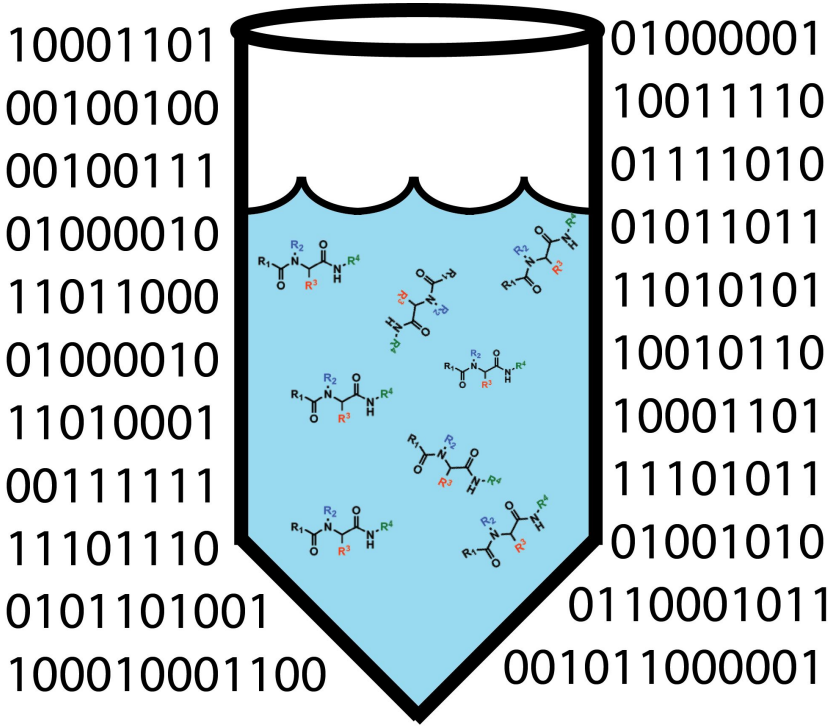
Storage and Computing with *Small* Molecules



PROF. BRENDA RUBENSTEIN

Joukowsky Family Assist. Prof. of Chemistry

MolPIGS, 2020

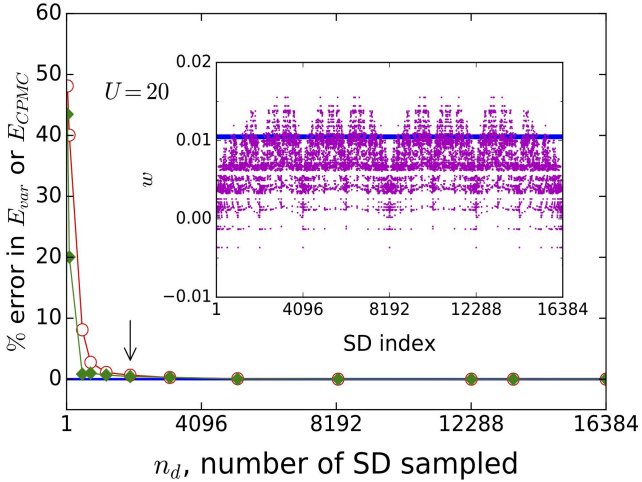
BROWN

THE RUBENSTEIN GROUP @ BROWN

Theoretical/Computational Chemistry and Physics

BROWN

Focus 1: Quantum Statistical Mechanics



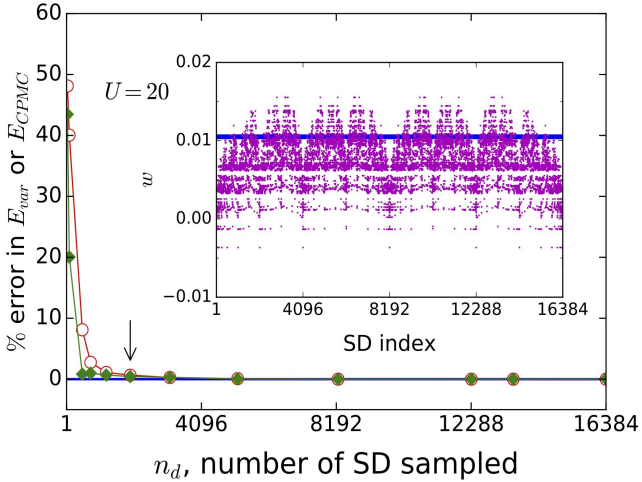
How Can Schrodinger's Equation Be Solved Faster and More Accurately?

THE RUBENSTEIN GROUP @ BROWN

Theoretical/Computational Chemistry and Physics

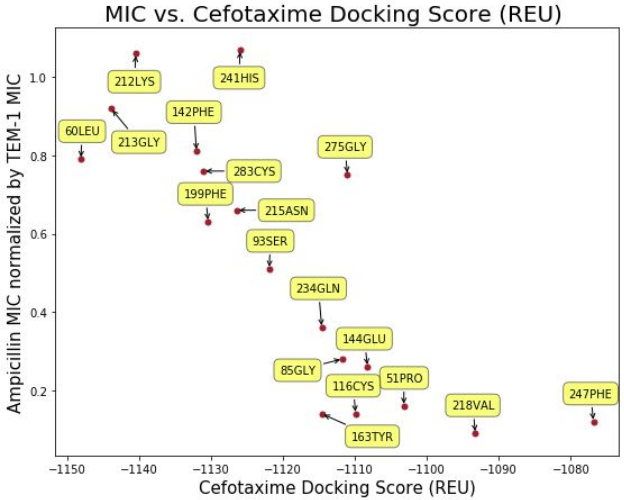
BROWN

Focus 1: Quantum Statistical Mechanics



How Can Schrodinger's Equation Be Solved Faster and More Accurately?

Focus 2: Computational Biophysics



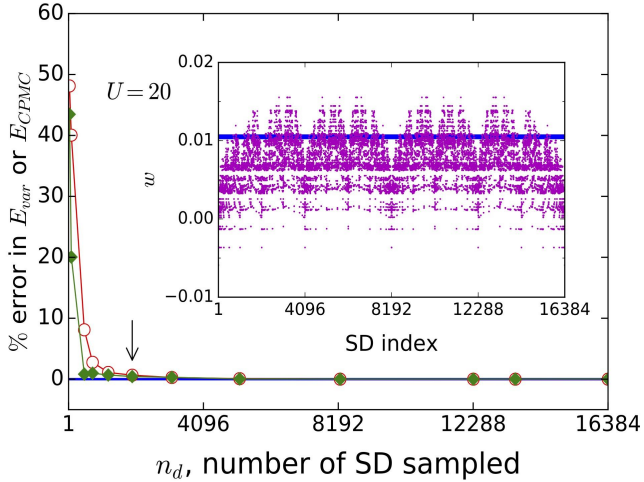
Can Thermodynamic Properties Be Used to Predict Organismal Fitness?

