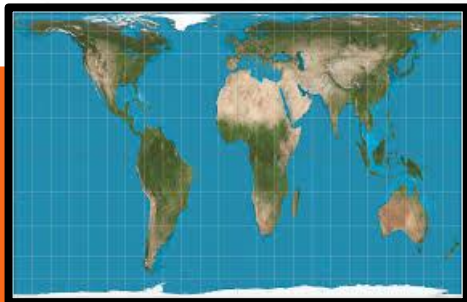
Icosahedron (icosphere level 0)



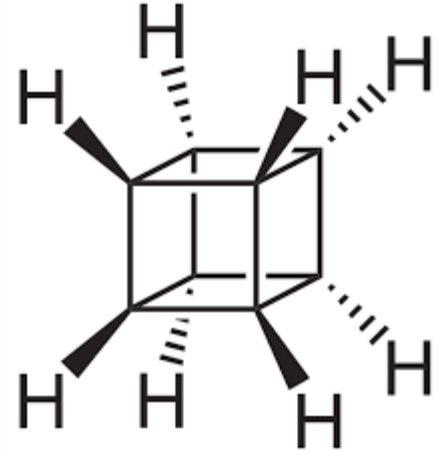
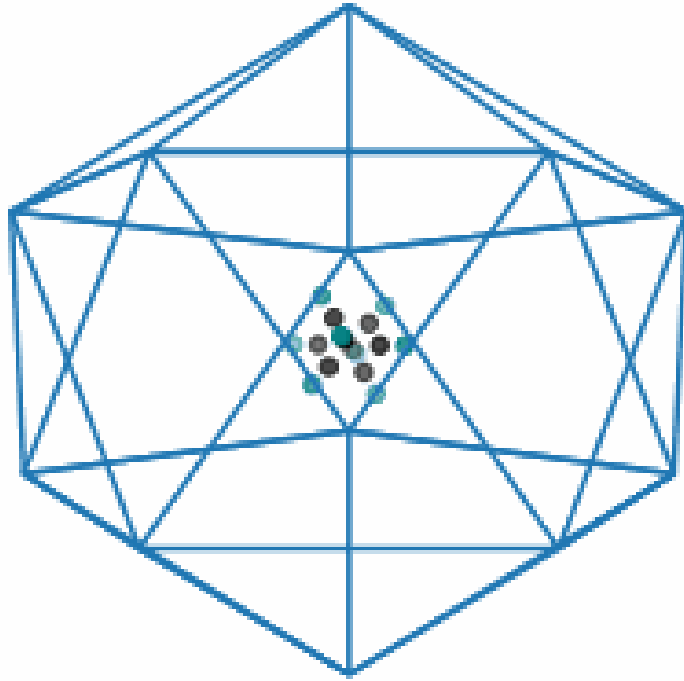
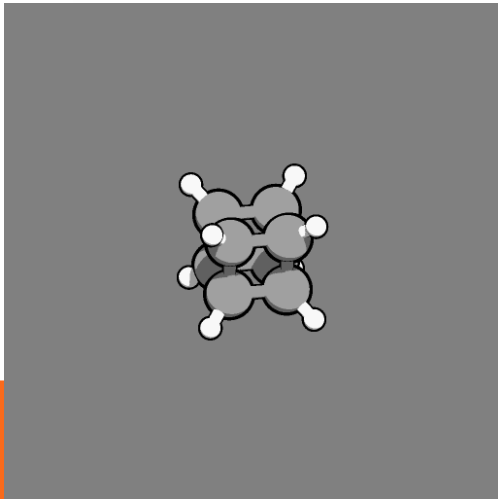
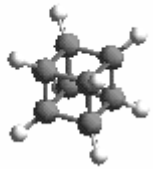
Icosahedron nets



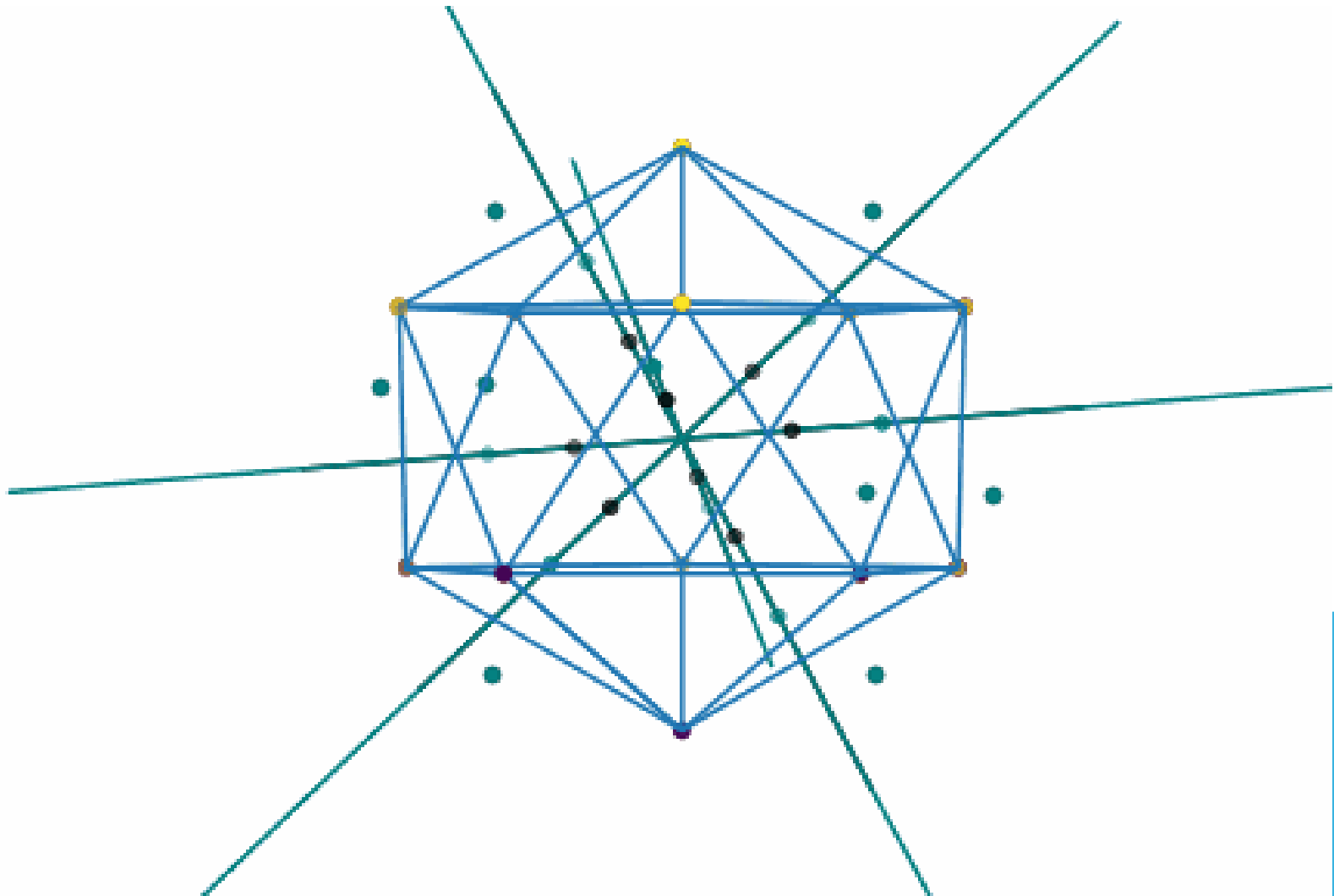
Irving Fisher. A world map on a regular icosahedron by gnomonic projection.
Geographical Review, 33(4):605–619, 1943



