NATIONAL COMPUTATIONAL INFRASTRUCTURE

Machine Learning in Chemistry and Materials Science

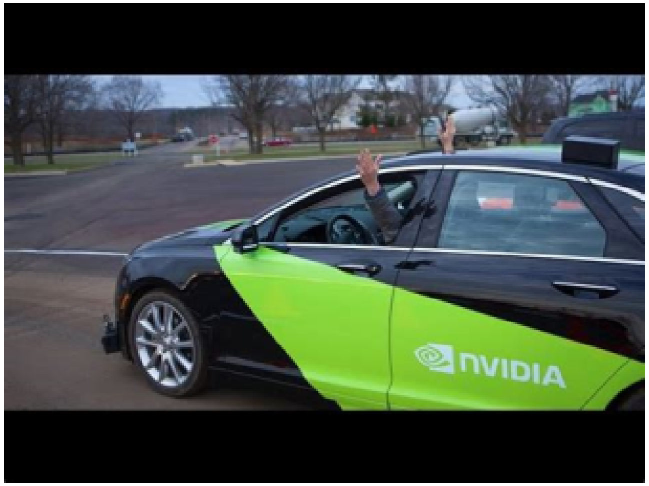
Rika Kobayashi

Quito, July 2019

Artificial Intelligence – Having machines exhibit human intelligence *i.e.* carry out tasks that humans can

Machine Learning – Having machines learn for themselves

Computer vision - image classification, image restoration, object detection in images and videos



Computer vision - image classification, image restoration, object detection in images and videos

Natural language processing - speech recognition, sentiment analysis, speech synthesis, language translation in text and audio

Computer vision - image classification, image restoration, object detection in images and videos

Natural language processing - speech recognition, sentiment analysis, speech synthesis, language translation in text and audio

Computer vision - image classification, image restoration, object detection in images and videos

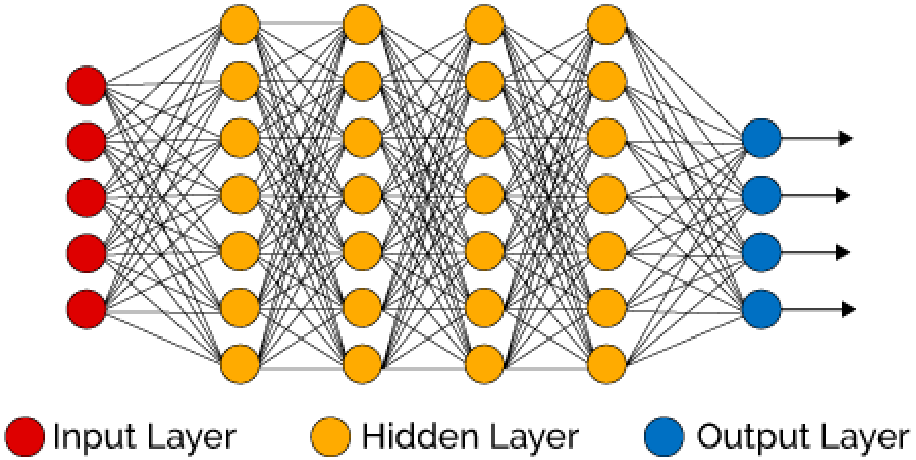
Natural language processing - speech recognition, sentiment analysis, speech synthesis, language translation in text and audio

Data mining - predicting market demand

Artificial Intelligence – Having machines exhibit human intelligence *i.e.* carry out tasks that humans can

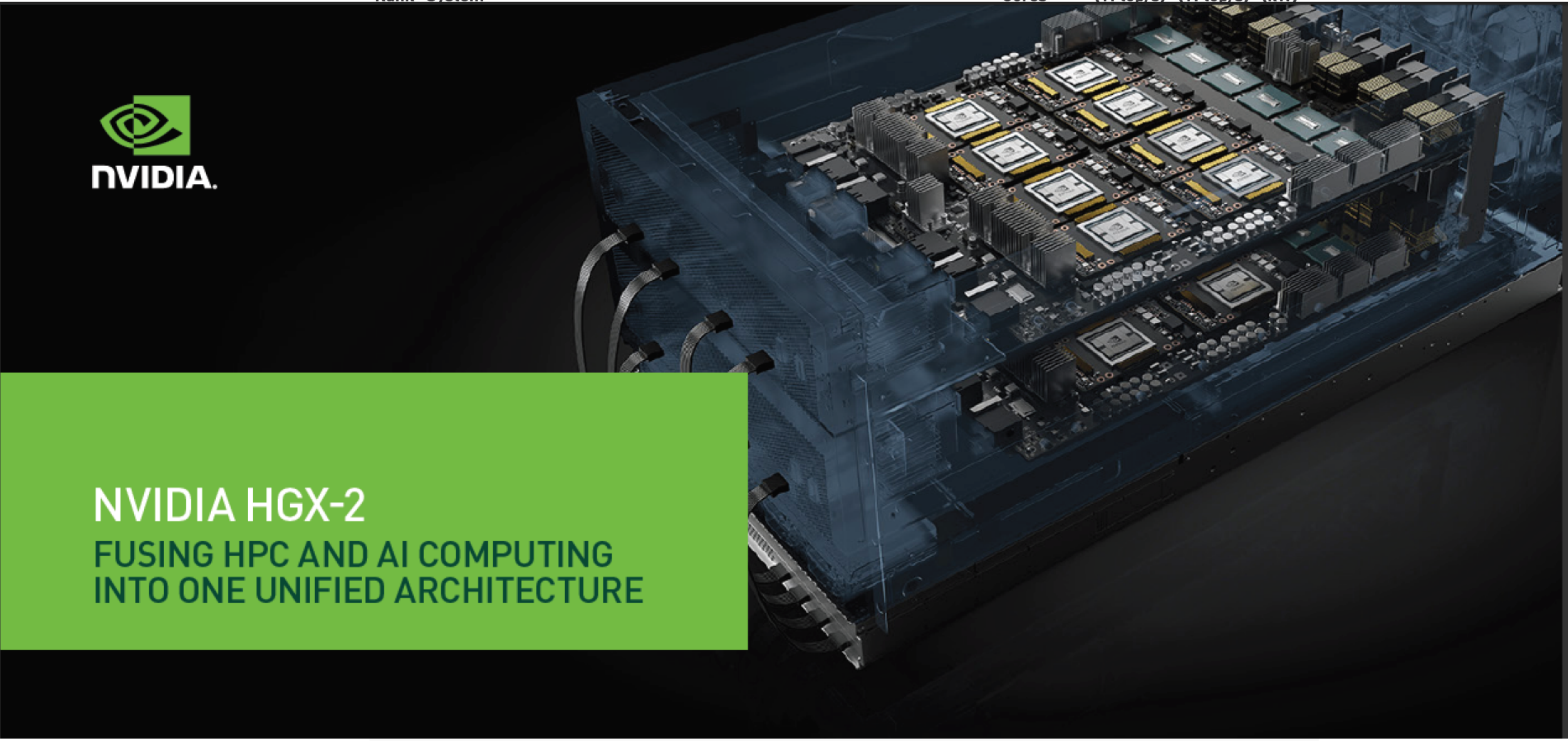
Machine Learning – Having machines learn for themselves

