From Structural Biology to AI: a Holistic Approach to Studying Molecular Machines

Michelle Gill, PhD
NVIDIA Professional Services

NVIDIA

NVIDIA Professional Services

Our goal is to *enable* broader customer *adoption* of *Deep Learning* on *NVIDIA-accelerated* platforms

Enable

NVIDIA helps organizations get started or overcome roadblocks

Adoption

NVIDIA equips teams with the skills to plan, manage, and deliver projects going forward

Deep Learning

NVIDIA Professional Services is focused on Deep Learning

NVIDIA-Accelerated

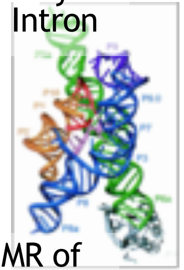
NVIDIA supports customers adopting NVIDIA-accelerated infrastructure (on-premises, in the cloud, or at the edge)

From Scientist to Data Scientist

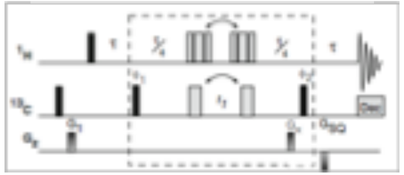
(and Sometimes Both)

Solution NMR Applications & Methods Development

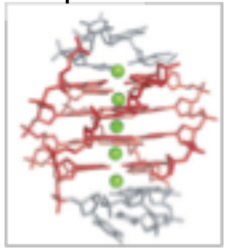
Biochemistry of Group I Intron



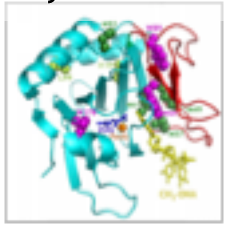
NMR Pulse Sequence Development



^{205}Tl NMR of G-Quadruplex

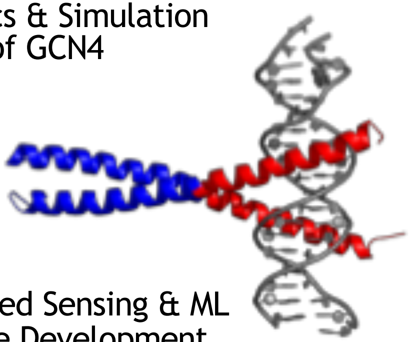


Dynamics of AlkB DNA Methyltransferase



Compressed Sensing & Molecular Dynamics

Dynamics & Simulation of GCN4



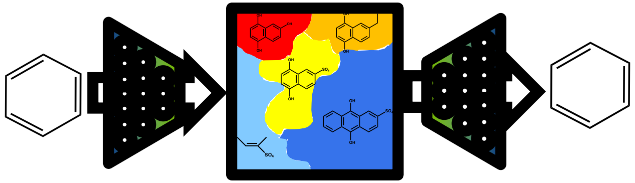
Compressed Sensing & ML Software Development

NESTA-NMR

fast and accurate reconstruction of NUS data

Data Science & Deep Learning

Deep Learning in Materials Science & Pharma



DeepChem Committer (Early Days)