THE RUBENSTEIN GROUP @ BROWN

Theoretical/Computational Chemistry and Physics

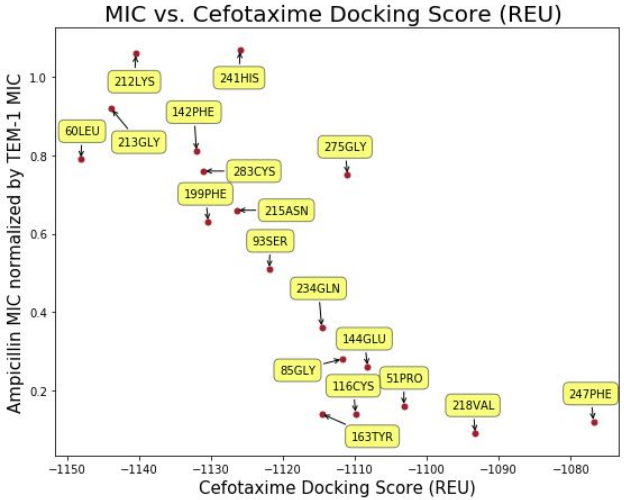
BROWN

Focus 1: Quantum Statistical Mechanics



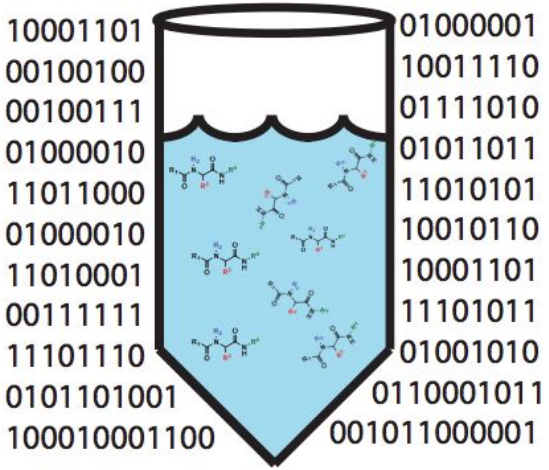
How Can Schrodinger's Equation Be Solved Faster and More Accurately?

Focus 2: Computational Biophysics



Can Thermodynamic Properties Be Used to Predict Organismal Fitness?

Focus 3: Quantum and Molecular Computing



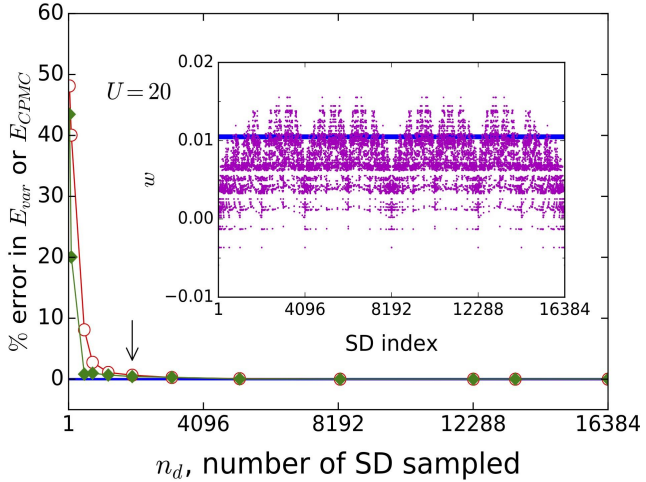
Can Molecules/Trapped Ions Be Used to Store and Compute on Information?

THE RUBENSTEIN GROUP @ BROWN

Theoretical/Computational Chemistry and Physics

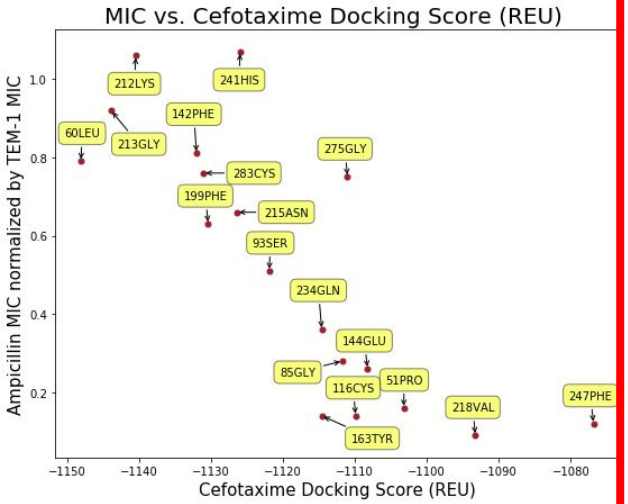
BROWN

Focus 1: Quantum Statistical Mechanics



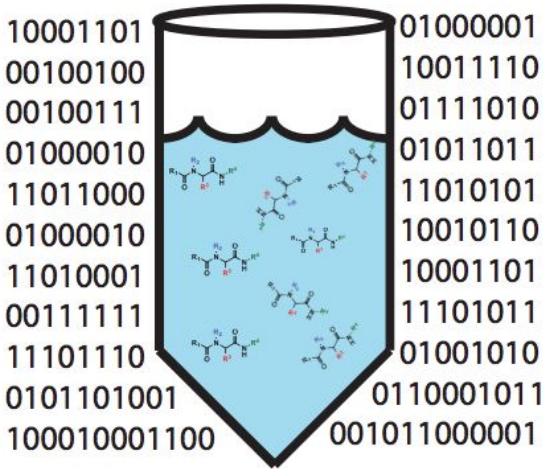
How Can Schrodinger's Equation Be Solved Faster and More Accurately?

Focus 2: Computational Biophysics



Can Thermodynamic Properties Be Used to Predict Organismal Fitness?

Focus 3: Quantum and Molecular Computing



Can Molecules/Trapped Ions Be Used to Store and Compute on Information?