Deep Learning – use of artificial neural networks with multiple layers allowing "deep" connections



Top 500 Supercomputer Sites – June 2019

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)



United States

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
8	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2570 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	32,576.6	1,649

Next generation deep learning

- Medical screening
- Weather forecasting and event detection
- Geographic Information Systems for satellite image analysis
- Bioinformatics

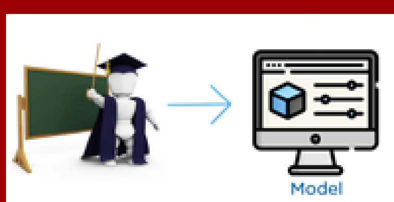
Deep learning in chemistry

- DeepChem - deep-learning in drug discovery, quantum chemistry and biology
- neural network force fields at DFT accuracy
- Kohn-Sham density from machine learning

Types of Machine Learning

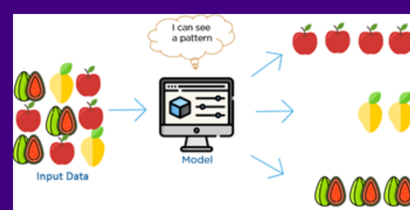
Supervised Learning

- Labeled data
- Direct feedback
- Regression
- Classification



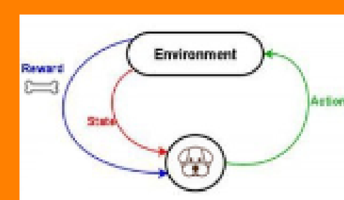
Unsupervised Learning

- Unlabeled data
- No feedback
- Clustering



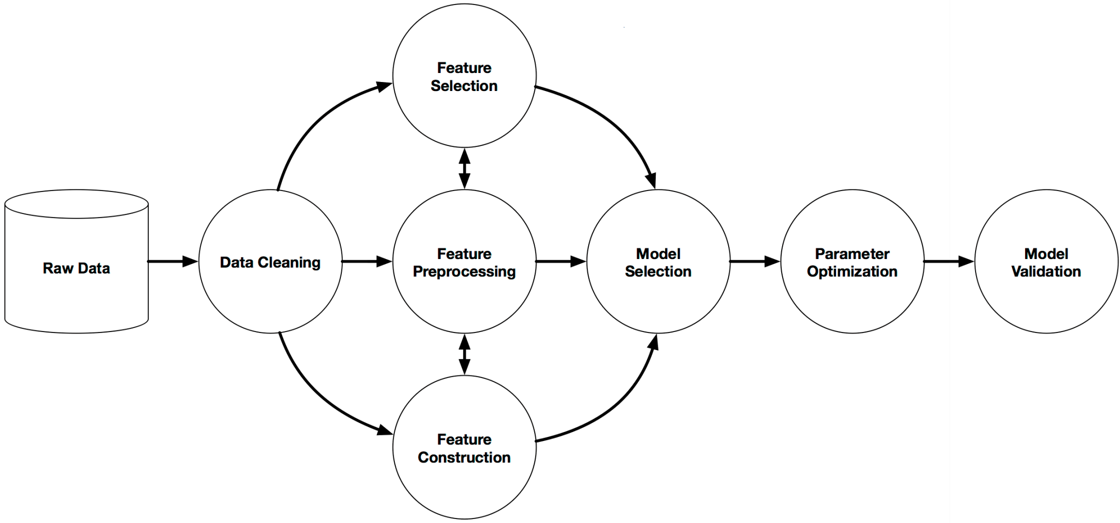
Reinforcement Learning

- Trial and error
- Reward based



Steps in a supervised machine learning workflow

1. Load the data
2. Explore the data
3. Preprocess the data
4. Run model
5. Evaluate model
6. Refine model
7. Predict



Deep Learning Frameworks are building blocks for the design, training and validation of deep neural networks through a high -level programming interface