deepchem

Graduate & Postdoctoral Research → Staff Scientist at NCI → Data Scientist

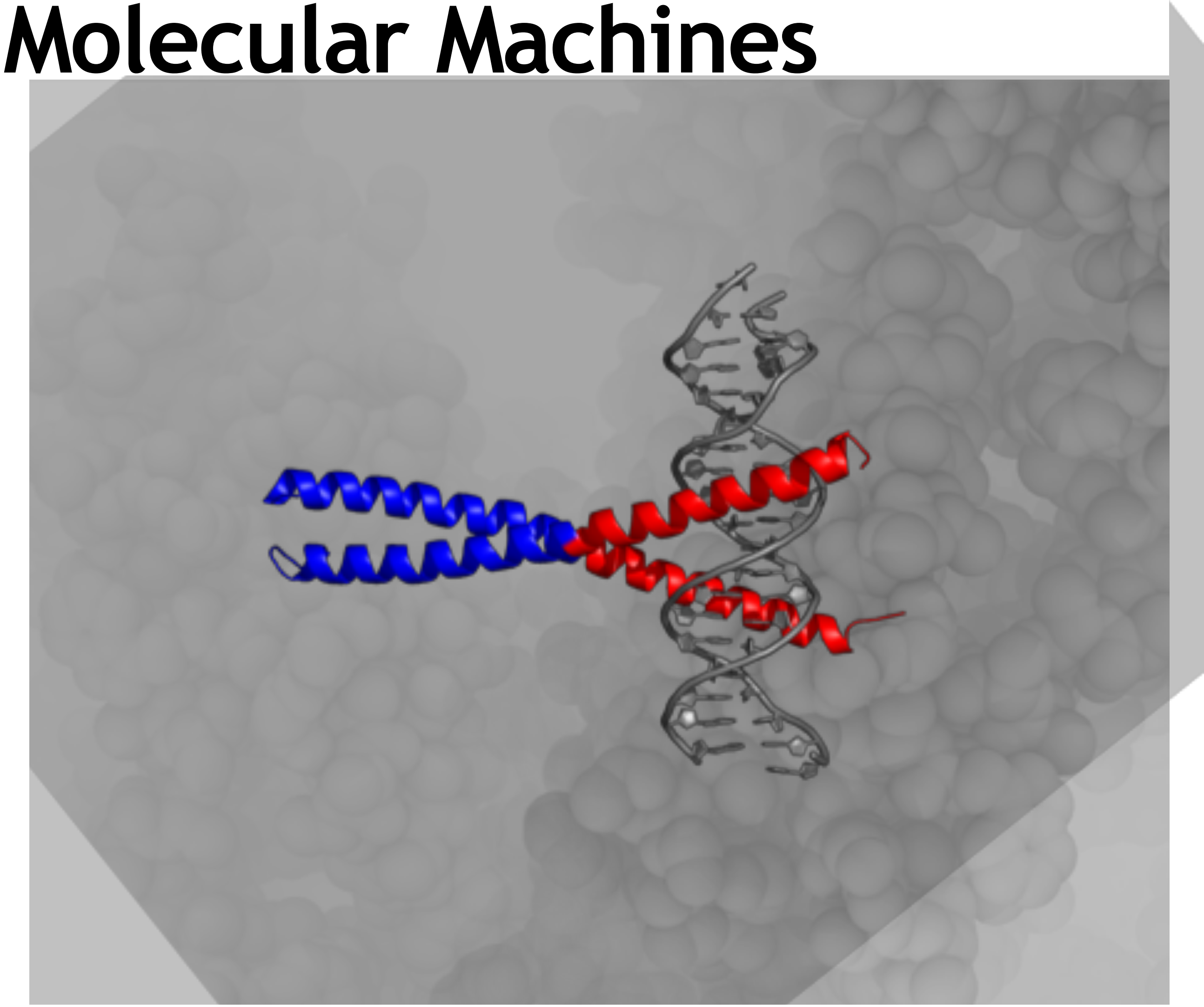
Enzymes are Molecular Machines

Enzymes are nature's machines

Found throughout an organism

Perform the chemical reactions that make life possible

Malfunction of enzymes can contribute to disease states

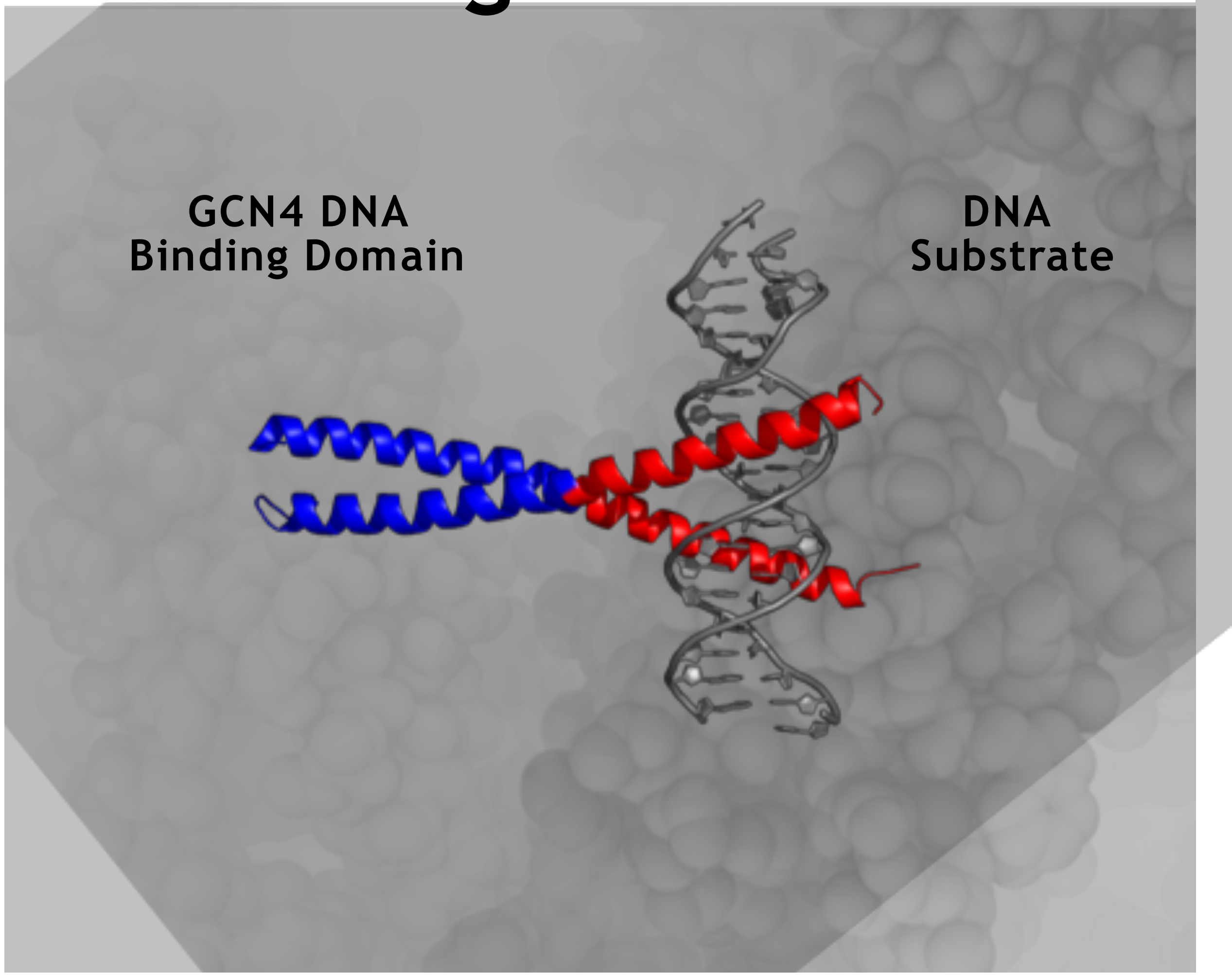


Molecular Dynamics
Control Substrate Recognition

GCN4 is an enzyme that belongs to the transcription factor family

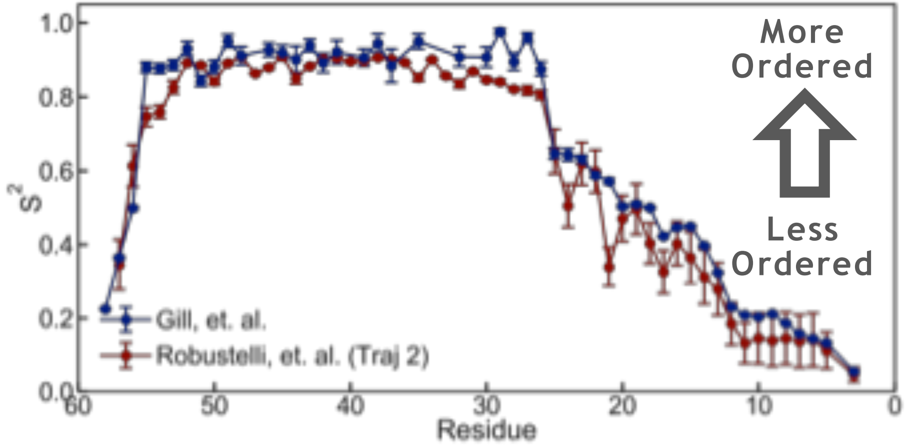
Binds to substrate (DNA) leading to expression of the nearby gene

Binding event requires motion of the enzyme — dynamics



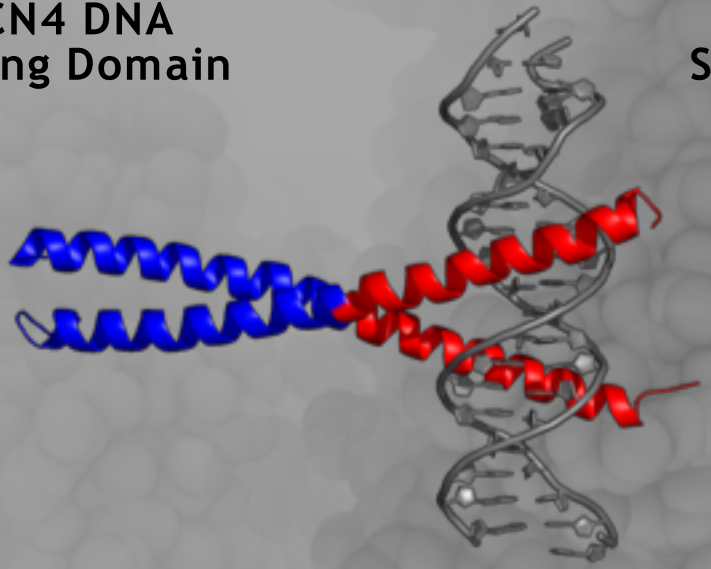
Molecular Dynamics Control Substrate Recognition

Order Parameters (S^2) from Solution NMR



GCN4 DNA Binding Domain

DNA Substrate



Gill, M.L., Byrd, R.A., Palmer, A.G. “Dynamics of GCN4 facilitate DNA interaction: a model-free analysis of an intrinsically disordered region“, *Phys. Chem. Chem. Phys.*, 2016, 18, 5839-5849.