TALK OUTLINE

BROWN

1. MOTIVATION FOR MOLECULAR STORAGE AND COMPUTING: Density, New Computing

2. MOLECULAR STORAGE: Via Mixtures

3. MOLECULAR COMPUTING: Molecular Perceptrons, Chemical Steganography

4. OUTSTANDING QUESTIONS

MOTIVATION

MY FIRST ACQUAINTANCE WITH COMPUTERS

Fujitsu's FACOM 655 Line Printer (1977)

- Satisfied the "need for computers with greater compactness"
- Printed 340 lines/min and 62 character types

MY FIRST ACQUAINTANCE WITH COMPUTERS

Fujitsu's FACOM 655 Line Printer (1977)

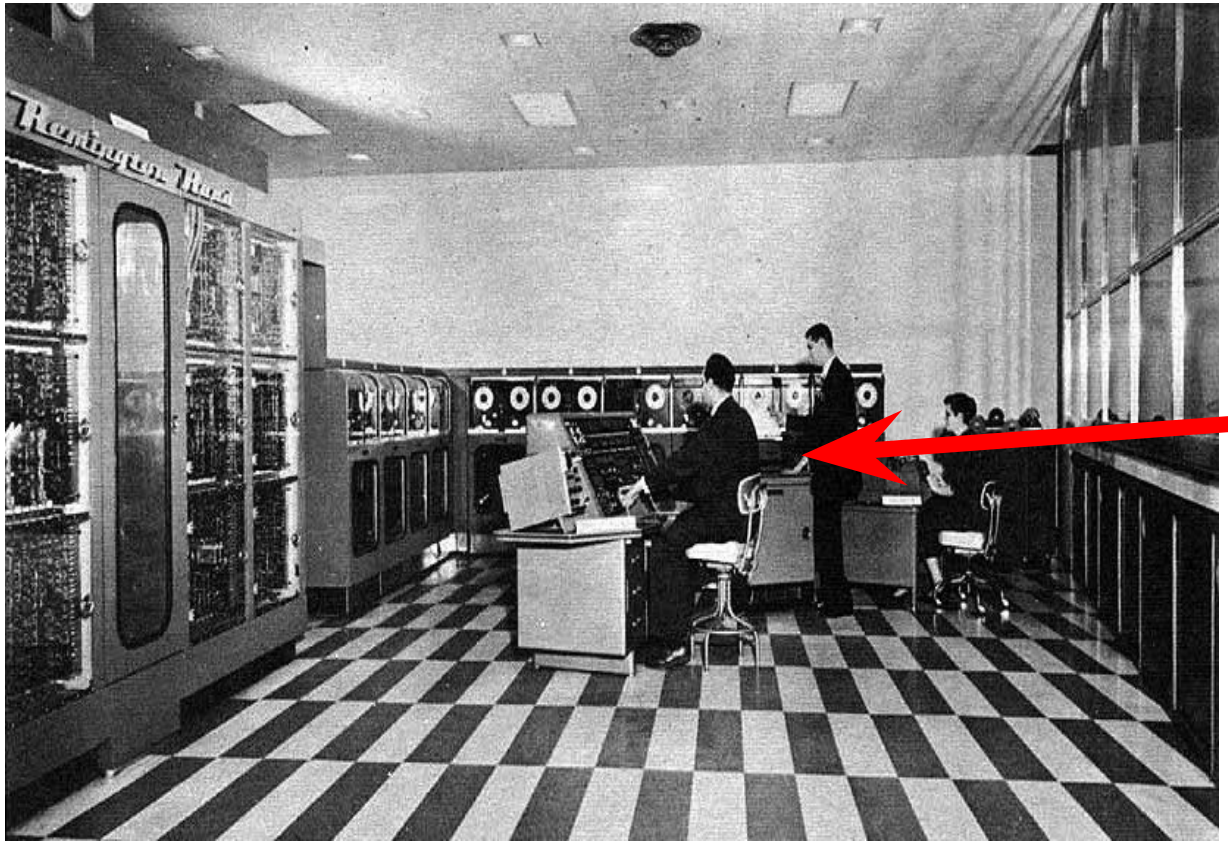
Totally Awesome Hiding Place!

MY FIRST ACQUAINTANCE WITH COMPUTERS

Fujitsu's FACOM 655 Line Printer (1977)

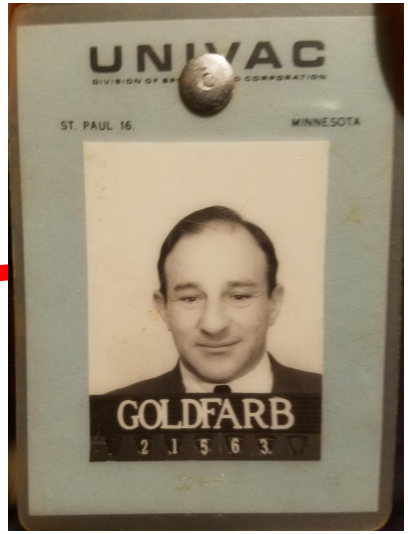
10.5-in. Magnetic Tapes, IBM Mainframes (1980s)

BUT, THEY COULD BE EVEN OLDER...



UNIVAC I, US Census Bureau (1952)

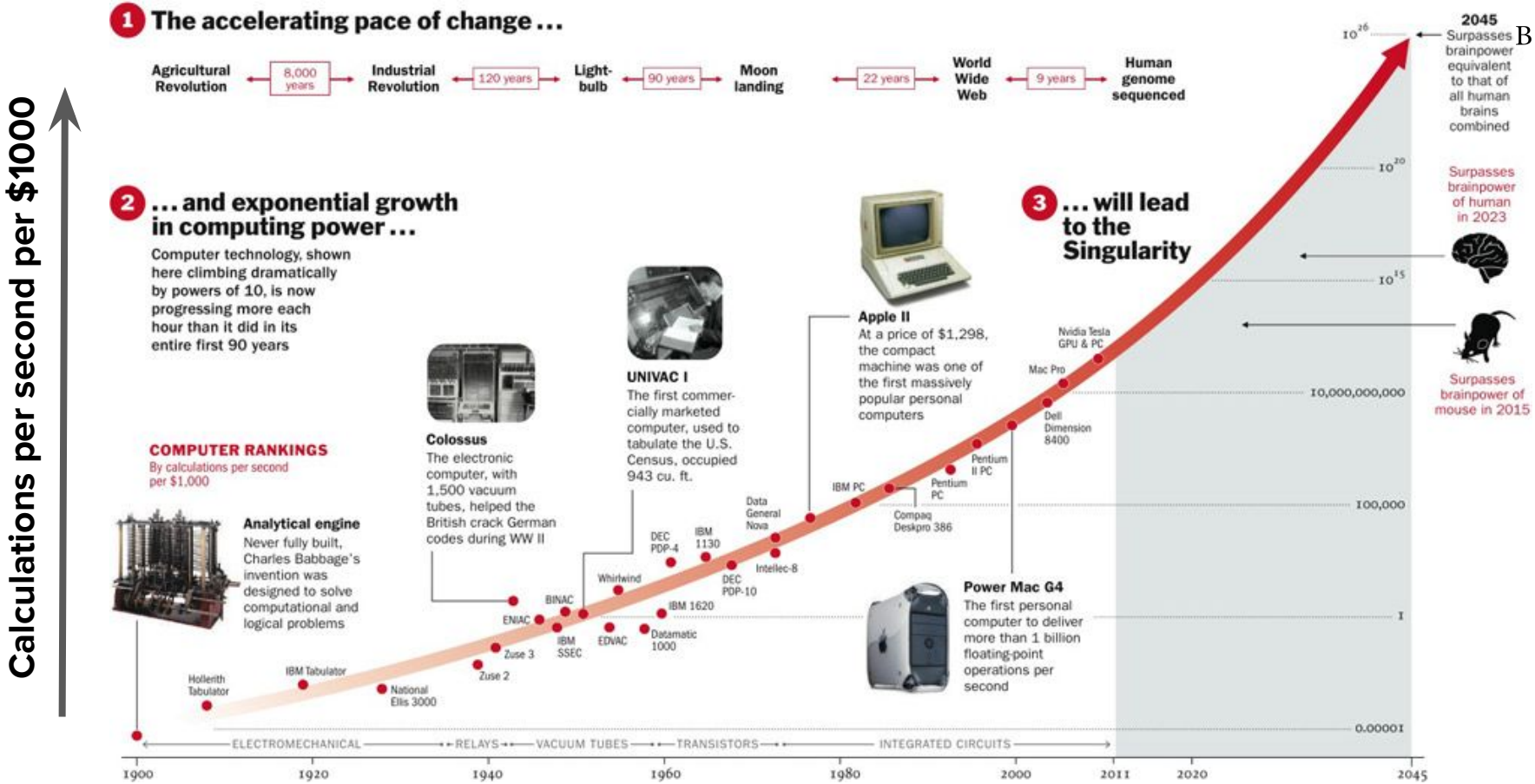
Correctly forecasted the 1952 election!