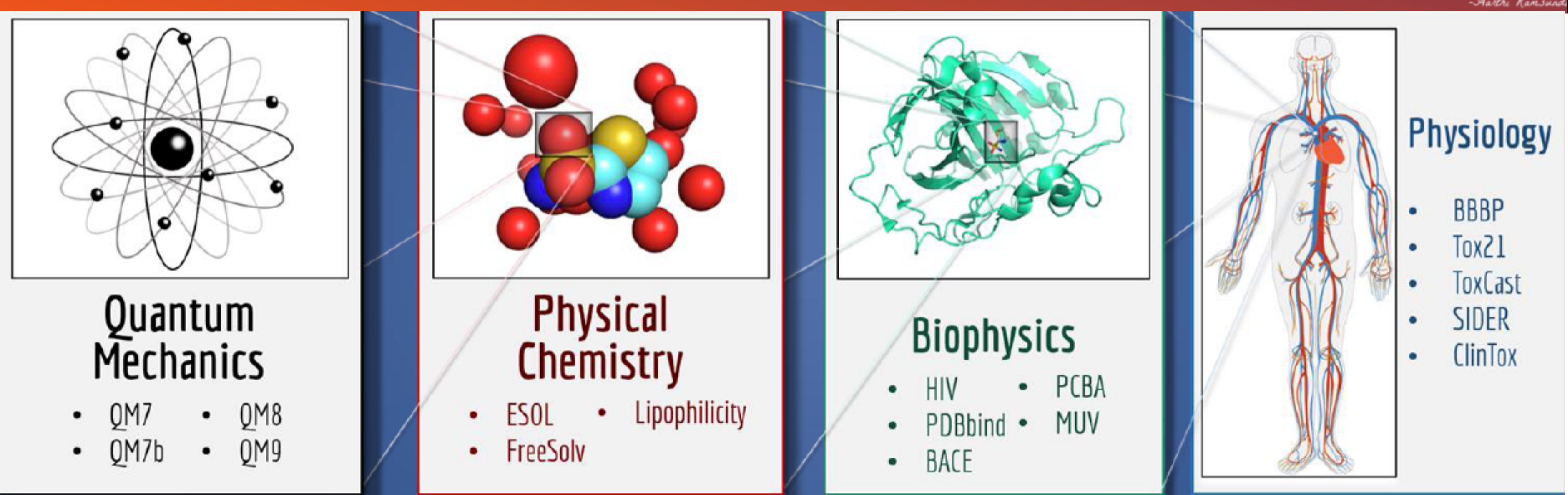
- TensorFlow
- Torch/PyTorch
- Caffe/Caffe2
- Microsoft Cognitive Toolkit/CNTK
- MXNet
- Scikit

deepchem

About Blog Tutorials Discuss Docs

Fork me on GitHub

DeepChem is a Python library democratizing deep learning for science.



Sam Hutchinson (3rd year student)

Springer Link

Journal of Computer-Aided Molecular Design
July 2014, Volume 28, Issue 7, pp 711–720 | Cite as

FreeSolv: a database of experimental and calculated hydration free energies, with input files

Authors | Authors and affiliations

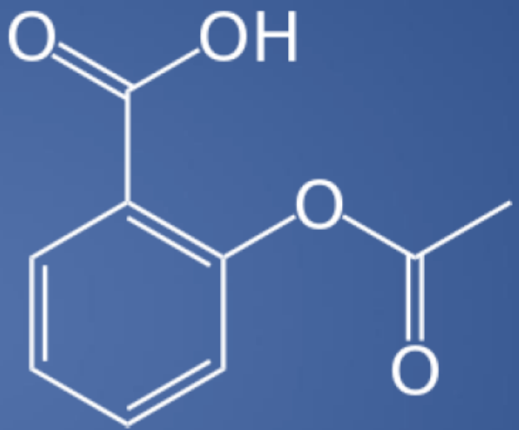
David L. Mobley, J. Peter Guthrie

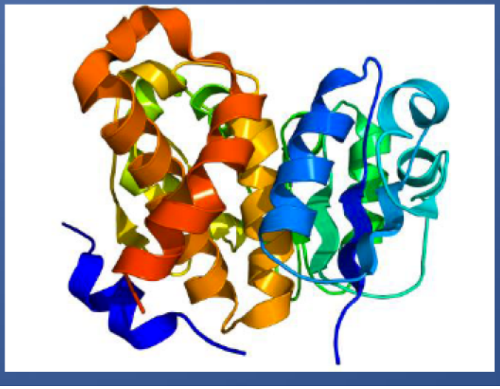
643 experimental and calculated hydration free energy of small molecules in water

Sam Hutchinson (3rd year student) – Feature Engineer

Molecular ML Challenge: Featurization

- Molecules come in many sizes and shapes.
- How can a molecule be transformed into a vector/matrix for machine learning?
- Turns out different representations needed for different problems.





Machine Learning and AI via Brain simulations

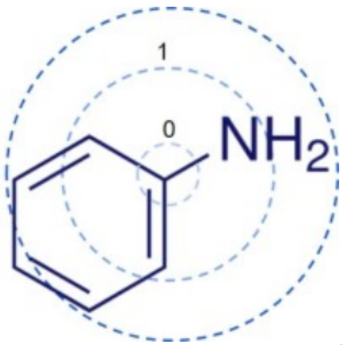
Andrew Ng

Stanford University

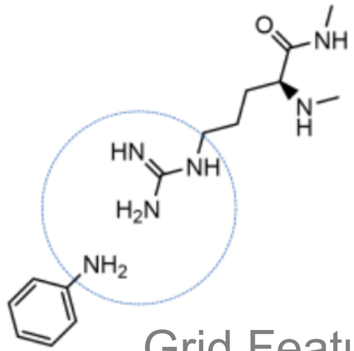
Coming up with features is difficult, time-consuming, requires expert knowledge.

"Applied machine learning" is basically feature engineering.