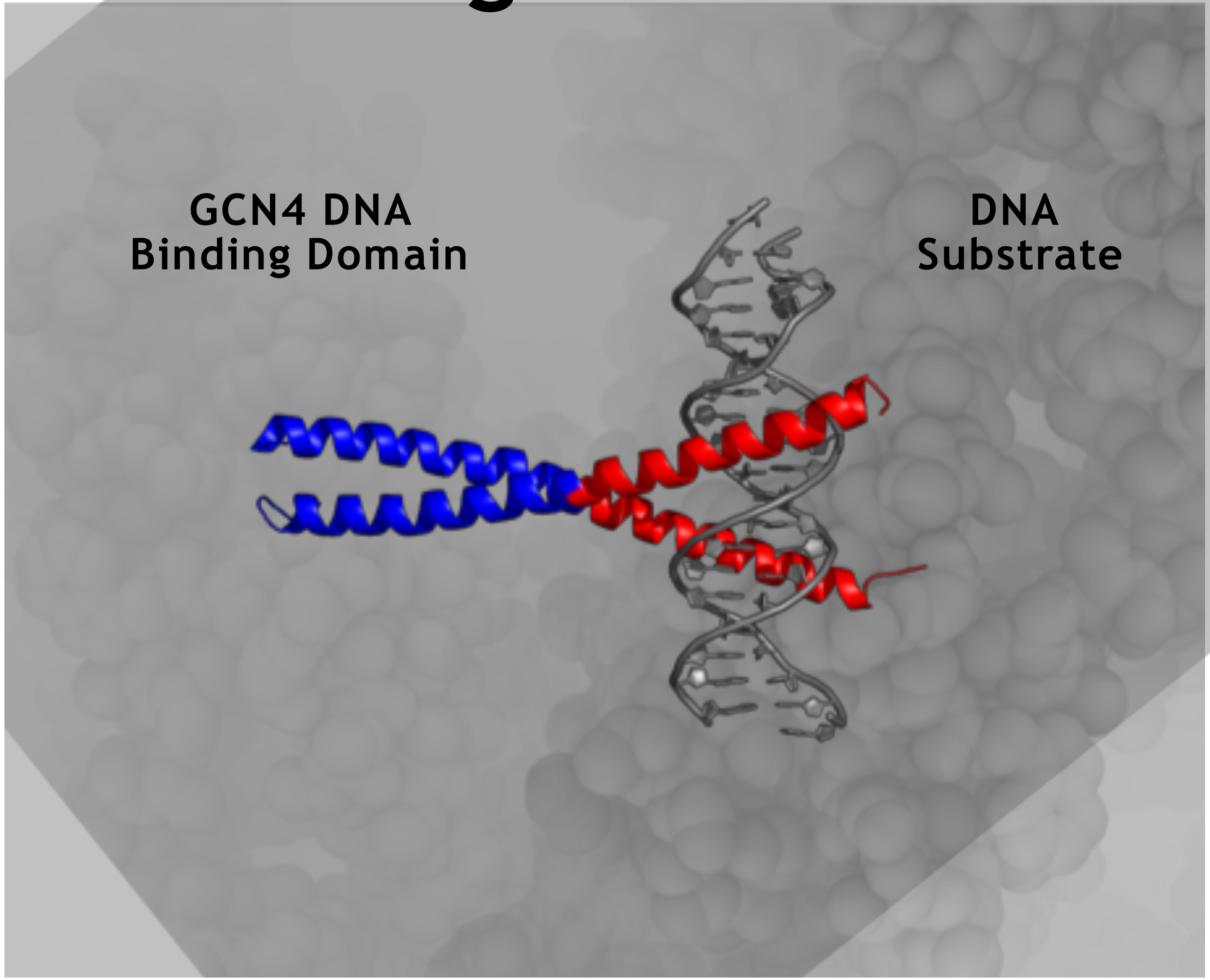
Robustelli, P., Trbovic, N. Friesner, R.A., Palmer, A.G. "Conformational dynamics of the partially disordered yeast transcription factor GCN4", *J. Chem Theory Comput.*, 2013, 9, 5190-5200.

Molecular Dynamics Control Substrate Recognition

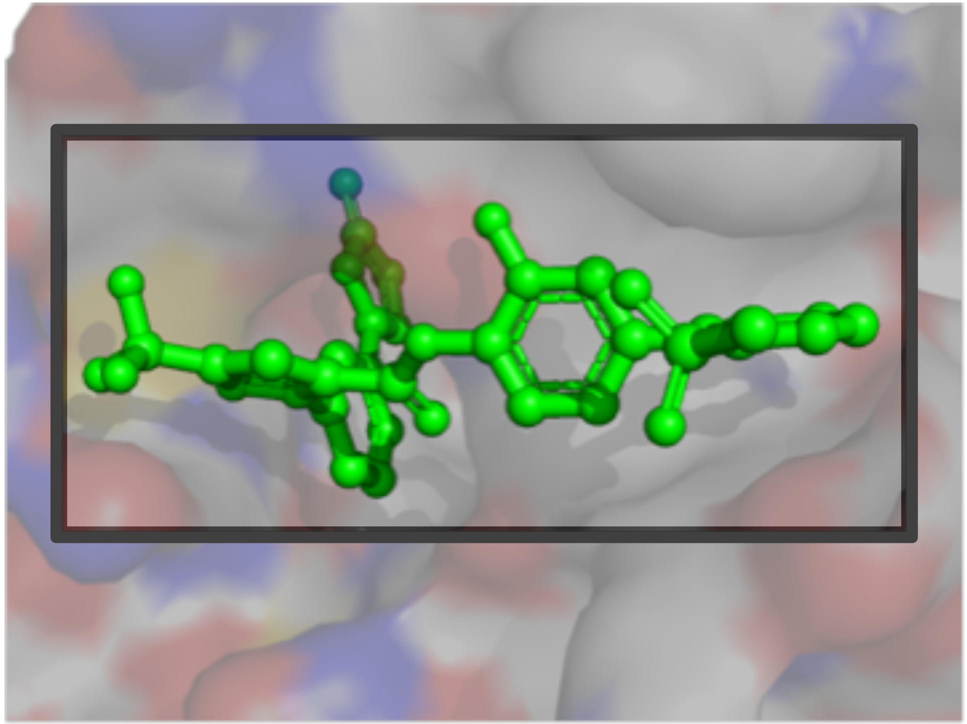
Over 30% of mammalian genome predicted to have regions of intrinsic disorder

Genes containing disordered regions associated with fundamental cellular processes (and cancer)

Disorder difficult to study with traditional structural biology methods



Enzyme Structure and Dynamics are Key for Therapeutic Discovery



Many successful therapeutics (drugs) bind in place of substrates

Understanding enzyme dynamics critical for substrate binding and development of therapeutics

Deep learning excels at recognition of complex patterns

Potential for AI-assisted drug discovery

Challenges Unique to Deep Learning With Chemistry

Limited availability of scientific data

Exploration of large chemical space

Representing chemical features

Limited Availability of Scientific Data

General Topics

Text, image, and sound data are readily available on internet

Limited Availability of Scientific Data

General Topics

Text, image, and sound data are readily available on internet

Chemistry Specific

Access to scientific data limited – difficult to acquire, not shared publicly

Exploration of Large Chemical Space

General Topics

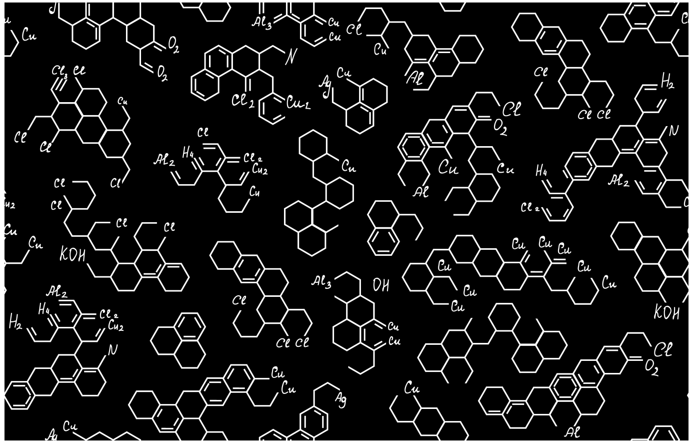
ImageNet classification uses 1000 categories (10^3 magnitude)