1. ENCAPSULATE MOLECULE IN AN ICOSAHDEDRON

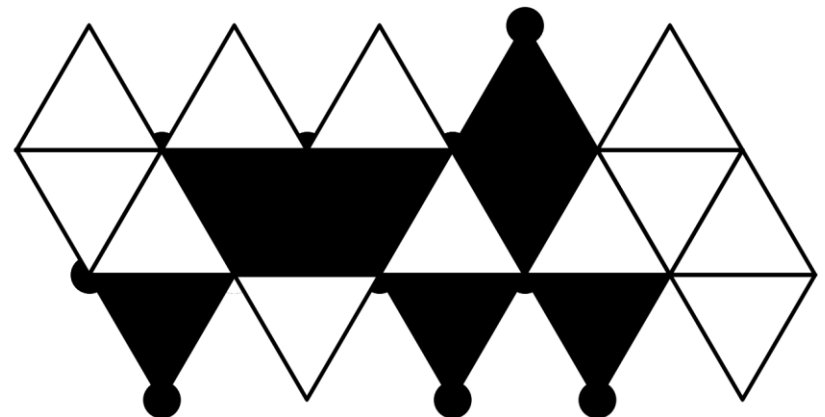
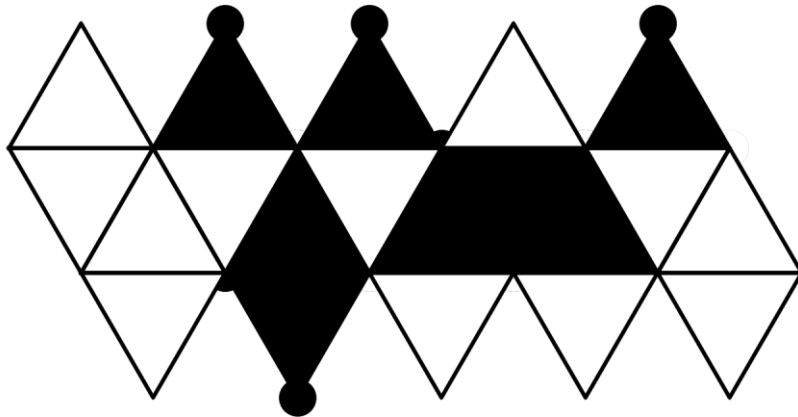
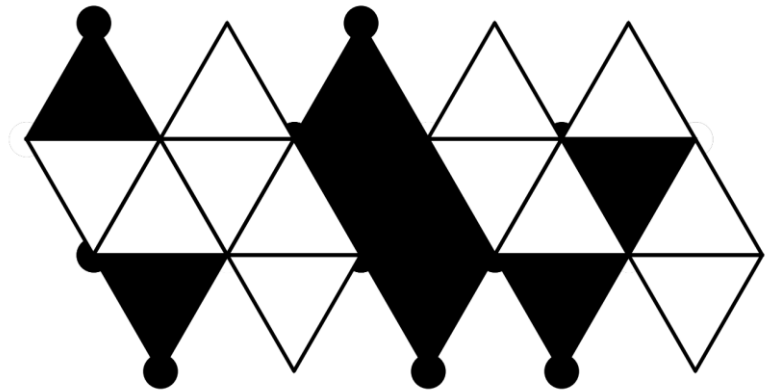
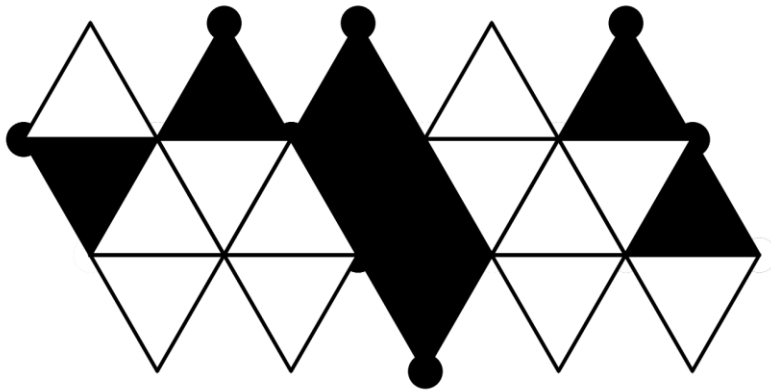


SCALE AND PROJECT ATOMS FROM CENTRE OF MASS TO SURFACE OF ICOSAHEDRON

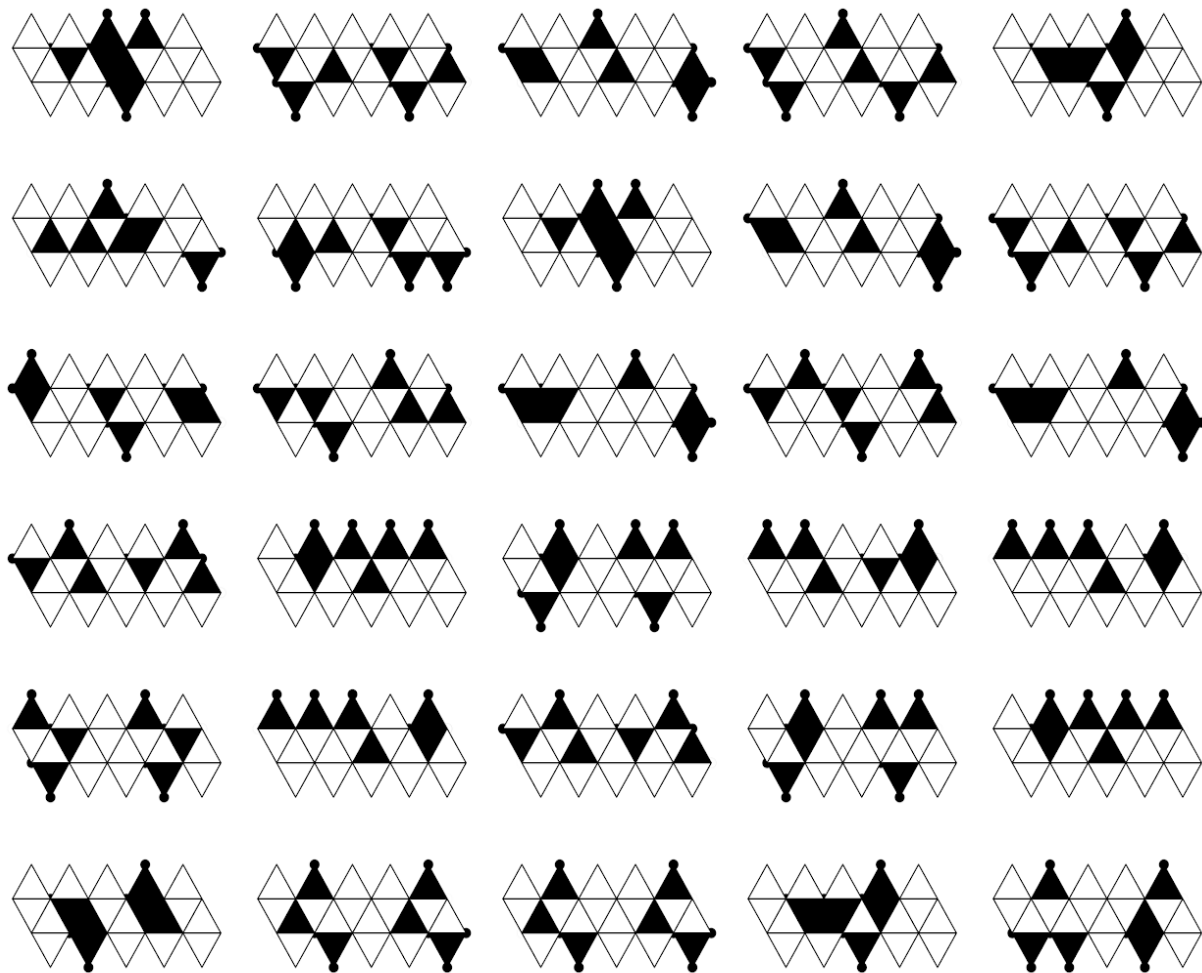
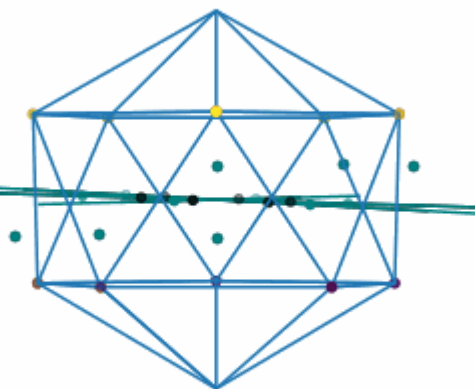
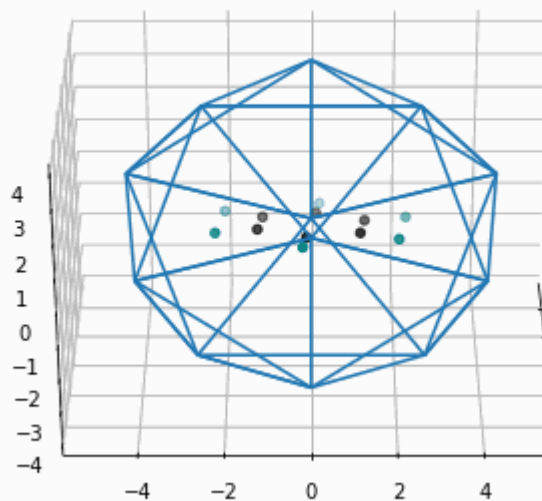