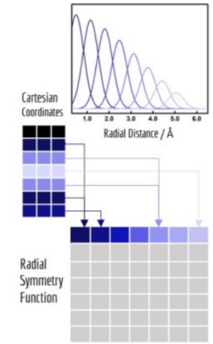
Sam Hutchinson (3rd year student) – Feature Engineer



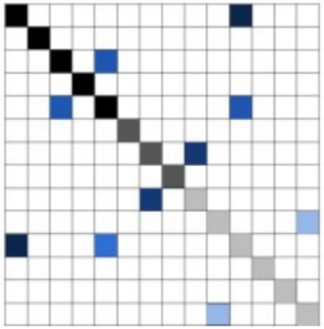
ECFP



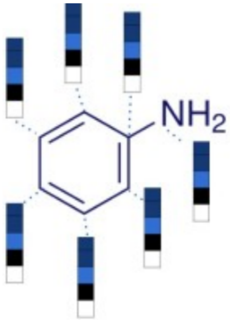
Grid Featurizer



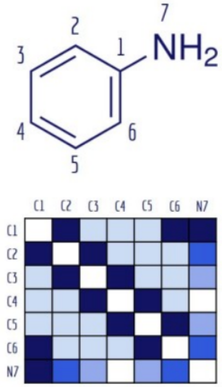
Symmetry Functions



Coulomb Matrix



Graph Convolutions



Weave

NCI

Sam Hutchinson (3rd year student) – Feature Engineer

ECFP vs **FCFP**

Extended **C**onnectivity **F**inger**p**rints

Based on **intra**molecular descriptors

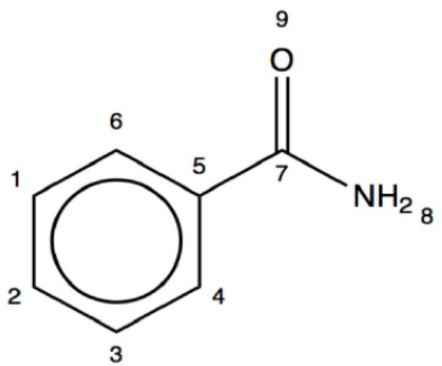
- atomic mass
- atomic number
- atomic charge
- valence minus number of hydrogens
- no. of directly attached heavy neighbours
- no. of directly attached hydrogens
- is it in a ring?

Functional **C**onnectivity **F**inger**p**rints

Based on **inter**molecular descriptors

- hydrogen bonding donor
- hydrogen bonding acceptor
- acidic
- basic
- aromatic
- halogenic

NCI



Atom	Acidic	Aromatic	Halogen	Basic	H-bond acceptor	H-bond donor	6-bit code/ Identifier
1	F	**T**	F	F	F	F	010000
2	F	**T**	F	F	F	F	010000
3	F	**T**	F	F	F	F	010000
4	F	**T**	F	F	F	F	010000
5	F	**T**	F	F	F	F	010000
6	F	**T**	F	F	F	F	010000
7	F	F	F	F	F	F	000000
8	F	F	F	**T**	F	F	000100
9	F	F	F	F	F	**T**	000001

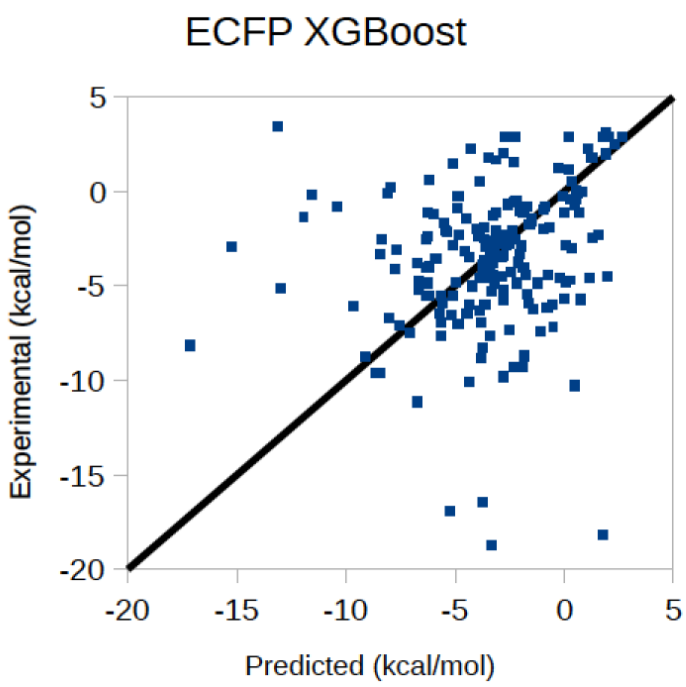
[1, 00000, 1, 000100, 1, 010000, 2, 000001]

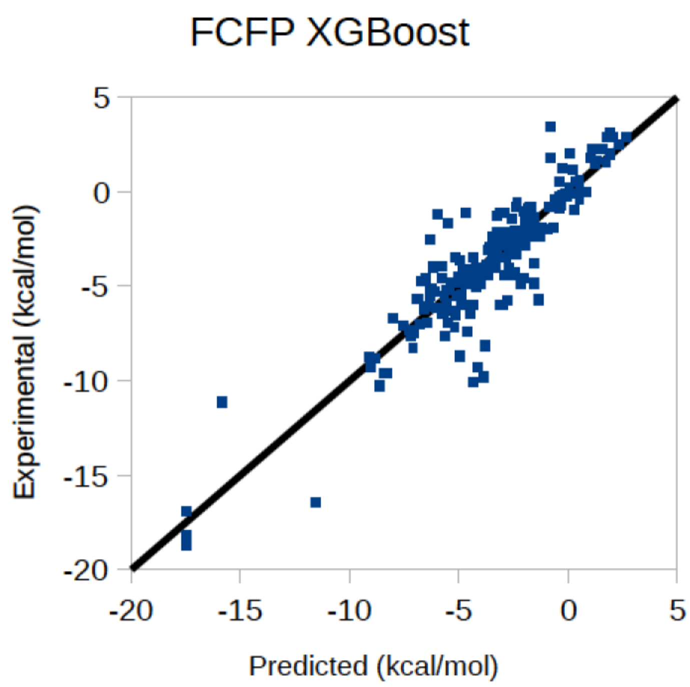
[iteration_no, central_atom, bond_order, atom_8, bond_order, atom_5, bond_order, atom_9]

crc32 hash function

1232502512

new fragment identifier





	EFCP	FCFP (this work)	FCFP (RdKit)
Train (R^2)	0.97±0.01	0.97±0.01	0.97±0.01
Valid (R^2)	0.74±0.13	0.78±0.08	0.82±0.04
Test (R^2)	0.78±0.05	0.81±0.03	0.83±0.04
Test (RMS) in kcal/mol	1.78±0.27	1.65±0.21	1.60±0.14