Exploration of Large Chemical Space

General Topics

ImageNet classification uses 1000 categories (10^3 magnitude)

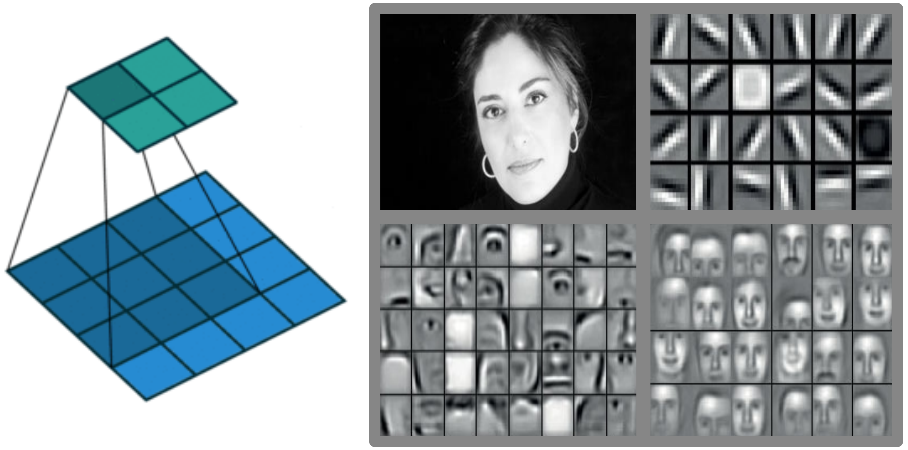
Chemistry Specific



Molecule space encompasses 10^{60} – 10^{300} possibilities (10^8 synthesized)

Representing Chemical Features

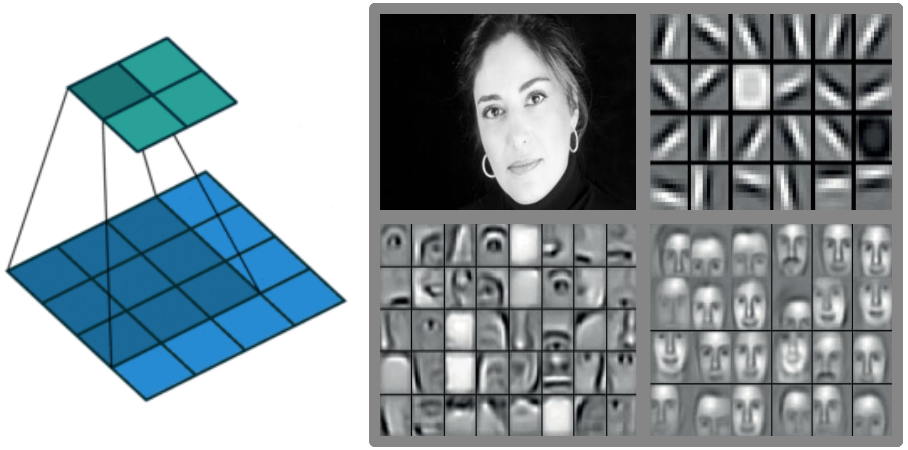
General Topics



Two-dimensional convolutional filters learn image features

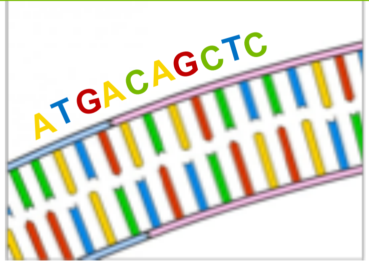
Representing Chemical Features

General Topics



Two-dimensional convolutional filters learn image features

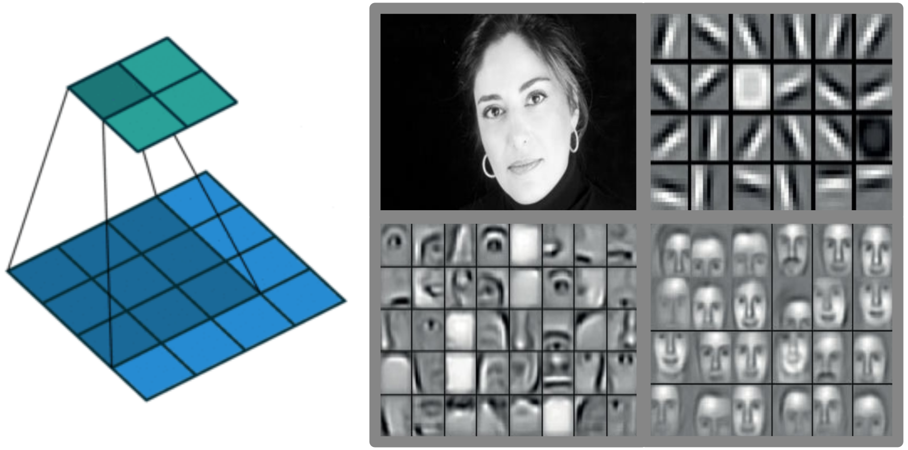
Chemistry Specific



Representations can be one-, two-, and three-dimensional (or combination)

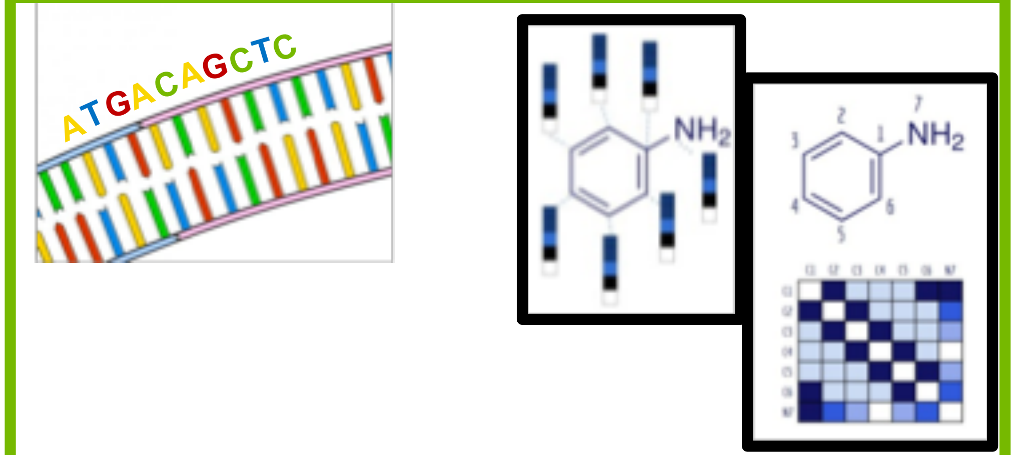
Representing Chemical Features

General Topics



Two-dimensional convolutional filters learn image features

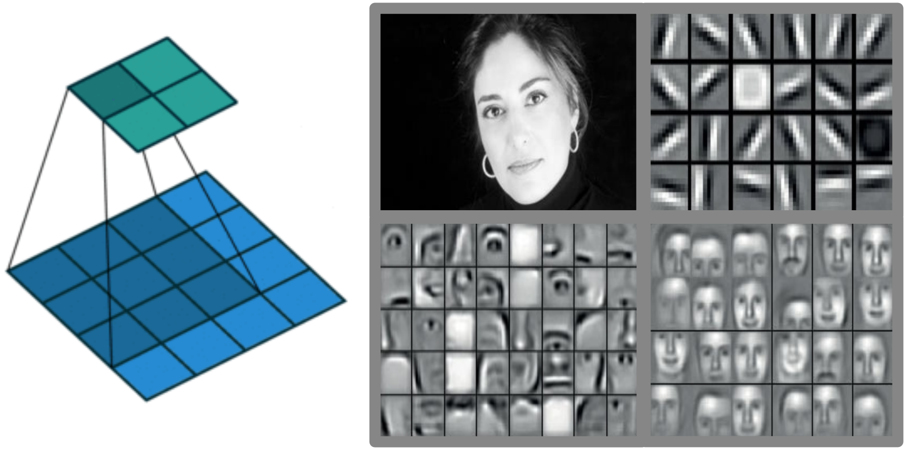
Chemistry Specific



Representations can be one-, two-, and three-dimensional (or combination)