COLOUR PANELS OF ICOSAHEDRON AND UNFOLD

There are 60 possible nets. Hydrogens not shown

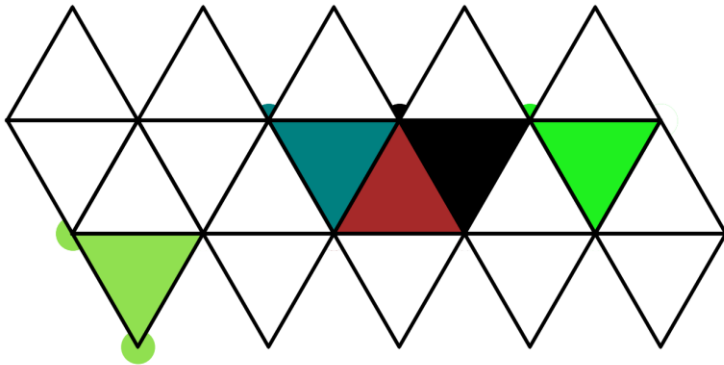


BENZENE Nets (unfoldings)

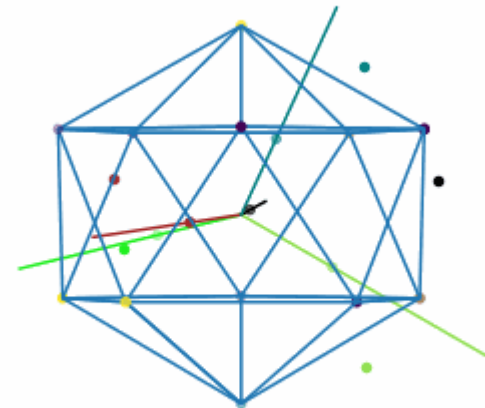
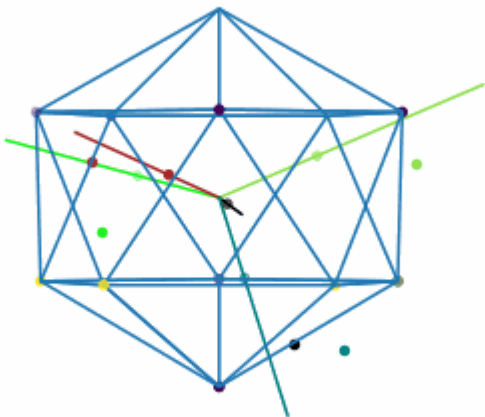
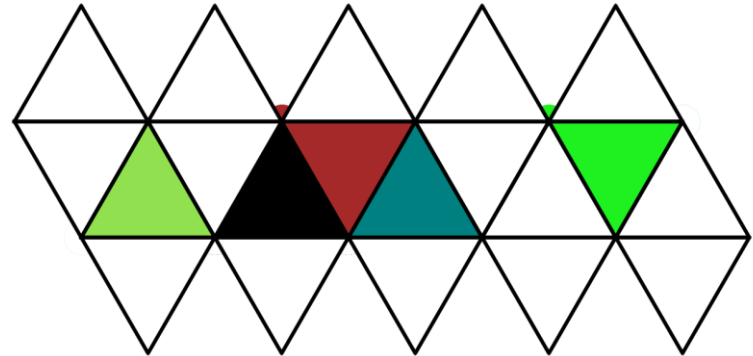


ENANTIOMERS

R-BClFmethane

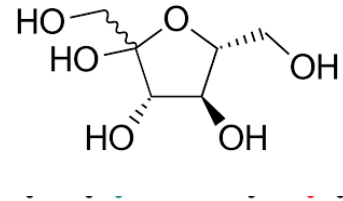
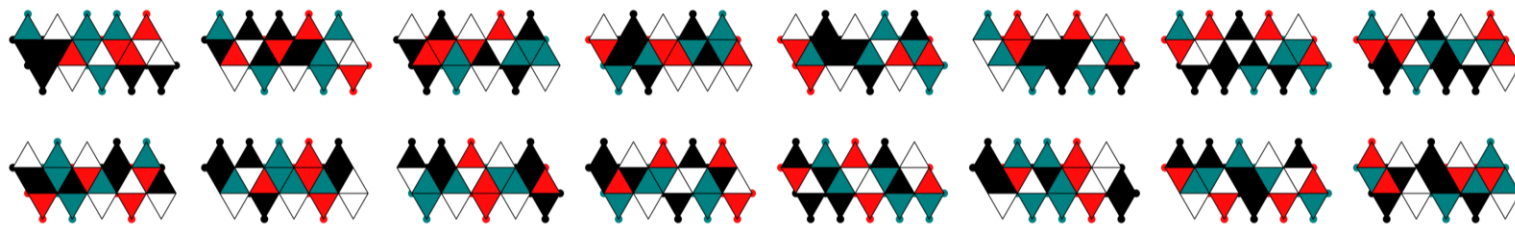