Grandpa Vick

Mathematician at Sperry-Rand, UNIVAC Programmer

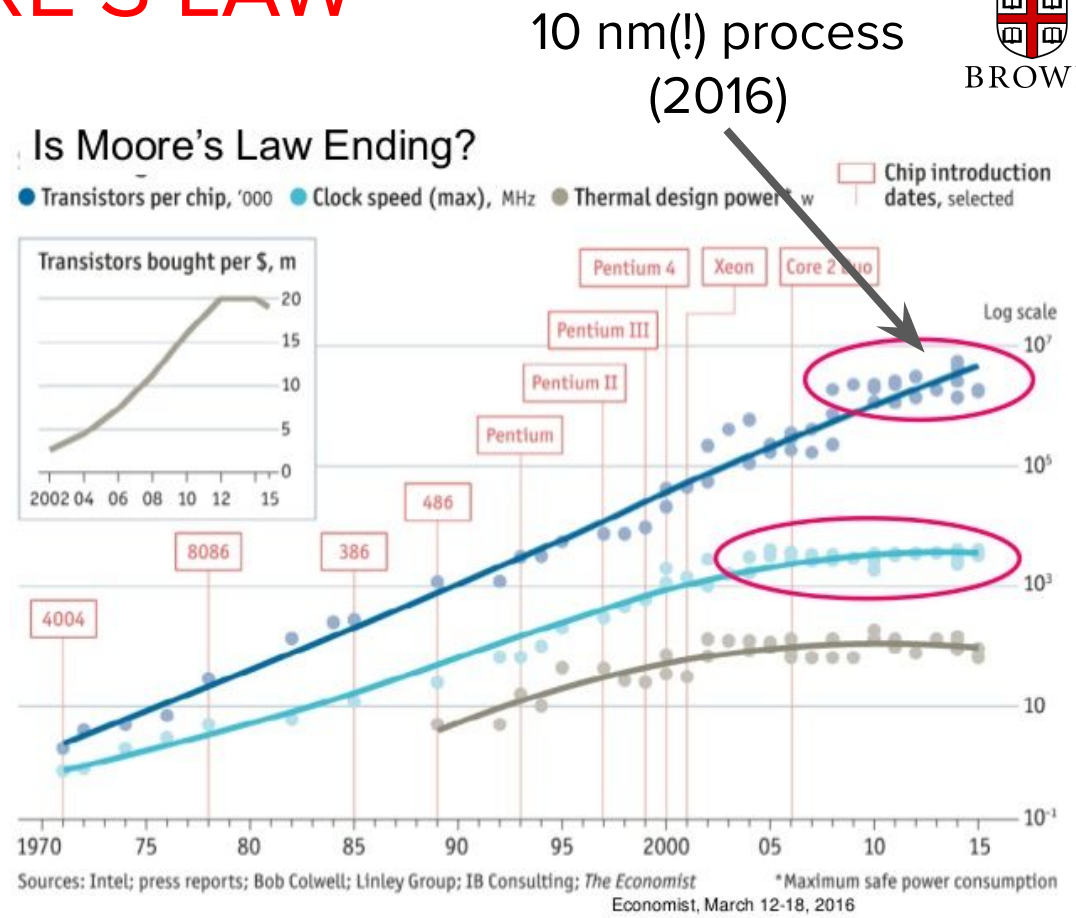
GROWTH OF COMP. POWER



L. Grossman, *Time Magazine* (2011).

THE END OF MOORE'S LAW

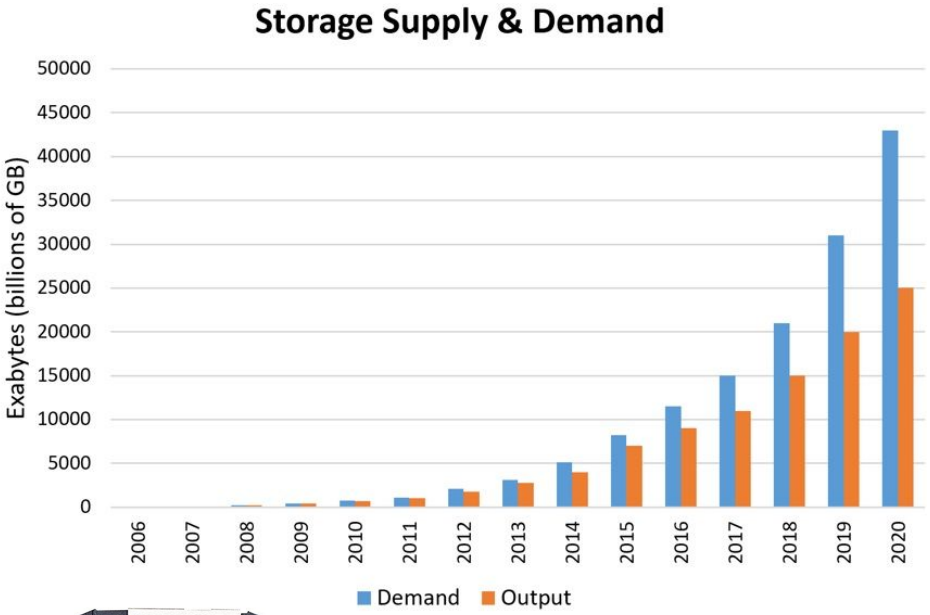
- Moore's Law: the number of transistors per chip will double every 18 months
- This law has held true since enunciated in 1965
- HOWEVER, WE MAY BE NEARING THE DEATH of MOORE'S LAW:
  - Transistors Too Small!
  - Expensive!
  - Massive power demands!