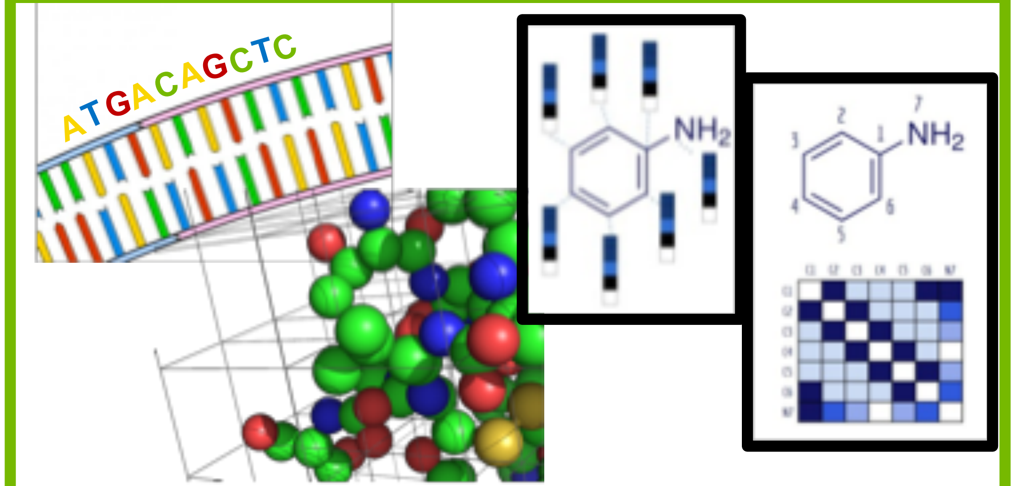
Representing Chemical Features

General Topics



Two-dimensional convolutional filters learn image features

Chemistry Specific



Representations can be one-, two-, and three-dimensional (or combination)

Deep Learning with Chemistry

Predicting enzyme-ligand binding using three-dimensional atomic convolutional networks

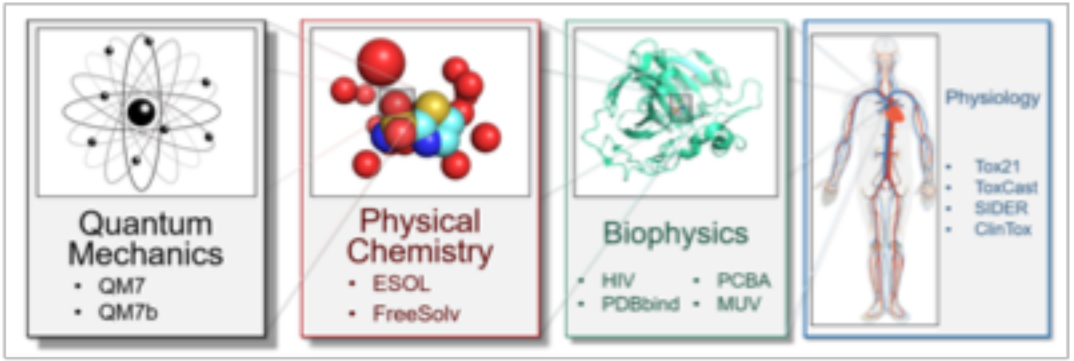
Generation of drug-like molecules using continuous latent spaces

(Bio)Chemistry Data for Deep Learning

Public and proprietary data are a useful combination — for transfer learning and as supplement

Utilized MoleculeNet — benchmark datasets for chemistry, biophysics, and physiology

Part of DeepChem deep learning library

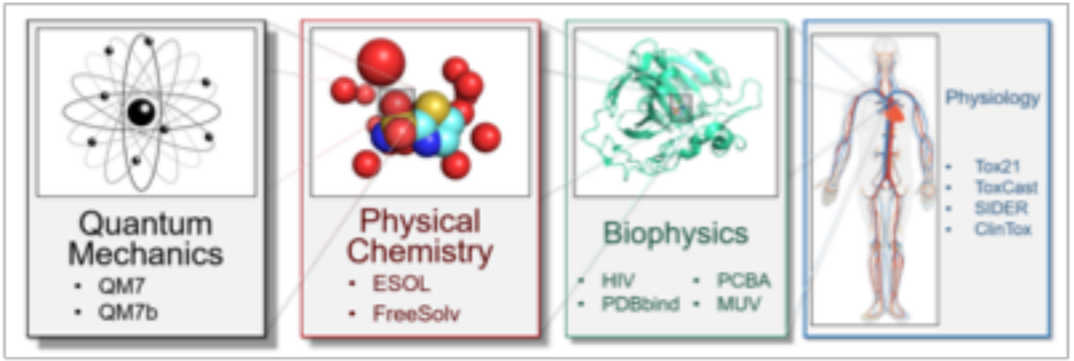


Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., Pande, V. *"MoleculeNet: A Benchmark for Molecular Machine Learning."* arXiv.org, 2017.

(Bio)Chemistry Data for Deep Learning

PDBbind — 12K published measurements and 3D structures of enzyme-ligand complexes

ZINC12 — 35M chemicals and properties for drug discovery



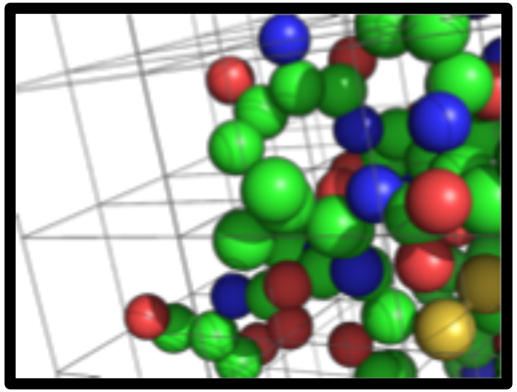
deepchem

Ligand Binding Featurization and Modeling

Utilizes a three-dimensional "atomic fingerprint"