	EFCP	FCFP (this work)	FCFP (RdKit)
Train (R^2)	0.97±0.01	0.97±0.01	0.97±0.01
Valid (R^2)	0.74±0.13	0.78±0.08	0.82±0.04
Test (R^2)	0.78±0.05	0.81±0.03	0.83±0.04
Test (RMS) in kcal/mol	1.78±0.27	1.65±0.21	1.60±0.14

Model	Training	Validation	Test
Random Forest	**0.80±0.03**	**2.12±0.68**	**2.03±0.22**
Multitask	**1.07±0.06**	**1.95±0.41**	**1.87±0.07**
XGBoost	**0.85±0.12**	**1.76±0.21**	**1.74±0.15**
KRR	**0.31±0.03**	**2.10±0.12**	**2.11±0.07**
GraphConv	**0.31±0.09**	**1.35±0.15**	**1.40±0.16**
DAG	**0.49±0.46**	**1.48±0.15**	**1.63±0.18**
Weave	**0.32±0.04**	**1.19±0.08**	**1.22±0.28**
MPNN	**0.31±0.05**	**1.20±0.02**	**1.15±0.12**

MoleculeNet benchmarks arXiv:1703.00564v3

Minnesota Solvation Database – version 2012

If this database is used for published work, the following citation should be given:

Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvation Database – version 2012, University of Minnesota, Minneapolis, 2012.

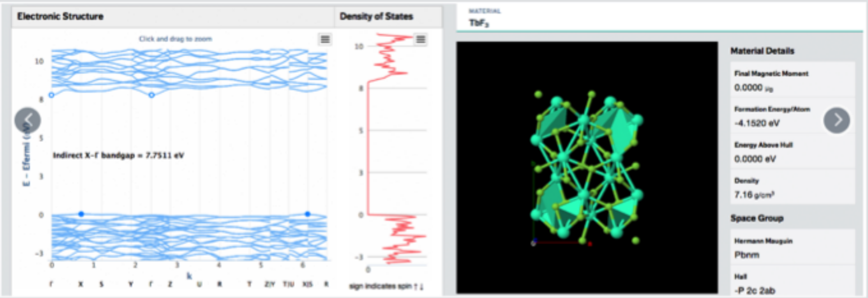
	EFCP	**FCFP**	**GraphConv**	**DAG**	**weave**
Octanol (ε=9.86)					
Train (R^2)	0.97±0.06	0.96 ±0.02	1.00±0.00	1.00±0.00	1.00±0.00
Valid (R^2)	0.80±0.08	0.89±0.06	0.85±0.05	0.96±0.02	0.90±0.07
Test (R^2)	0.79±0.10	0.79±0.13	0.94±0.01	0.96±0.02	0.94±0.03
Test (RMS) in kcal/mol	1.65± 0.52	1.54±0.24	1.10±0.42	0.76±0.29	1.04±0.30
Hexadecane (ε=2.05)					
Train (R^2)	0.89±0.14	0.84±0.20	0.97±0.01	1.00±0.01	0.99±0.01
Valid (R^2)	0.62±0.22	0.52±0.22	0.66±0.08	0.91±0.04	0.90±0.08
Test (R^2)	0.59±0.11	0.79±0.34	0.59±0.10	0.68±0.06	0.69±0.14
Test (RMS) in kcal/mol	1.22±0.51	1.19±0.66	1.13±0.14	1.05±0.09	1.03±0.24

Machine Learning in Materials Science

The Materials Project

Harnessing the power of supercomputing and state of the art electronic structure methods, the Materials Project provides open web-based access to computed information on known and predicted materials as well as powerful analysis tools to inspire and design novel materials.

Learn more | Tutorials | Sign In or Register to start using



EXPLORE MATERIALS
Search for materials information by chemistry, composition, or property

EXPLORE BATTERIES
Find candidate materials for lithium batteries. Get voltage profiles and oxygen evolution data.

VISUALIZE STABILITY
Generate phase and pourbaix diagrams to find stable phases and study reaction pathways

INVENT STRUCTURES
Design new compounds with our structure editor and substitution algorithms

CALCULATE
Calculate the enthalpy of 10,000+ reactions and compare with experimental values

Database Statistics

83,989	52,179	21,954	530,243
INORGANIC COMPOUNDS	BANDSTRUCTURES	MOLECULES	NANOPOROUS MATERIALS
7,676	1,002	3,628	16,128
ELASTIC TENSORS	PIEZOELECTRIC TENSORS	INTERCALATION ELECTRODES	CONVERSION ELECTRODES

Machine Learning and AI via Brain simulations

Andrew Ng

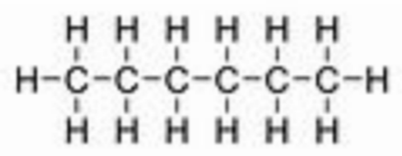
Stanford University

Coming up with features is difficult, time-consuming, requires expert knowledge.

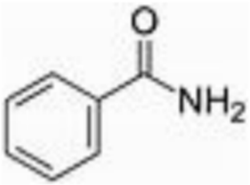
"Applied machine learning" is basically feature engineering.

Machine Learning in Materials Science?