S-BClFmethane

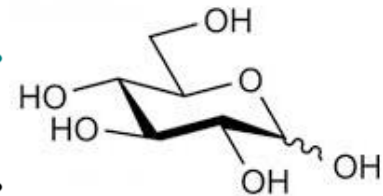
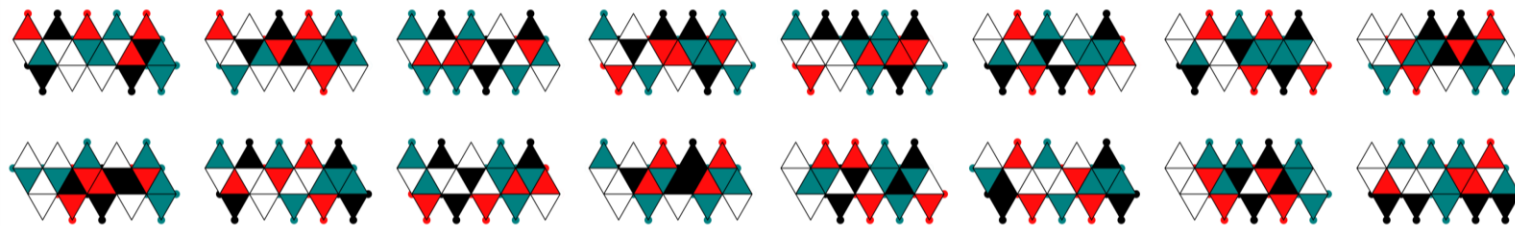


SUGAR EXAMPLE

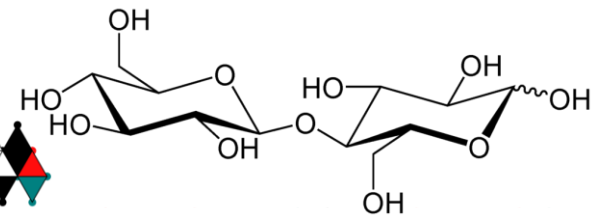
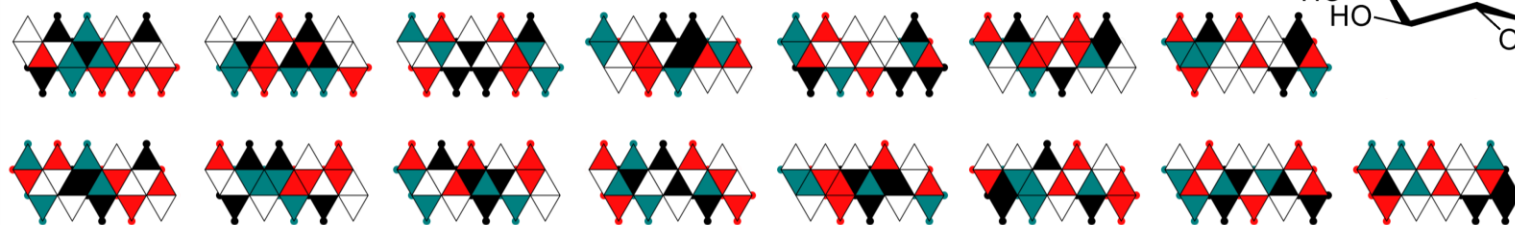
Fructose