Atom type (element) and distance calculated within threshold and pooled

Atom Type
+
Distance



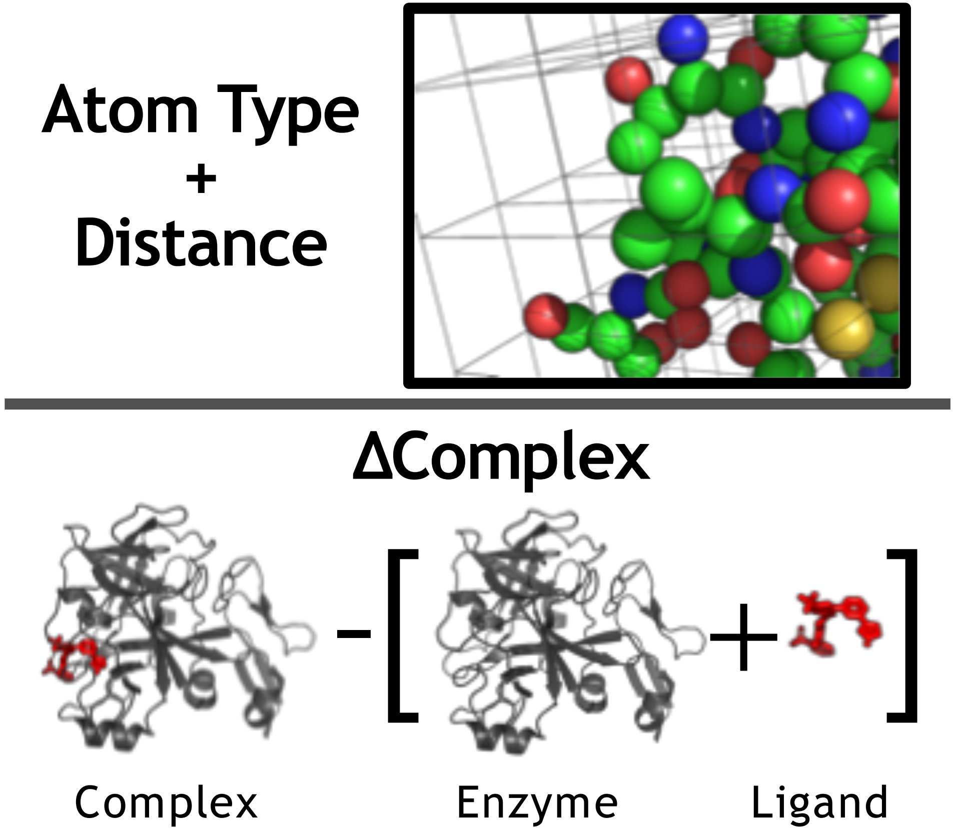
Featurization based on: Gomes, J., Ramsundar, B., Feinberg, E. N., & Pande, V. S. "Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity.", *arXiv,* 2017.

Ligand Binding Featurization and Modeling

Utilizes a three-dimensional "atomic fingerprint"

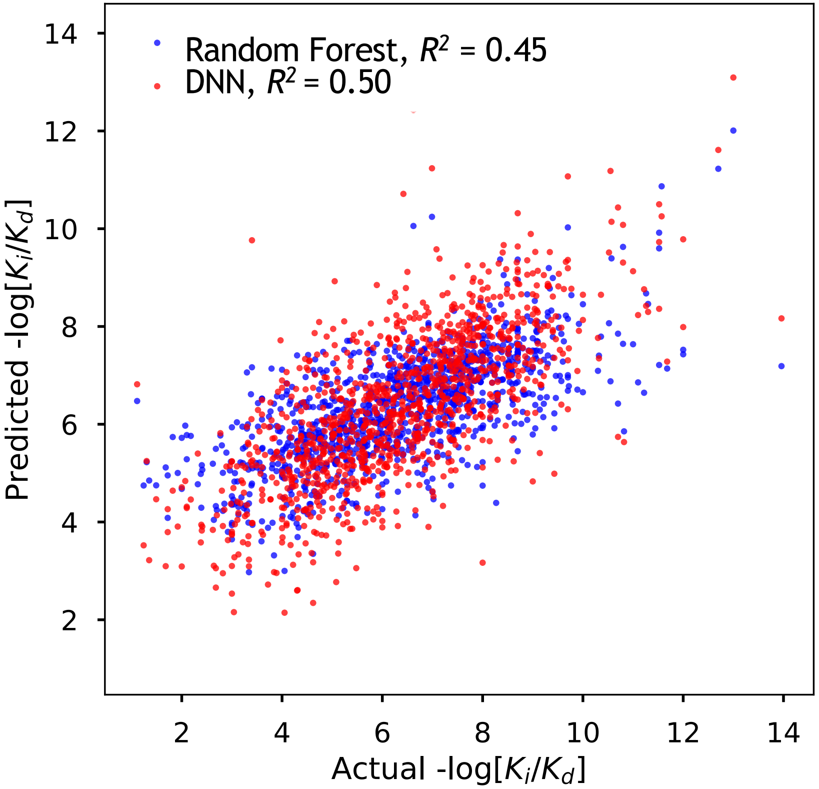
Atom type (element) and distance calculated within threshold and pooled

Three interlinked neural networks emphasize changes introduced by ligand binding


Atom Type + Distance

ΔComplex

Predicting Ligand Affinity



Compared binding predictions with Random Forest and DNN regression models

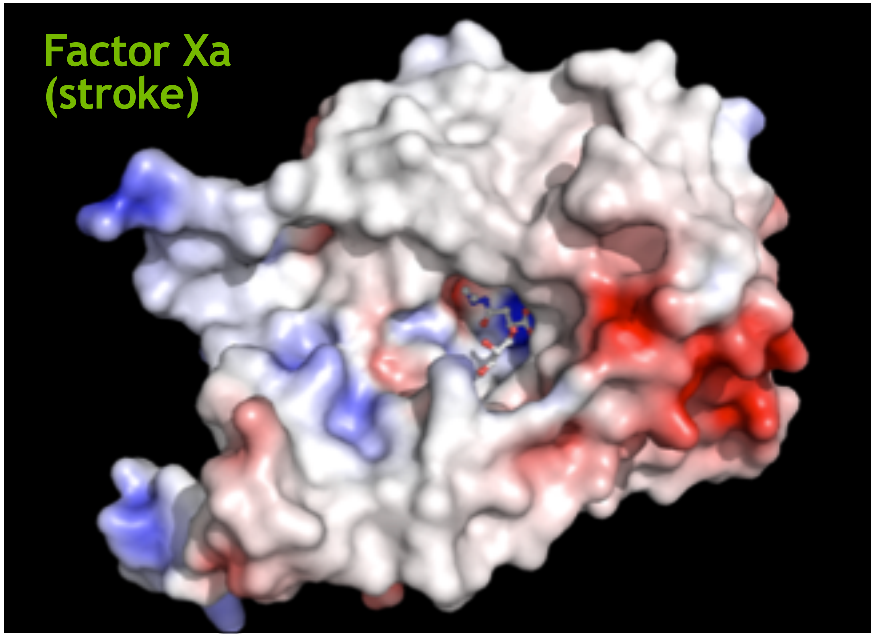
R^2 indicates DNN performs slightly better than Random Forest