THE PROBLEM: BIG DATA

- We produce 2.5 exabytes (10^{18} bytes!) of data per day
- Our data production is growing at a rate of 40% per year
- In 2013, the size of the total digital universe was 4.4 zettabytes (10^{21} bytes). In 2020, it is estimated to become 44 zettabytes.

THE BIGGER PROBLEM: WHERE TO STORE?



MBs

TBs

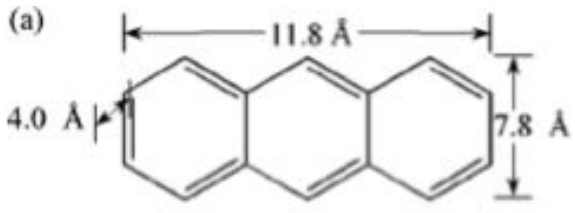
- Storage capacity is growing far, far slower than the digital universe.
- In 2013, average storage capacity could hold 33% of the digital universe. In 2020, it will be able to hold less than 15% of the digital universe.
- Storage capacity doesn't even address challenges with I/O speeds or the ability to randomly access data.

Recode, http://www.eetimes.com/author.asp?section_id=36&doc_id=1330462.

THE DAWN OF MOLECULAR COMPUTING

BROWN

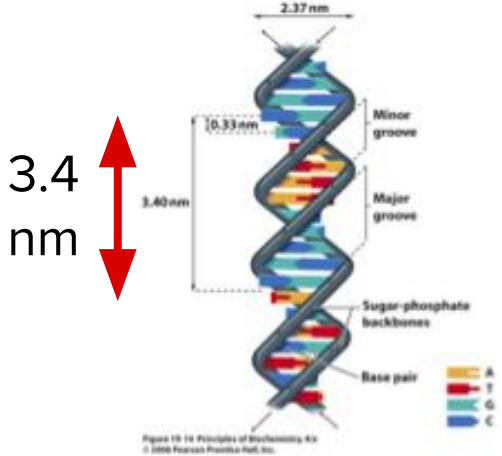
- It will be difficult to miniaturize transistors below 5 nm
  - Quantum mechanical tunneling intercedes
- BUT, MOLECULES ARE NATURALLY OF THESE DIMENSIONS!
- *Why not compute using molecules?*



Anthracene

1.09 Å

Nitrogen



3.4 nm

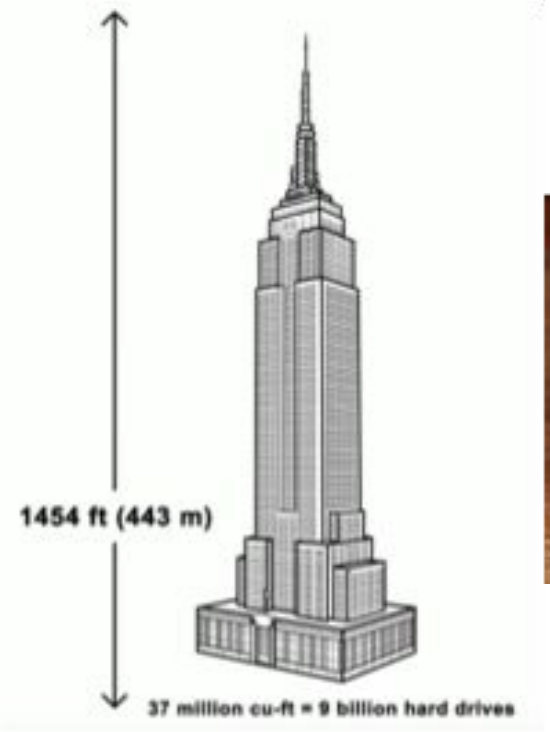
DNA

THE DAWN OF MOLECULAR COMPUTING

BROWN

ADVANTAGES of MOLECULES

- Found Everywhere
- Low Power Demands
- *Billions of Billions of Billions* of Them in Small Volumes
- Embarrassingly Parallel
- *We Know How To Manipulate with Chemistry!*