Glucose

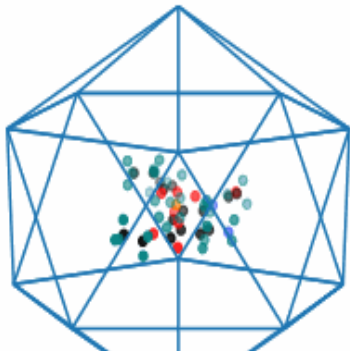


Cellubiose



Input

Icosahedron



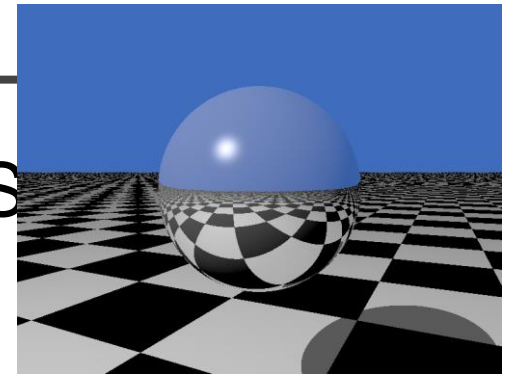
Icosphere 1

ENCOD



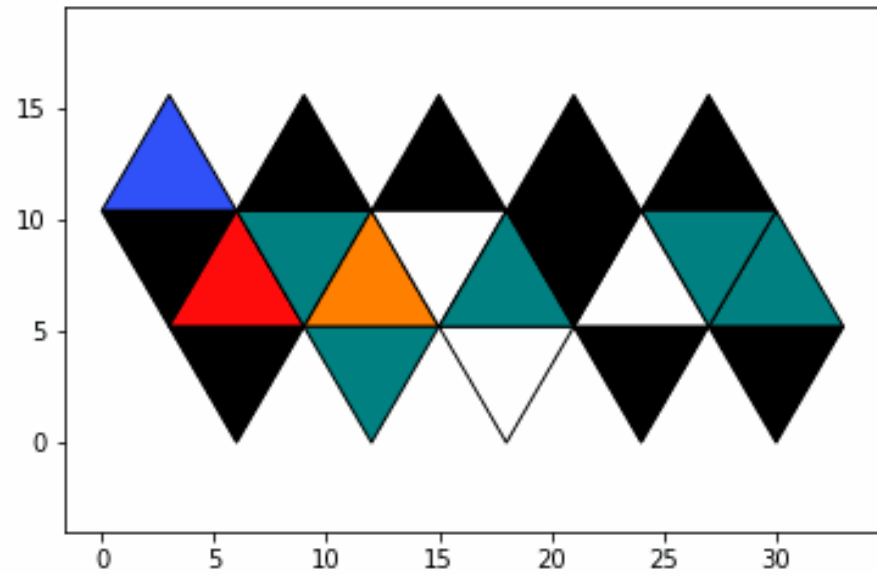
Tamiflu

CALS

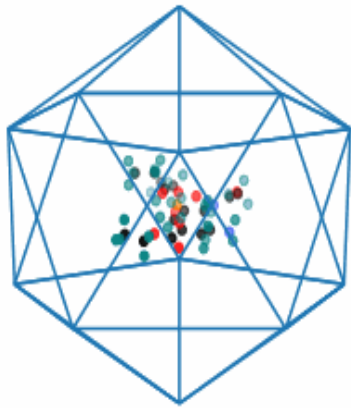


Icosphere 1

Ic

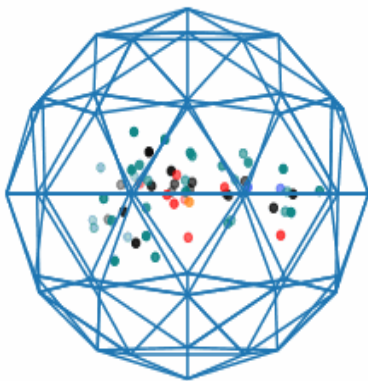


COARSE-GRAINING

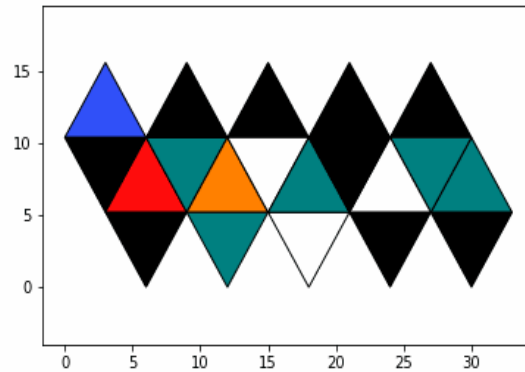


Icosahedron

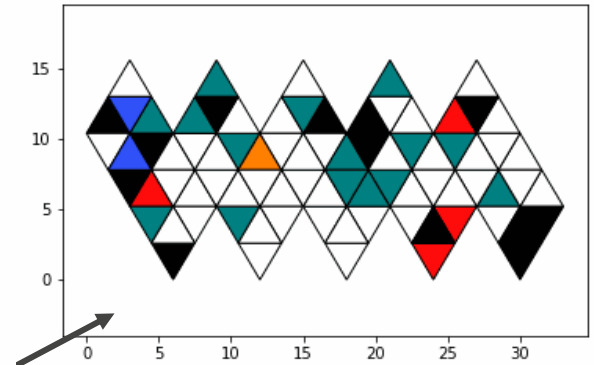
Icosphere 1



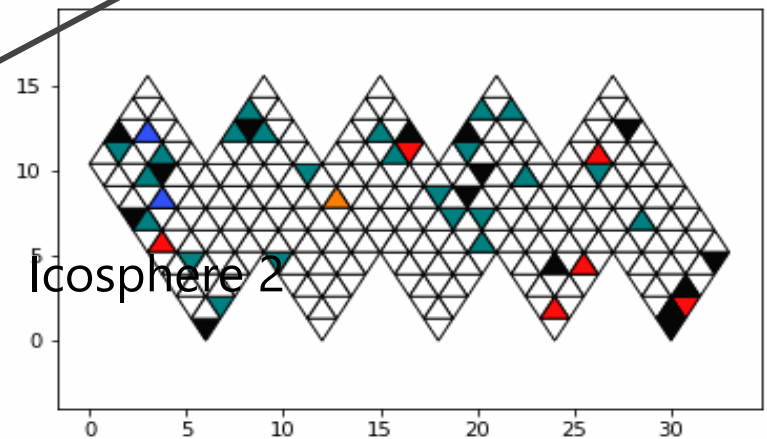
Tamiflu



Icosphere 1



Icosphere 2

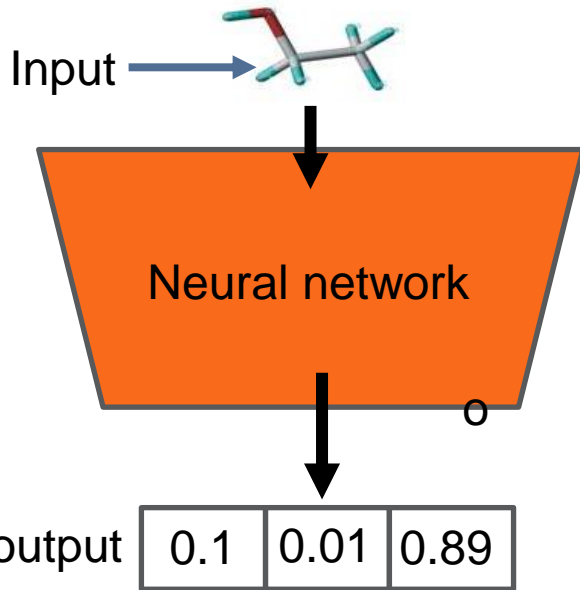


METHODOLOGY

(PRELIMINARY)

TYPES OF DEEP NEURAL NETWORK TASKS AND ARCHITECTURE

Classification



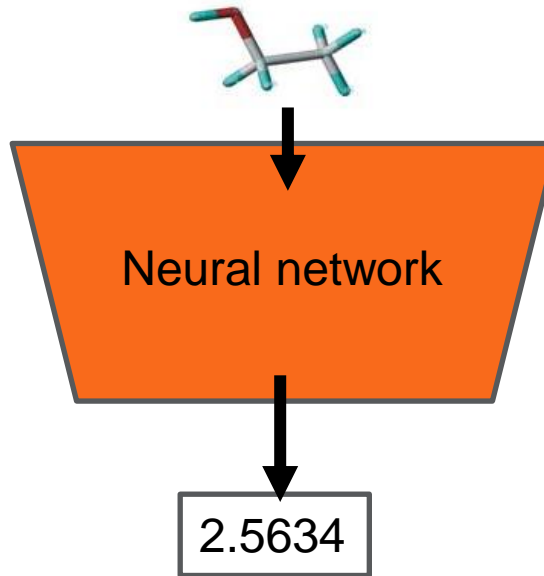
Ground truth
(correct answer)

0.0	0.0	1.0
-----	-----	-----

$[0,0,1]$ = alcohol

Interpretation

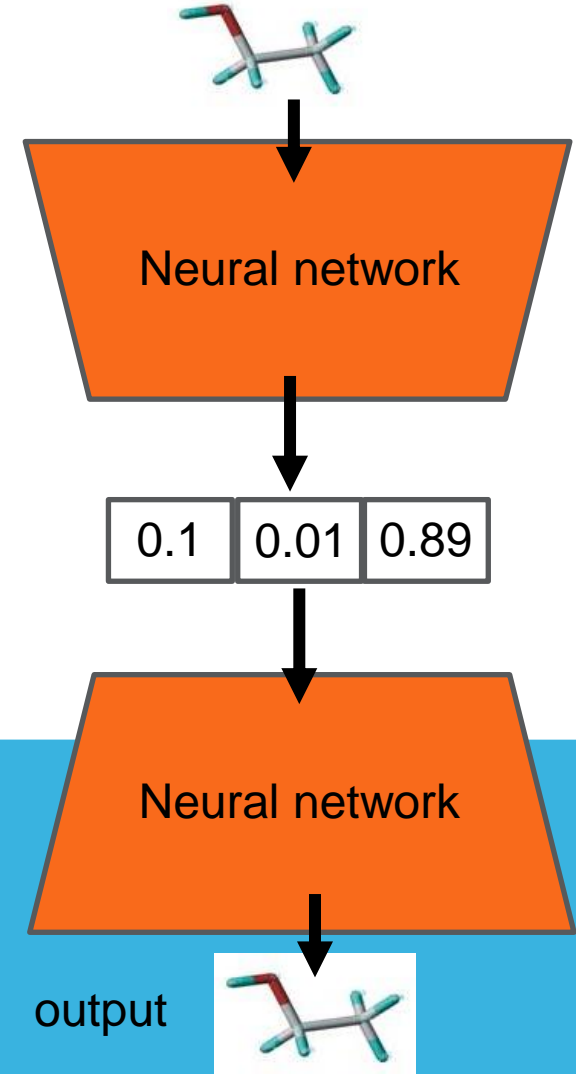
Regression



2.6102

Radius of
gyration

Autoencoder



Neural network

output

EXPERIMENTAL ARCHITECTURES

SET-UP

SMILES string



**3D-Structure
(minimised)**



Icosphere nets



Spherical NN



Output target

CONTROL

SMILES string



**Morgan fingerprint
(binary data)**



Neural network



Output target

EXPERIMENTAL DETAILS

DATA SETS

- a. 1000 nets for 1000 molecules**
 - a. 60,000 nets for 1000 molecules**
- Icospheres level 1, 2 and 3 tested**

TESTING

Molecular ID:

10% of nets per molecule taken as test set

Everything else:

10% of molecules taken as test set

Generalisation to new molecules

(has the NN learned the general rule?)