DNN performance expected to improve with more data

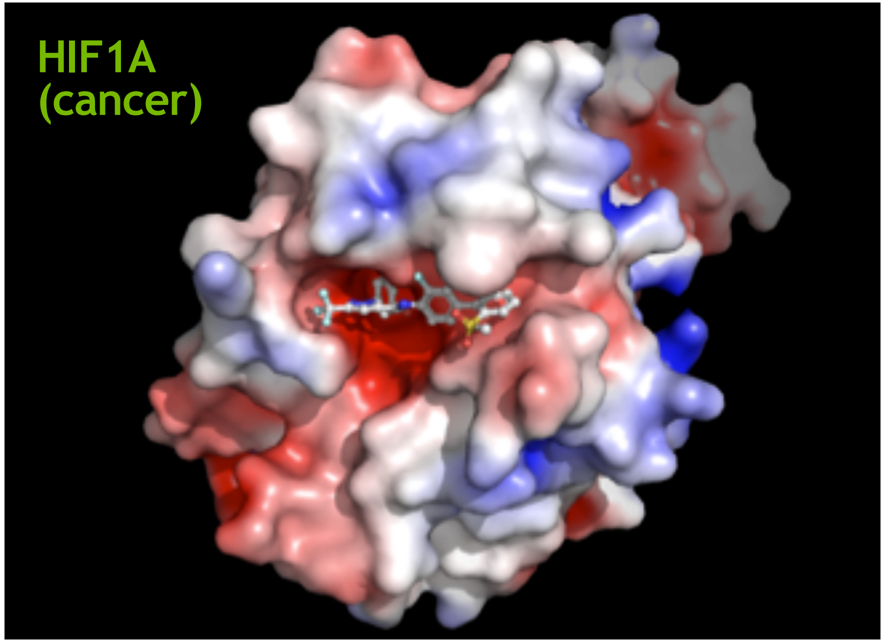
Predicting only successful ligand binding events (challenge)

Visualizing Enzyme-Ligand Binding

Tightest Binding Ligand



Weakest Binding Ligand



Opposite Protein-Ligand Charges Attract (+/-)

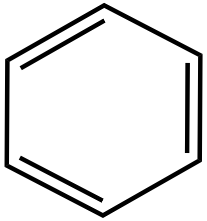
Deep Learning with Chemistry

Predicting enzyme-ligand binding using three-dimensional atomic convolutional networks

Generation of drug-like molecules using continuous latent spaces

Featurizing Chemicals for Deep Learning

Chemical



SMILES String

c1ccccc1

One-Hot Encoding

	c	1	n	...
c	1	0	0	0
1	0	1	0	0
c	1	0	0	0
c	1	0	0	0
c	1	0	0	0

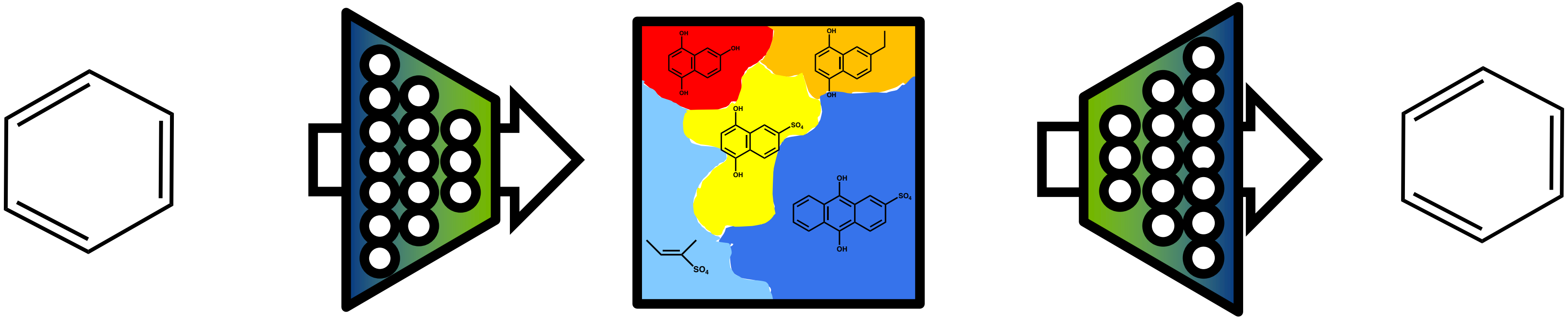
Chemicals first converted to SMILES strings

SMILES characters used to create one-hot encoded vectors

Vectors used as features for neural network

Character-level encodings don't always capture inherent rules of chemistry

Deep Learning Assisted Chemical Design



Chemical Input

Encoder Neural Network

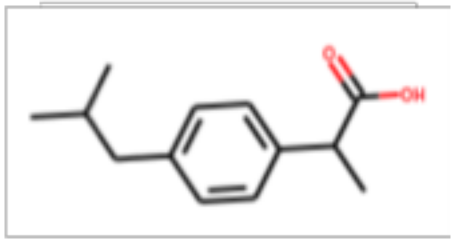
Continuous Molecular Representation

Decoder Neural Network

Chemical Output

Network based on: Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Augilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., Aspuru-Guzik, A. *"Automatic chemical design using a data-driven continuous representation of molecules."* *ACS Cent. Sci.*, 2018.