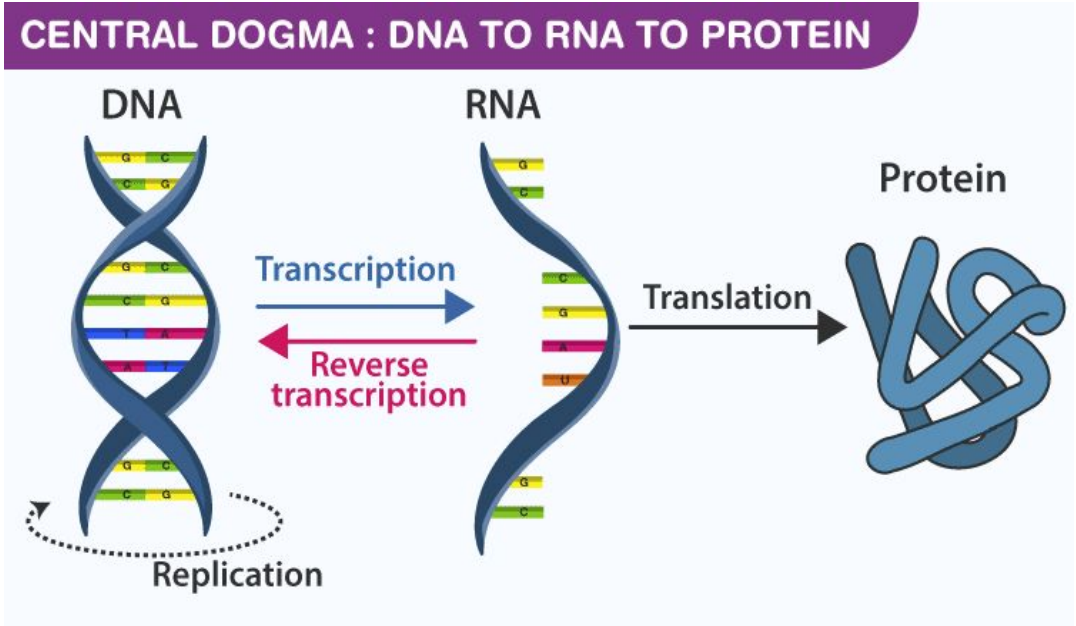


1 Beaker of Water May Be Able to Store 200 Empire State Buildings Worth of Data!

DNA AS THE ULTIMATE STORAGE MEDIUM

BROWN

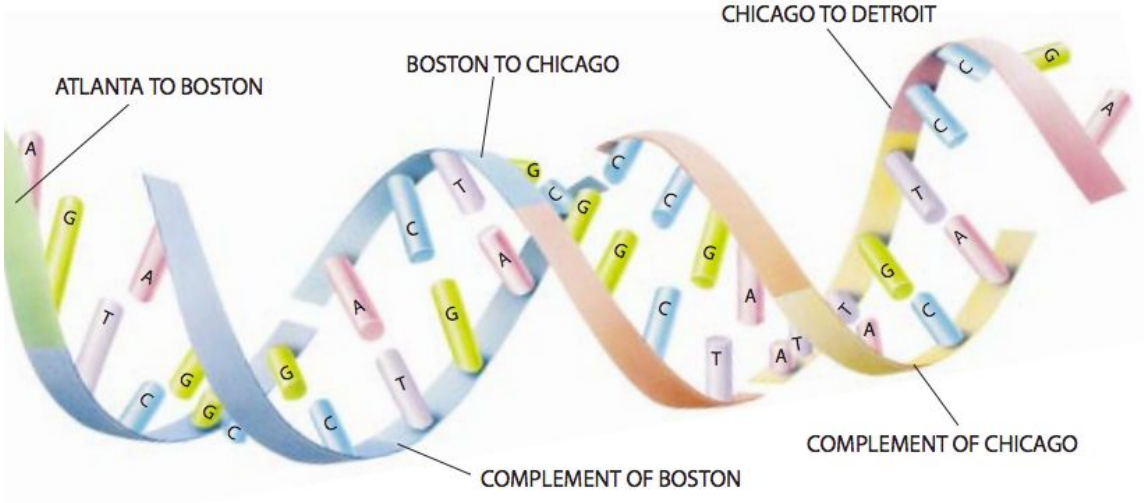
- Biomolecules, such as DNA, store large amounts of information (1.5 GB in human genome!) in a small space
- Millions/Billions of years of evolution have led to Low-Error DNA Processing Techniques (“Bio Software”)

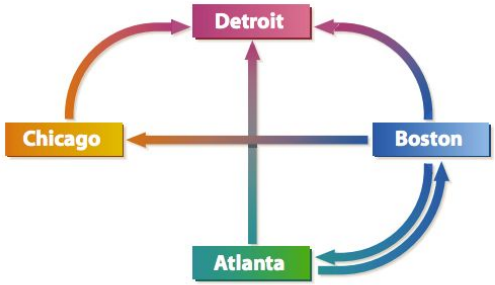


DNA COMPUTING

Useful for Solving Problems That Can Be Constructed In Terms of Path Minimization (SAT Problems, Traveling Salesman Problem)

The Traveling Salesman Problem





CITY	DNA NAME	COMPLEMENT
ATLANTA	ACTTGCAG	TGAACGTC
BOSTON	TCGGACTG	AGCCTGAC
CHICAGO	GGCTATGT	CCGATACA
DETROIT	CCGAGCAA	GGCTCGTT

FLIGHT	DNA FLIGHT NUMBER
ATLANTA - BOSTON	GCAGTCGG
ATLANTA - DETROIT	GCAGCCGA
BOSTON - CHICAGO	ACTGGGCT
BOSTON - DETROIT	ACTGCCGA
BOSTON - ATLANTA	ACTGACTT
CHICAGO - DETROIT	ATGTCCGA

Adelman, *Science* (1994); *Scientific American* (1998); Lipton, *Science* (1995).

THE CHALLENGE

BROWN

Can We Develop Ways of Storing Information Within and Computing With *Small* Molecules?

THE CHALLENGE

BROWN

Can We Develop Ways of Storing Information Within and Computing With *Small* Molecules?

- What are the tradeoffs that accompany small unordered mixtures vs. larger ordered polymers?

THE CHALLENGE

BROWN

Can We Develop Ways of Storing Information Within and Computing With *Small* Molecules?

- What are the tradeoffs that accompany small unordered mixtures vs. larger ordered polymers?
- What *pressures* does storage place on synthesis and detection?

THE CHALLENGE

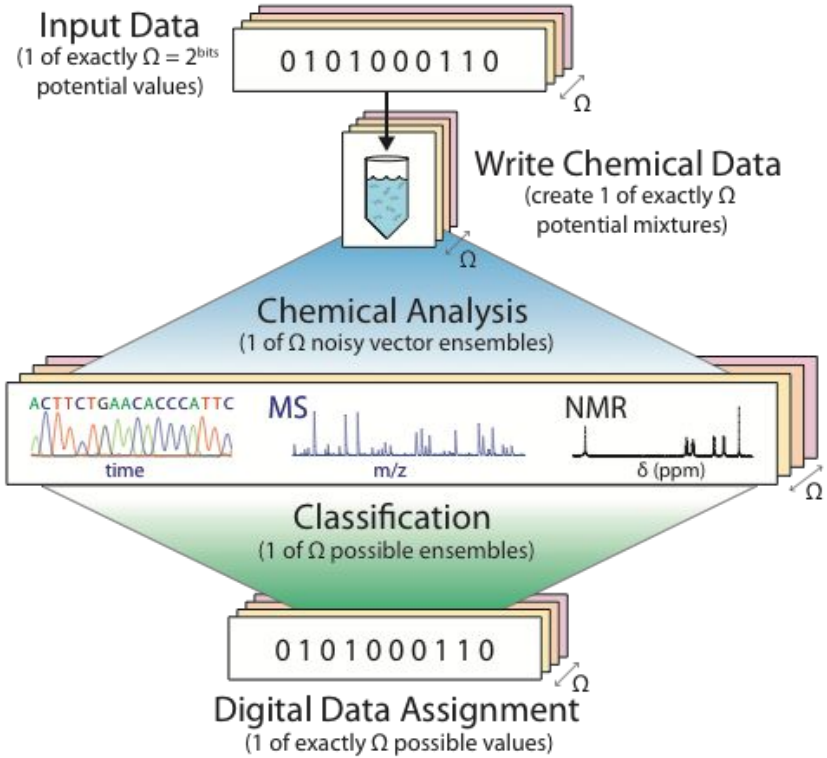
BROWN

Can We Develop Ways of Storing Information Within and Computing With *Small* Molecules?

- What are the tradeoffs that accompany small unordered mixtures vs. larger ordered polymers?
- What *pressures* does storage place on synthesis and detection?
- What are fundamentally *molecular* computations?

MOLECULAR STORAGE

READING AND WRITING DATA IN CHEMICAL MIXTURES



Writing chemical data means creating 1 out of $\mathbf{\Omega}=2^{\text{bits}}$ possible chemical states (mixtures or sets of mixtures).

Reading a chemical dataset corresponds to figuring out which 1 out of the $\mathbf{\Omega}$ states it is.

Rosenstein *et al.*, *Trans. NanoBioSciences* (2020).

OUR APPROACH

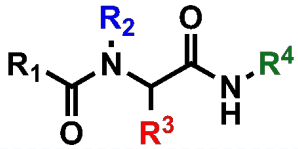
BROWN

Goals

- To store abstract digital data by taking advantage of small molecule chemical diversity
- To develop leading-edge synthesis and analysis techniques that can scale to *billions* of molecules/day
- To establish scalable strategies for solution-phase information processing using small-molecule reactions

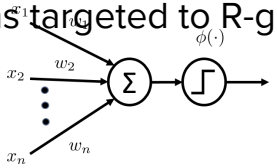
Synthesis:

We will encode information in collections of small molecules.