**Actual input data
is:**

RGB where

**R = mass of inner
most atom**

**G = mass of outer
most atom**

**B = sum of atom
masses**

Eg:

C-H is

[12.0107, 1.0078, 13.0185]

C=O is

[12.0107, 15.999, 28.0097]

O-H is

[15.999, 1.0078, 17.0068]

Also:

Problems can be hard because the NN is underpowered or the dataset is too small

e.g.:
Classifying 1000 classes from 60,000 datapoints is harder than classifying 10 classes from 60,000 data points

DIFFICULTY OF PROBLEM

EASY:

- Entirely easily derivable from input data only.

(Ella can figure out the rules to do it in under 5 minutes)
e.g. counting atom types,
summing atomic mass

MEDIUM:

- Requires some chemical knowledge that might not be given

e.g. counting valence electrons,
learning what a heteroatom is (to count them)
counting H donors etc

HARD:

- Requires a spatial, electronic or quantum knowledge

e.g. calculating shape parameters like principal moments of inertia (and derived quantities like...)
MolLogP



RESULTS

(PRELIMINARY)

Task: classification
/ autoencoder
Level: **hard**

Input data:

Input no: 1000 /
60,000

Level: 2

No. classes: **1000**

Output target:
ID number

Test set: unseen
nets

Training details:
Epochs: 160

HAS IT LEARNED TO 'READ' THE NETS?

Task	Augmentation	Acc. On train	Acc. On test
Random (baseline)	n/a	0.1%	0.1%
Classification	no	100%	1%
Classification	yes	100%	60%

Conclusion:

Yes. But it can only differentiate 60% of them.

Task: regression

Level: easy

Input data:

Input no: 1000

Level: 2

Augmentation: no

Output target:

Exact molar weight

Test set: unseen
molecules

Results:

$R^2 = 0.812$

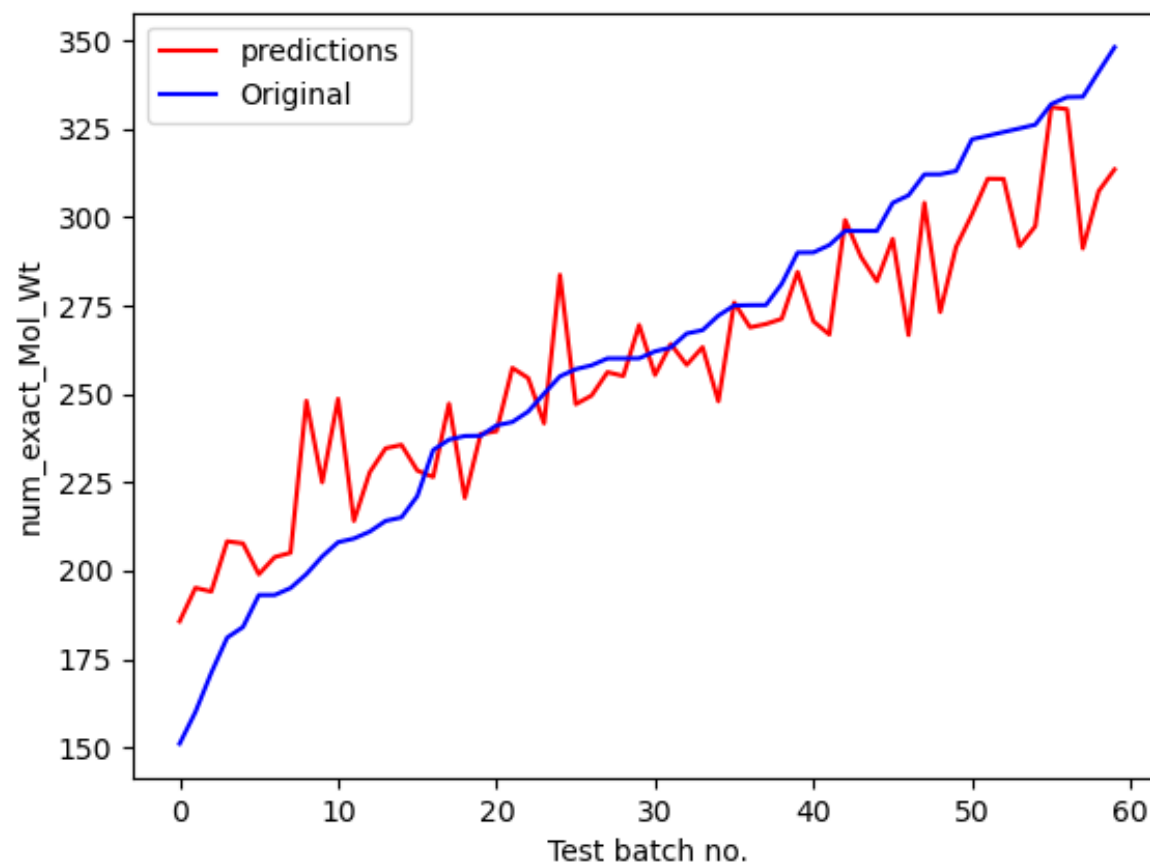
MSE = 580

Training details:

Epochs: **60**

Plotted: test set

EFFECT OF AUGMENTATION 1: WITHOUT AUGMENTATION



Task: regression

Level: Easy

Input data:

Input no: 1000

Level: 2

Augmentation: **yes**

Output target:

Exact molar weight

Test set: unseen
molecules

Results:

$R^2 = \mathbf{0.979}$

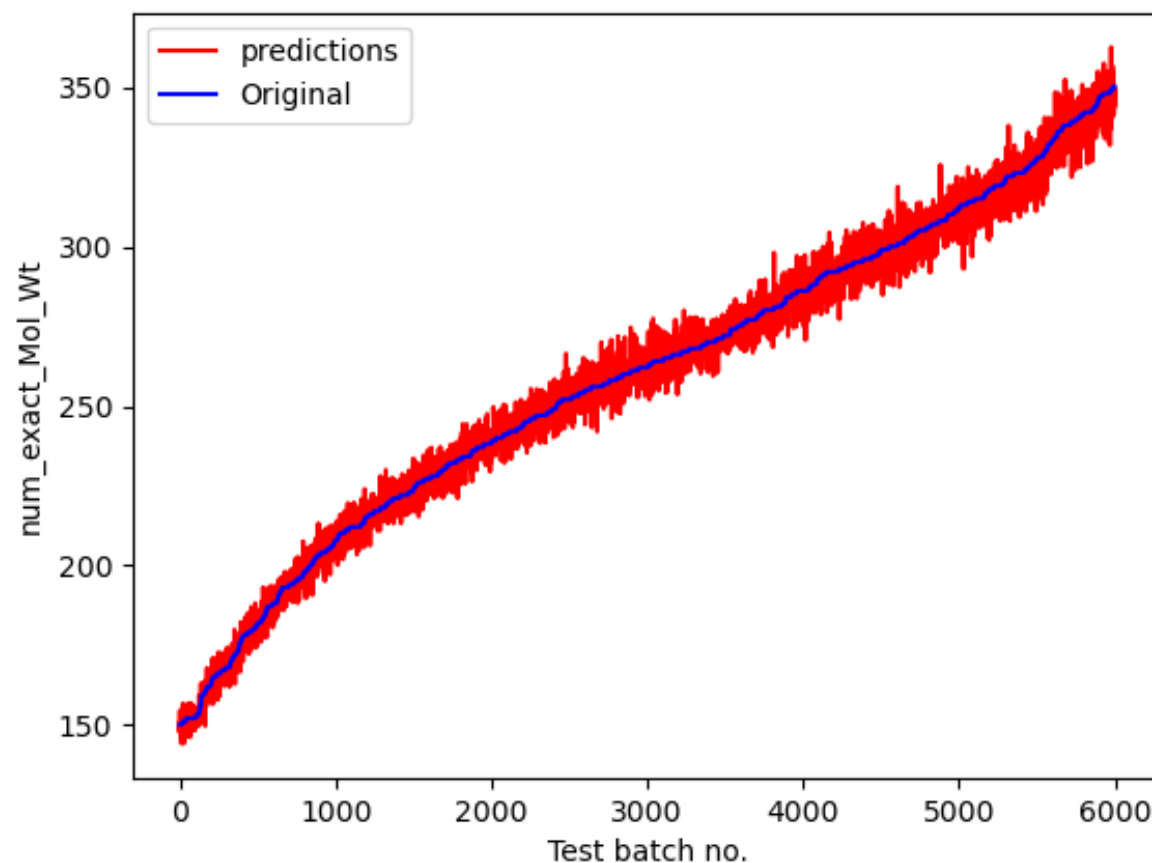
MSE = **23**

Training details:

Epochs: **60**

Plotted: test set

EFFECT OF AUGMENTATION: WITH AUGMENTATION



Task: regression

Level: Medium?

Input data:

Input no: 1000

Level: 2

Augmentation: yes

Output target:

Number of valence
electrons

Test set: unseen
molecules

Results:

$R^2 = 0.9824$

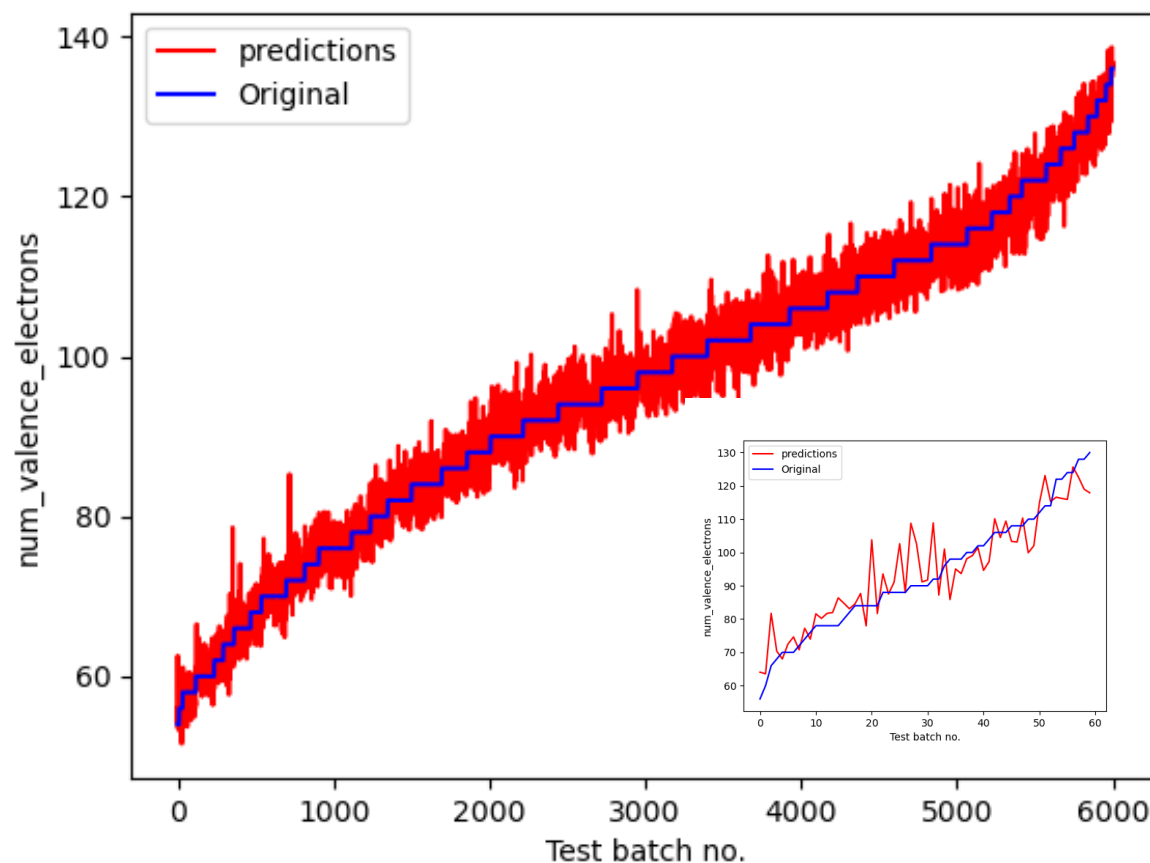
MAE = 21.9

Training details:

Epochs: 160

Plotted: test set

EFFECT OF AUGMENTATION: WITH AUGMENTATION



Task: regression
Level: hard

Input data:

Input no: 1000

Level: 2

Augmentation:
yes

Output target:

Radius of gyration

Test set: unseen
molecules

Results:

$R^2 = 0.886$

$MAE = 0.138$

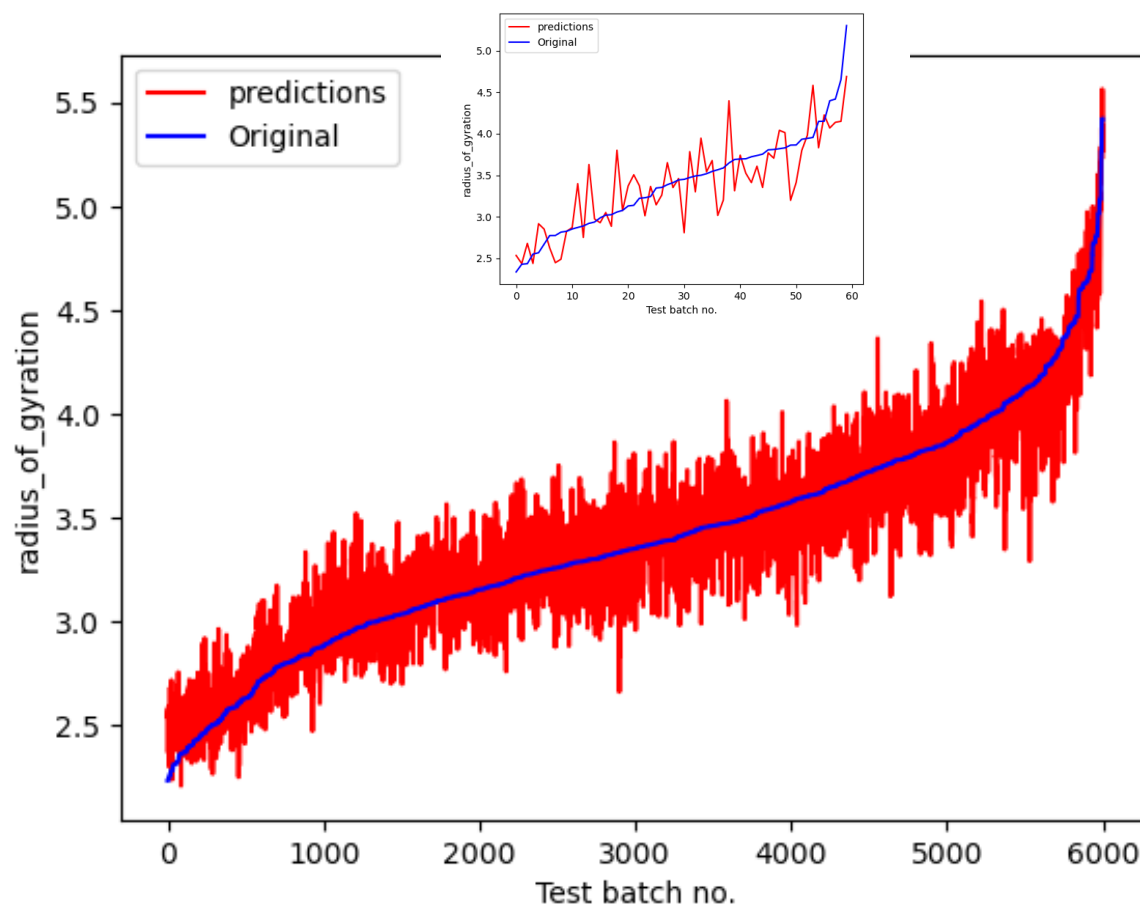
Training details:

Epochs: 160

Plotted: test set

CAN IT LEARN SHAPE PARAMETERS?

(I.E. THOSE BASED ON MOMENTS OF INERTIA)



Task: regression

Level: hard

Input data:

Input no: 1000 /
60,000

Level: 2

Output target:

Radius of gyration

Test set: unseen
molecules

Training details:

Epochs: 160

MATCH THE INPUT DATA TO THE PROBLEM - 1

Input	Output	Augmentation	Difficulty	MAE	R2
SMILES	Radius of gyration	no	Hard-impossible?	1.1	-0.1167
Icosphere level 2	Radius of gyration	no	medium	0.258	0.667
Icosphere level 2	Radius of gyration	yes	medium	0.138	0.886

Nets contain mass and relative angle information →
NN can infer about shape

SMILES strings do not contain shape information
(without the application of chemical know how
anyway) → NN cannot learn the underlying function

Task: classification

Level: easy -
medium

Input data:

Input no: 1000 /
60,000

Level: 2

No. classes: 4

Output target:
charge

Test set: unseen
molecules

Training details:

Epochs: 300

MATCH THE INPUT DATA TO THE PROBLEM - 2

Input	Output	Augmentati on	Difficulty	Acc. On test
SMILES	charge	no	easy	86%
Icosphere level 2	charge	no	medium	0.49%
Icosphere level 2	charge	yes	medium	48.9%

Both SMILES and nets contain the relevant information.

SMILES → charge is the easier problem.

SUMMARY

- Presented a method for data augmentation
 - Presented a method to include information about 3D structure and symmetry
 - Creating small benchmark datasets (to be done before attempting the MoleculeNet benchmarks)
-

FURTHER DIRECTIONS

Further data augmentation can be done by:

- Symmetry breaking by distorting along vibrational directions (normal modes)
- Including rotation of the molecule (not tested yet)

Testing the method

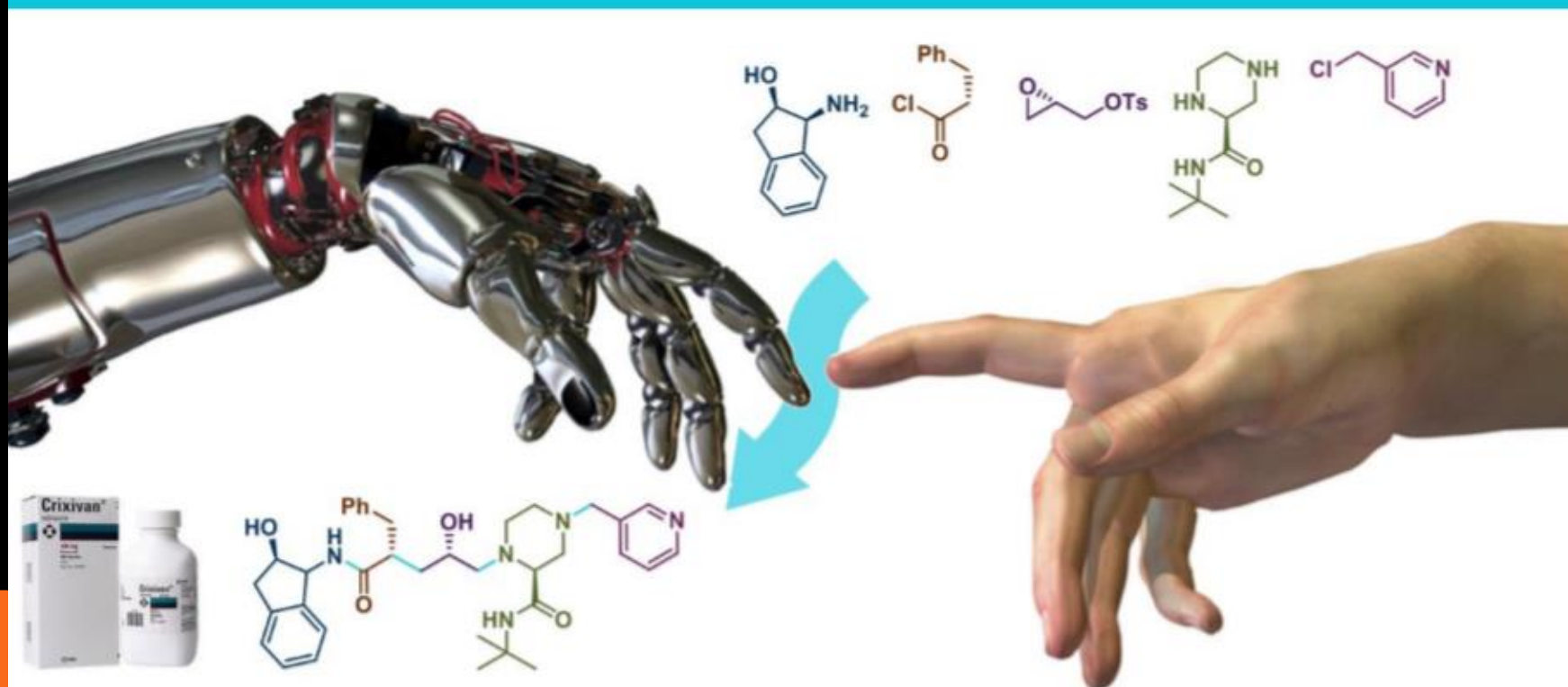
- Creating larger datasets of molecules (>1000)
- Trying out the MoleculeNet benchmarks
- Attempting de novo molecule suggestion etc

Developing the input data method

- Finding out what information we need to include, mass, charge, valence, distance?
 - Testing unscaled icospheres
-



EPSRC Centre for Doctoral Training in Technology Enhanced Chemical Synthesis



Enhancing Chemical Synthesis with Technology