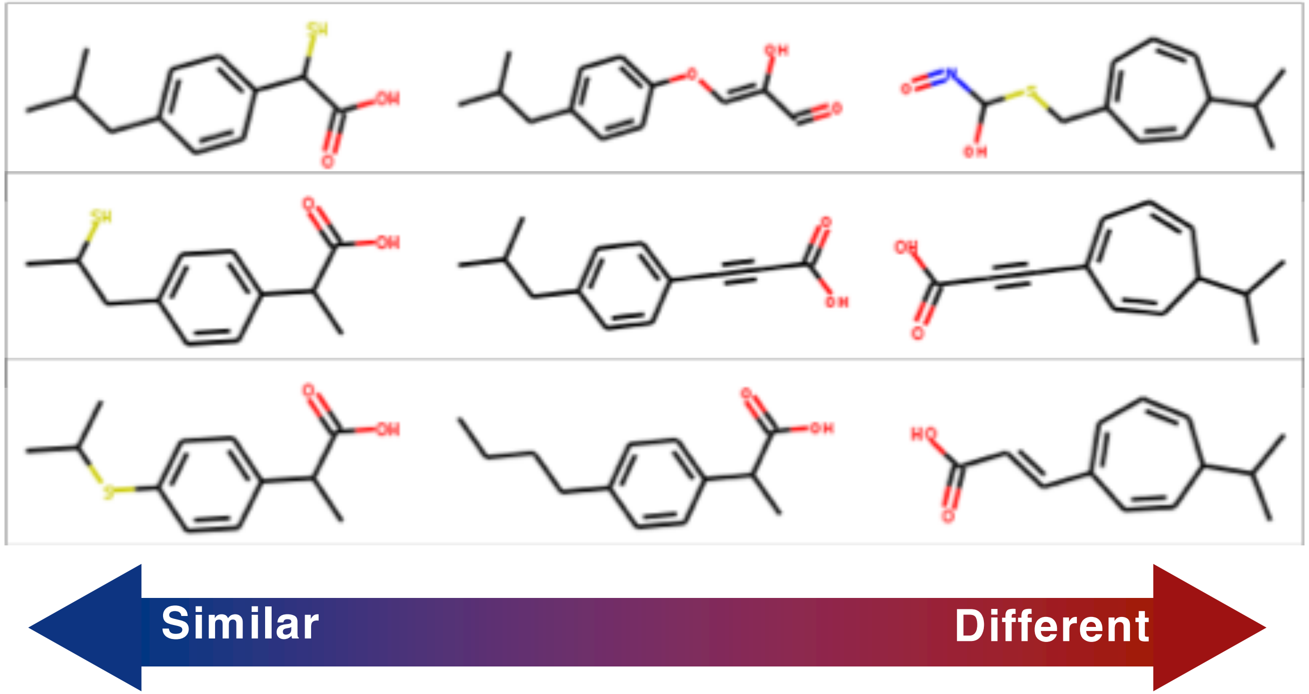
Exploring Chemical Space



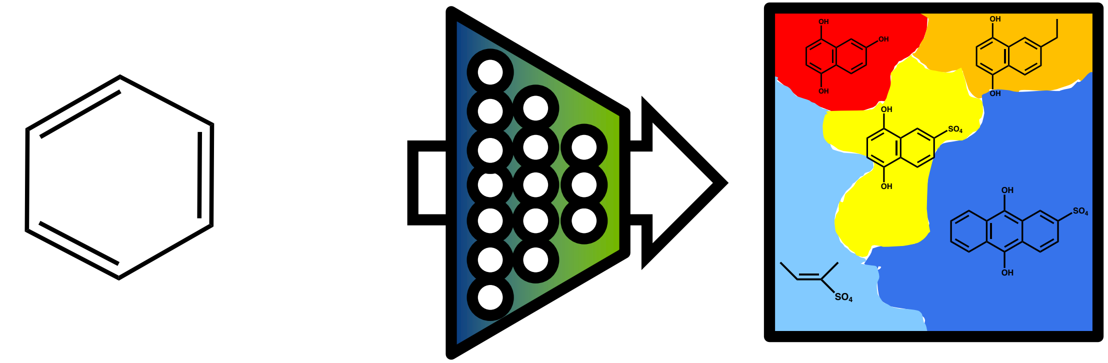
Ibuprofen

Molecules Sampled in Neighborhood of Ibuprofen

0.5 – 1.0 2.0 – 4.0 5.0 – 7.0



Generating Novel Chemicals with Deep learning

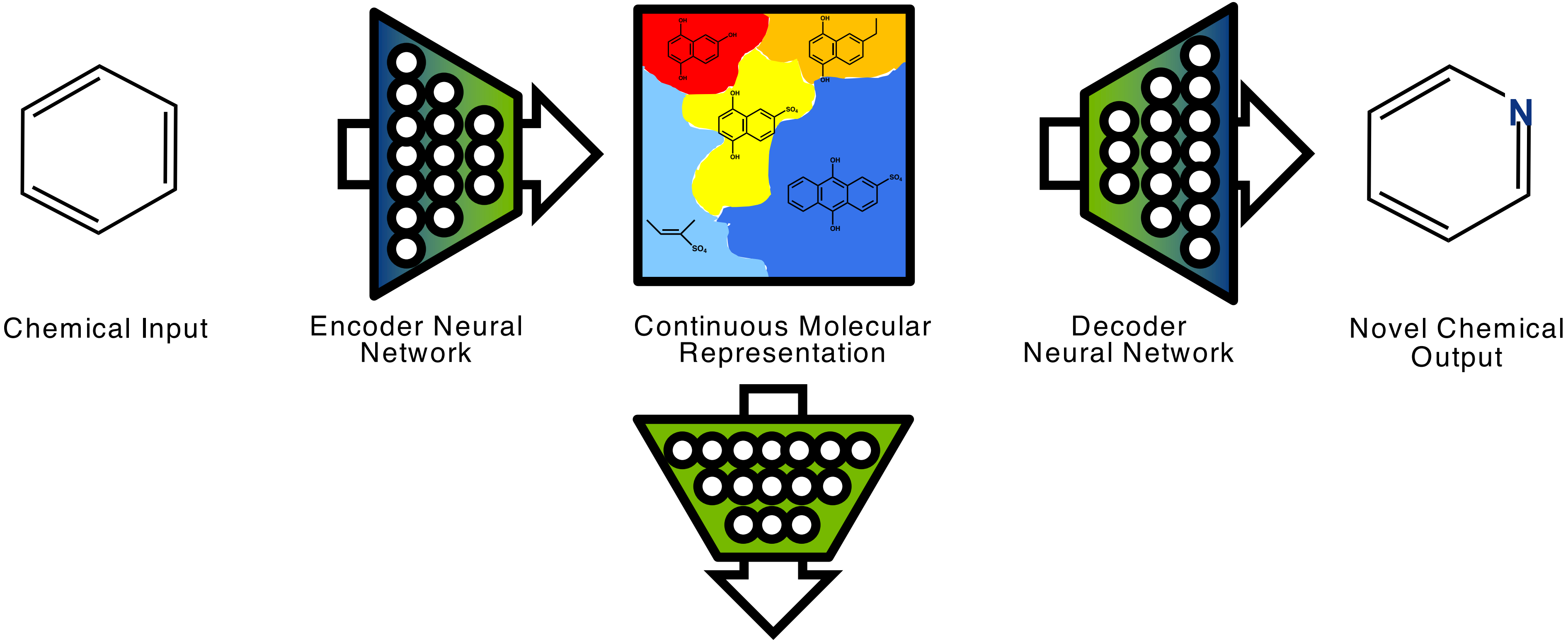


Chemical Input

Encoder Neural Network

Continuous Molecular Representation

Generating Novel Chemicals with Deep learning



Chemical Input

Encoder Neural Network

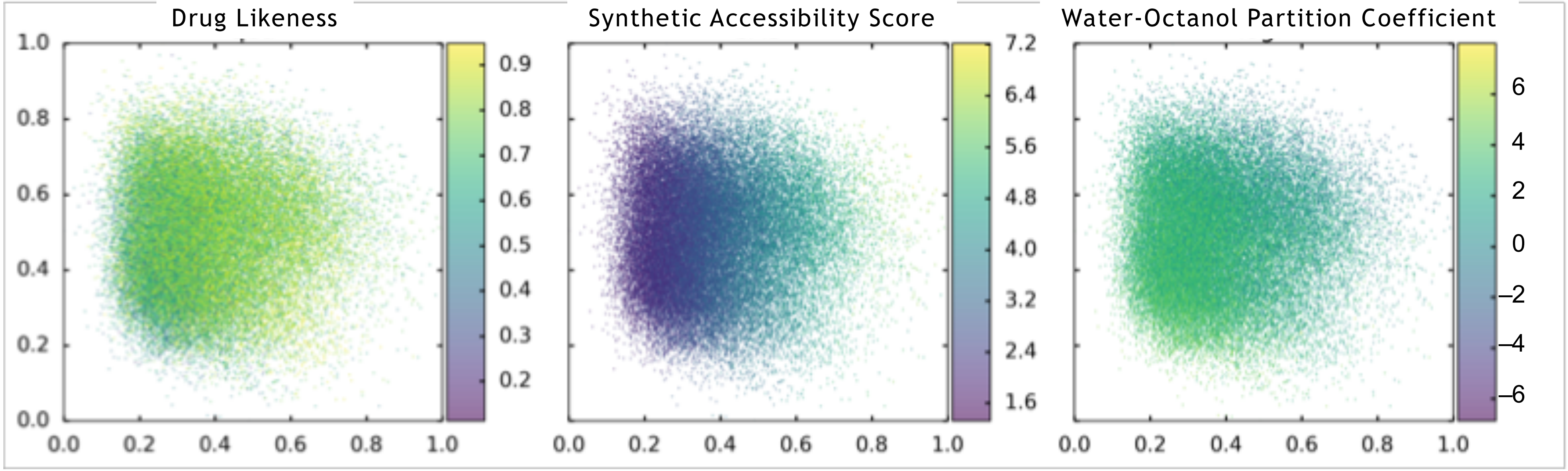
Continuous Molecular Representation

Decoder Neural Network

Novel Chemical Output

Chemical Property Prediction

Chemical Properties in Molecular Space



Conclusions

Artificial intelligence provides a nascent but powerful toolkit for accelerating chemistry-focused innovation

Combining deep learning with experimental and simulation data further accelerates this iterative process

Tomorrow's basic and applied science breakthroughs will blend AI and traditional methodologies

NVIDIA