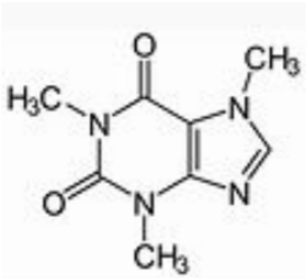
SMILES - **s**implified **m**olecular-**i**nput **l**ine-**e**ntry **s**ystem



CCCCCC



C1=CC=C(C=C1)C(=O)N



CN1C=NC2=C1C(=O)N(C(=O)N2C)C

Features used in Materials Science

number of atoms/ions

atomic number

atomic mass

atomic/ionic radius

period/group in Periodic Table

valency

electron affinity

electronegativity

ionization energy

van der Waals radius

covalent radius

melting point

boiling point

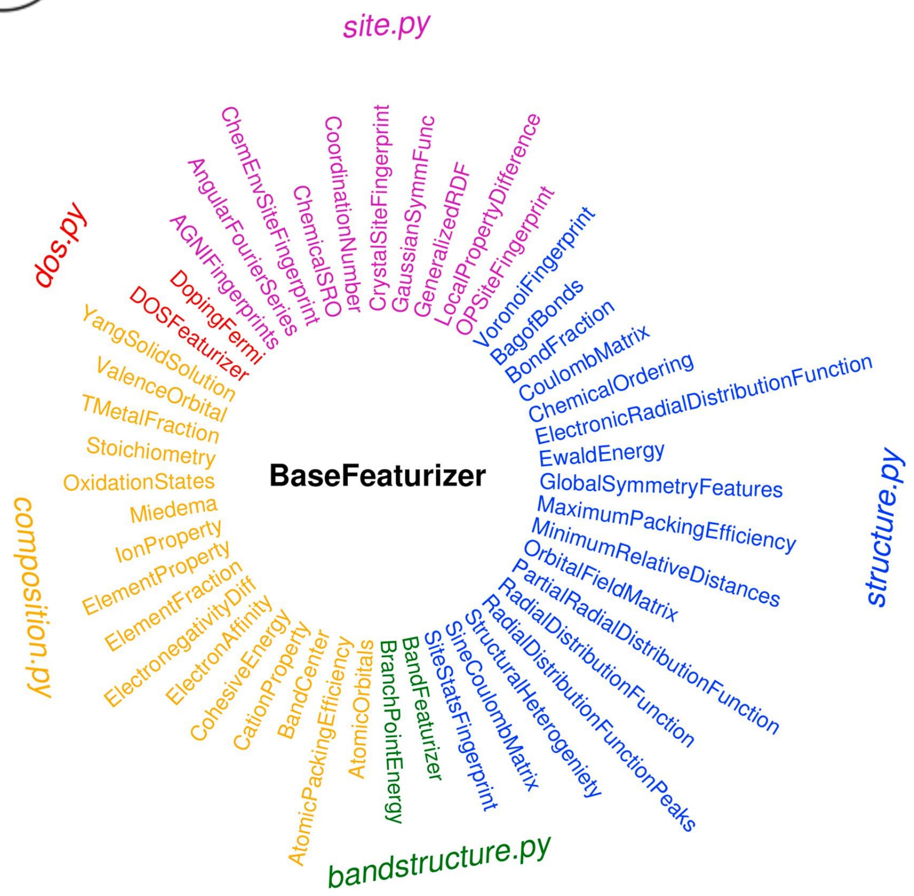
density

molar volume

thermal conductivity

specific heat

diffusivity


site.py
ChemEnvSiteFingerprint
CoordinationNumber
CrystalSiteFingerprint
GaussianSymmFunc
GeneralizedRDF
LocalPropertyDifference
AngularFourierSeries
ChemicalSRO
AGNIFingerprints
OPSiteFingerprint
VoronoiFingerprint
dos.py
DopingFermi
DOSFeaturizer
BagofBonds
BondFraction
CoulombMatrix
ChemicalOrdering
ElectronicRadialDistributionFunction
EwaldEnergy
GlobalSymmetryFeatures
MaximumPackingEfficiency
MinimumRelativeDistances
OrbitalFieldMatrix
PartialRadialDistributionFunction
RadialDistributionFunction
RadialDistributionFunctionPeaks
StructuralHeterogeniety
SineCoulombMatrix
SiteStatsFingerprint
structure.py
YangSolidSolution
ValenceOrbital
TMetalFraction
Stoichiometry
OxidationStates
Miedema
IonProperty
ElementProperty
ElementFraction
ElectronegativityDiff
ElectronAffinity
CohesiveEnergy
CationProperty
BandCenter
AtomicPackingEfficiency
AtomicOrbitals
composition.py
BaseFeaturizer
BranchPointEnergy
BandFeaturizer
bandstructure.py