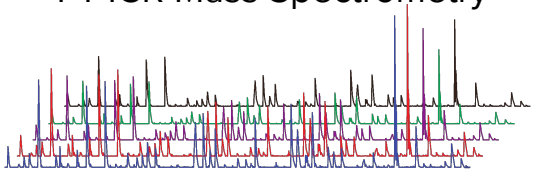
Computation:

Solution-phase classifiers using reactions targeted to R-groups.



Detection:

FT-ICR Mass Spectrometry

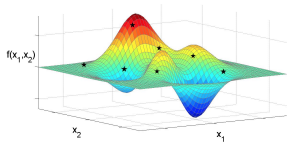


Data:

Images, audio, and weather series.

Analysis:

Custom CAD software to optimize synthesis, select optimal chemical encodings of datasets, and automate statistical data recovery from multi-dimensional datasets.

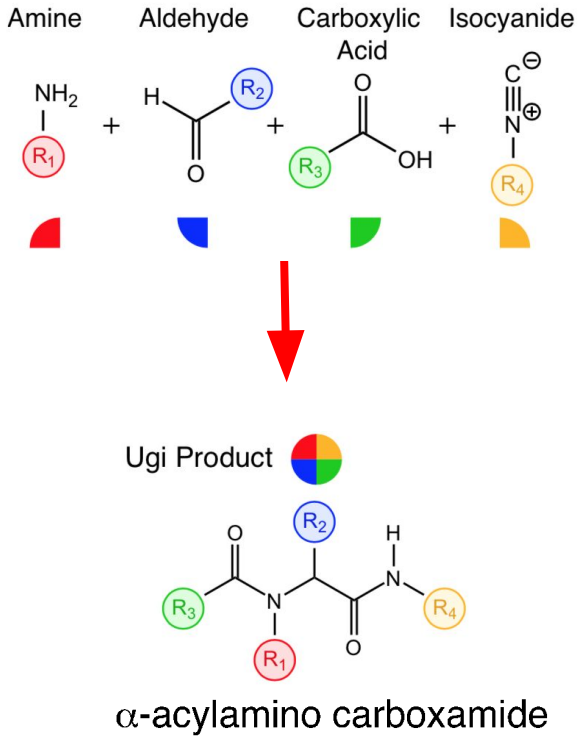


MULTICOMPONENT REACTIONS

We will encode our data in molecules produced via Ugi reactions, which are:

- High-yielding
- Combinatorial
- Soluble in reaction solvents (acetonitrile, methanol; water)
- Yield Stable, Nonpolar Products
- Readily characterized by NMR and mass spectrometry
- Amenable to tandem reactions
- Like peptides!

Ugi 4-Component Reaction



Combinatorial:
If 10 R^1s, R^2s, R^3s, and R^4s,
10,000 Different Compounds

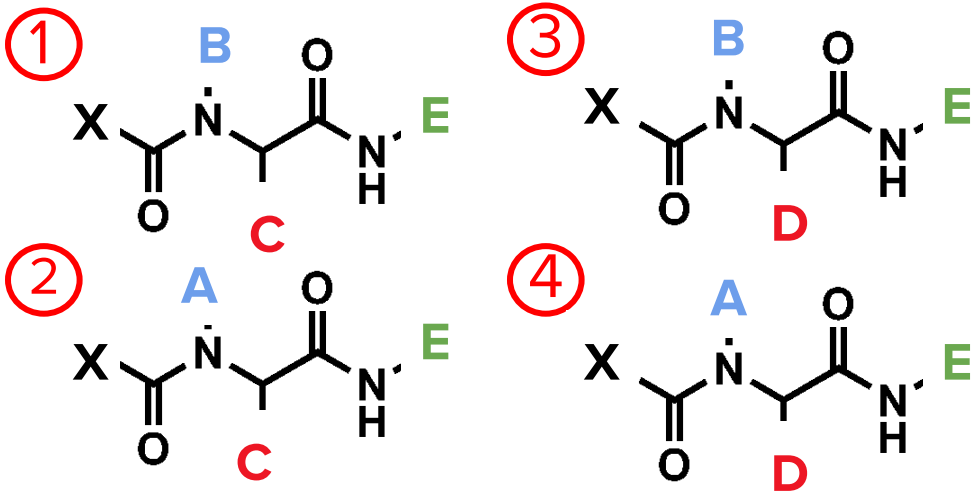
HOW WOULD YOU WRITE SOMETHING?

BROWN

Alphanumeric: Rhode Island ACS

Binary: 0101 0010 0110 1000 0110 1111 0110 0100
0110 0101 0010 0000 0100 1001 0111 0011
0110 1100 0110 0001 0110 1110 0110 0100
0010 0000 0100 0001 0100 0011 0101 0011

Molecular:



Let There Be Two Groups Per Site:

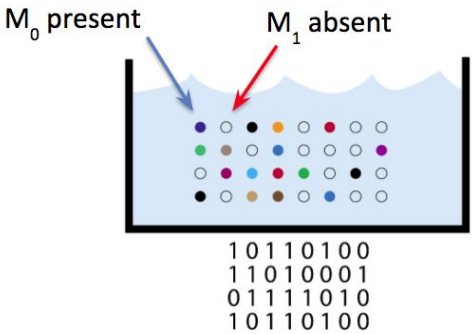
R_1: X (0), Y (1)

R_2: A (0), B (1)

R_3: C (0), D (1)

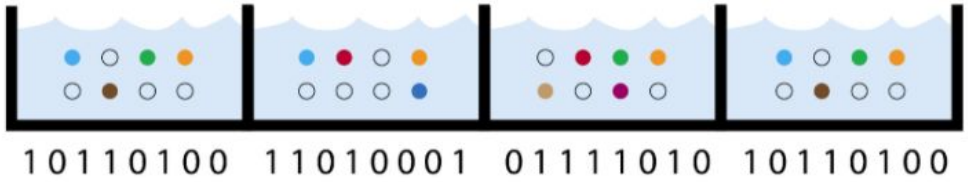
R_4: E (0), F (1)

DATA ENCODING AND TRADEOFFS



One Mixture

M = 800,000,000
unique molecules



Many Smaller Mixtures

D separate mixtures with different combinations of M unique molecules:

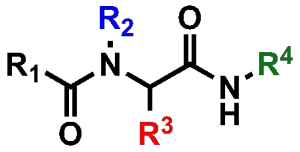
e.g., D = 25,000
100MB / 25,000 = 4KB
M = 32,000 unique molecules