NCI

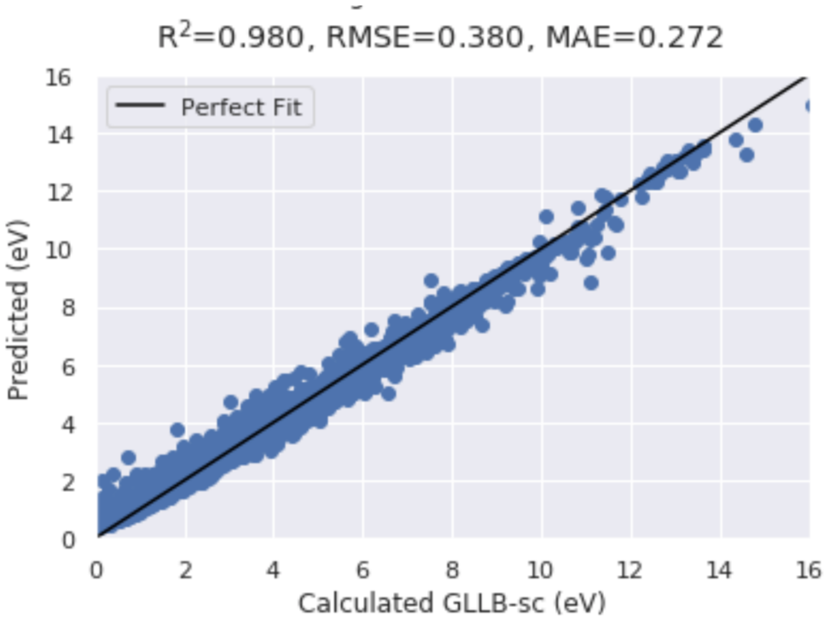
www.MaterialsViews.com

www.advenergymat.de

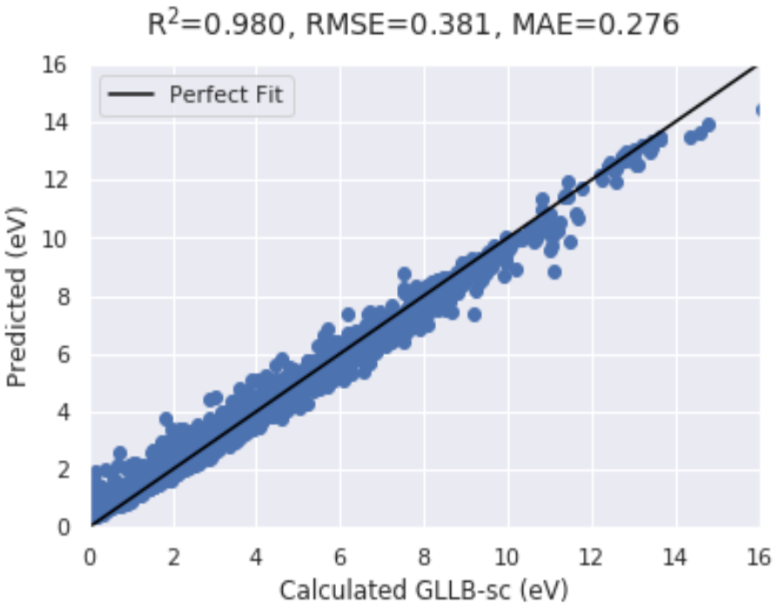
New Light-Harvesting Materials Using Accurate and Efficient Bandgap Calculations

Ivano E. Castelli, Falco Hüser, Mohnish Pandey, Hong Li, Kristian S. Thygesen, Brian Seger, Anubhav Jain, Kristin A. Persson, Gerbrand Ceder, and Karsten W. Jacobsen*

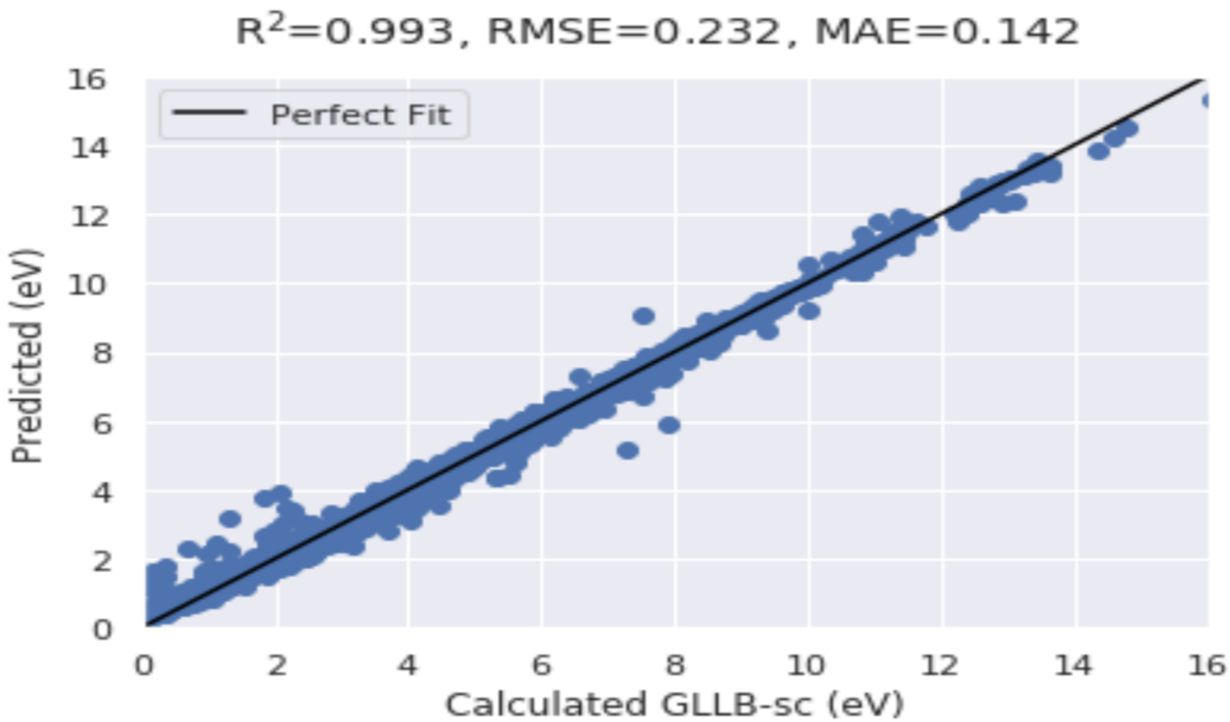
This contains GLLB-sc computed band gaps of around 2400 experimentally known materials showing a band gap at the GGA level and their corresponding Materials Project identifier which was used to download 2254 structures from the Materials Project repository.



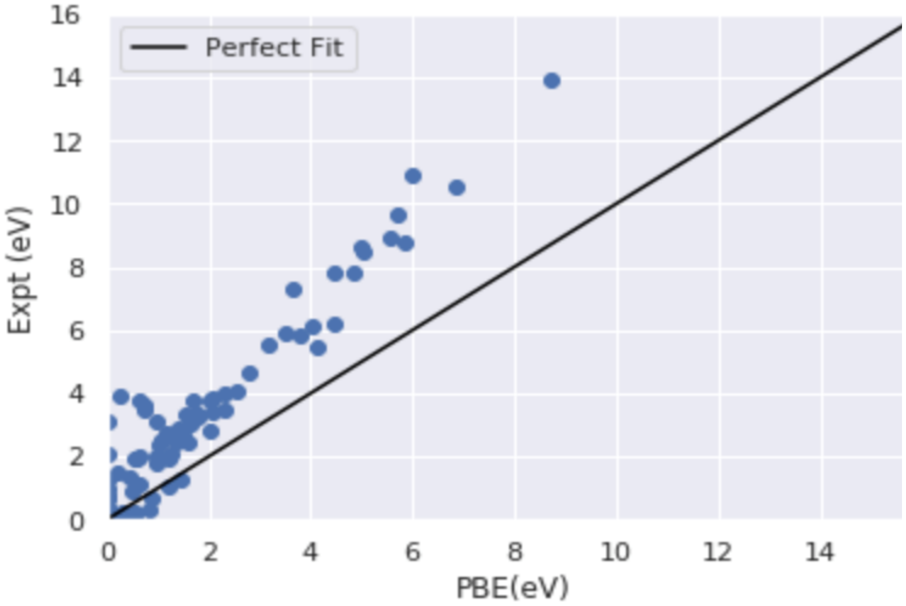
Elemental descriptors



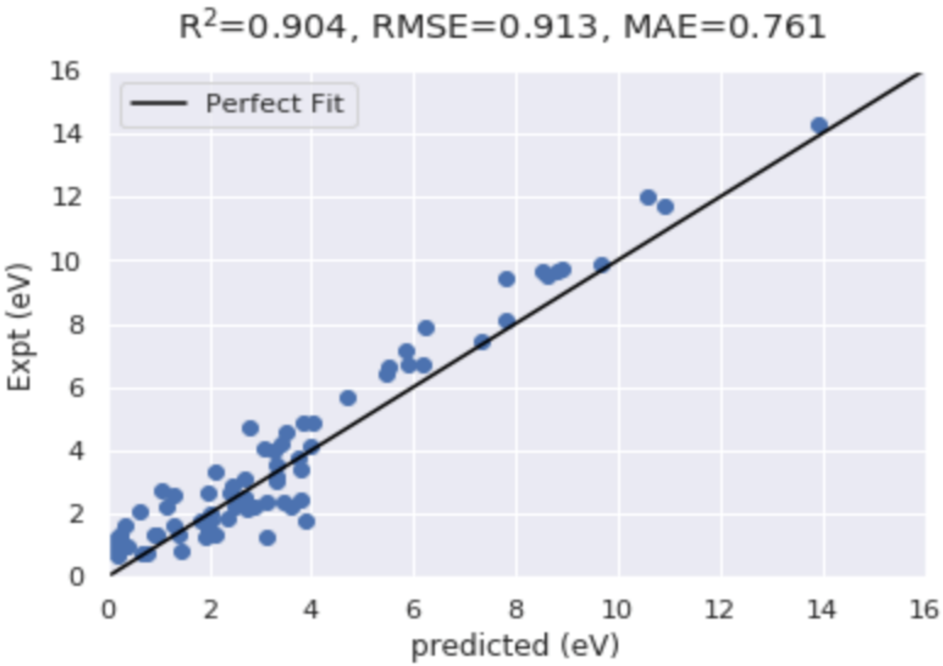
Including structural features



Including PBE estimate



Experimental vs PBE band gaps



Experimental vs ML band gaps

Conclusions

Conclusions from Solvation study

– Putting in more "chemistry" did not make much of a difference

– Featurisers without as much "chemistry" perform better

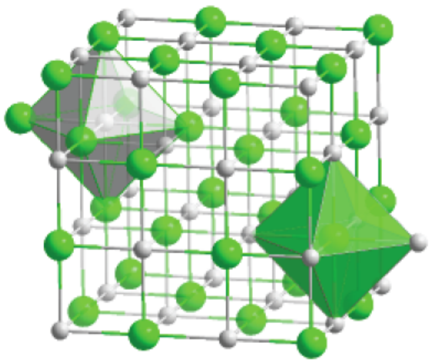
– Are we putting in the right "chemistry"?

Conclusions from Materials study

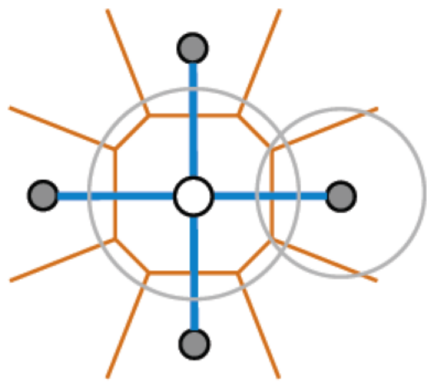
– Putting in more "structure" did not make much of a difference

– Better performance was gained by including crude *ab initio* descriptors

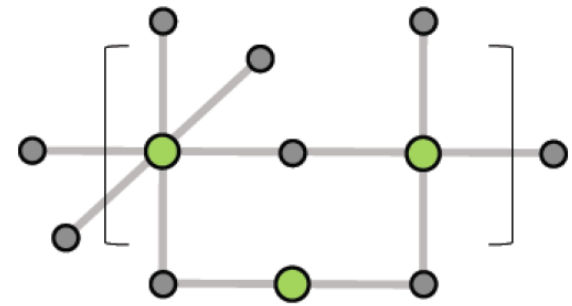
a) crystal structure



b) Voronoi tessellation and neighbors search



c) infinite periodic graph construction and property labeling

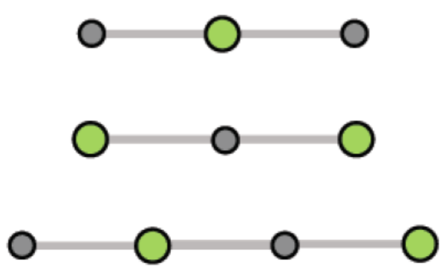


d) decomposition to fragments

nodes (atoms)

edges (bonds)

path fragments of length *l*, *l* = 2, 3, ...



circular fragments (polyhedrons)