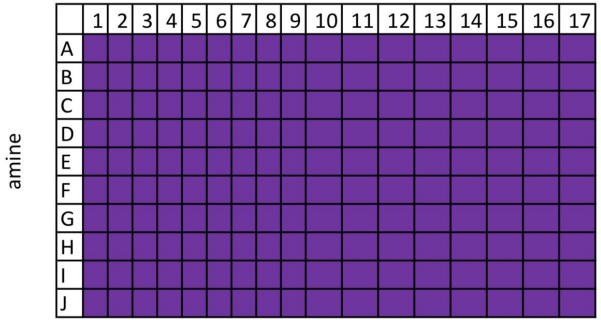
>30,000 Molecules still pushes MS bounds and analysis speed

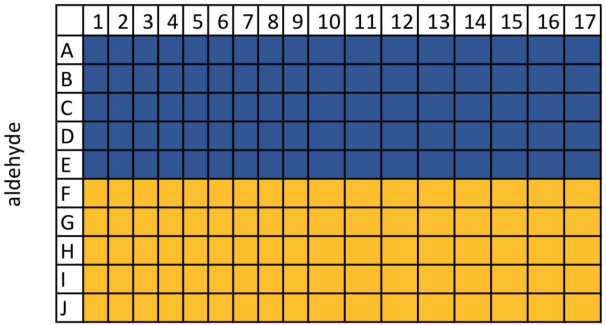
COMBINATORIAL LIBRARY SYNTHESIS

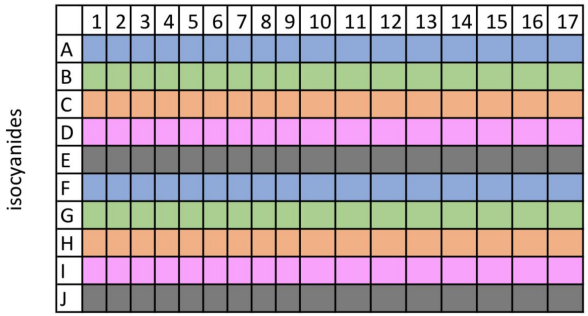
BROWN

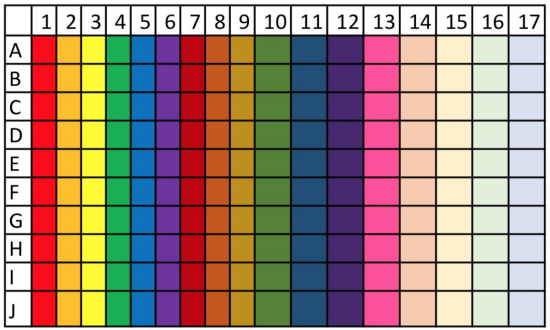
Example: 1 amine, 2 aldehydes, 17 carboxylic acids, 5 isocyanides
(1x2x17x5 = 170 Ugi compounds)



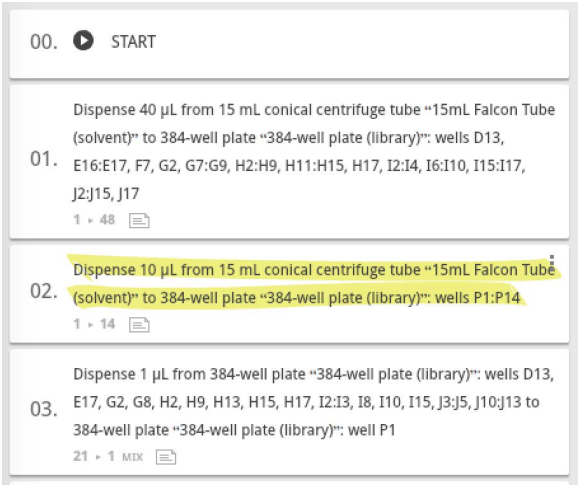








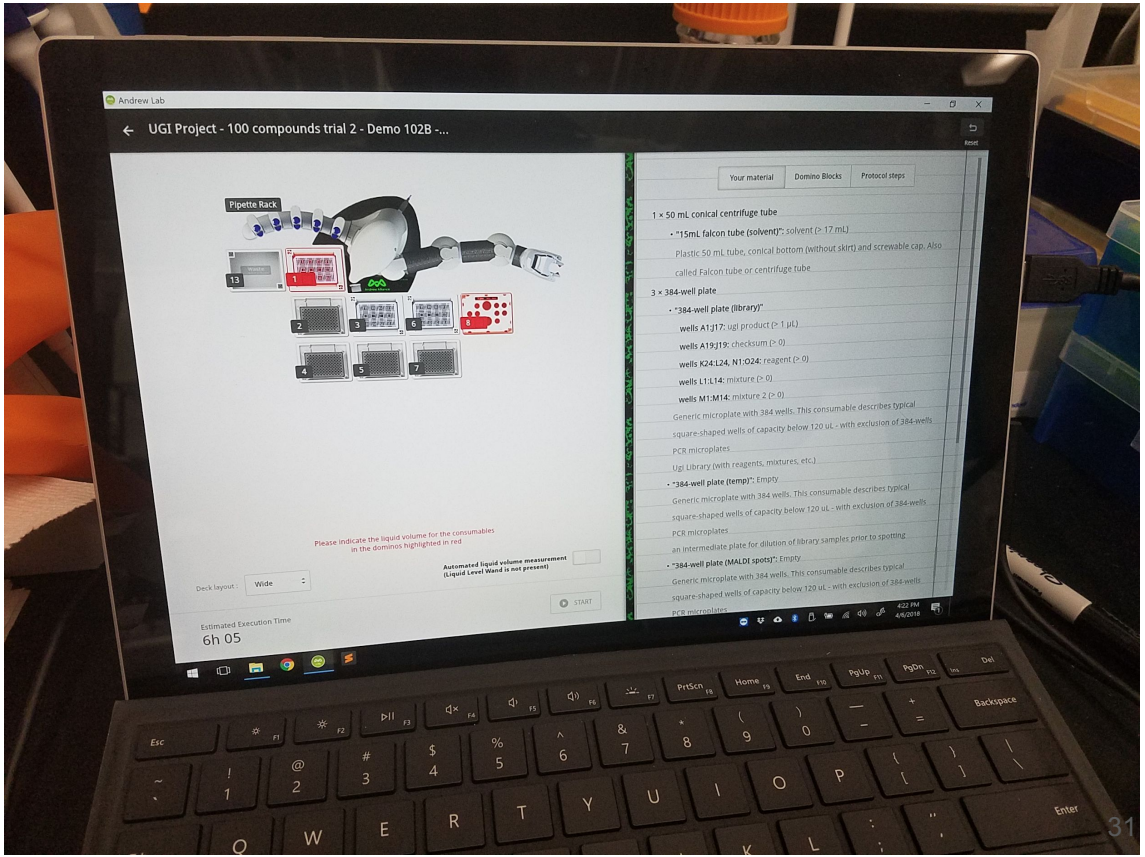
→ Translate into automation protocol



[…]

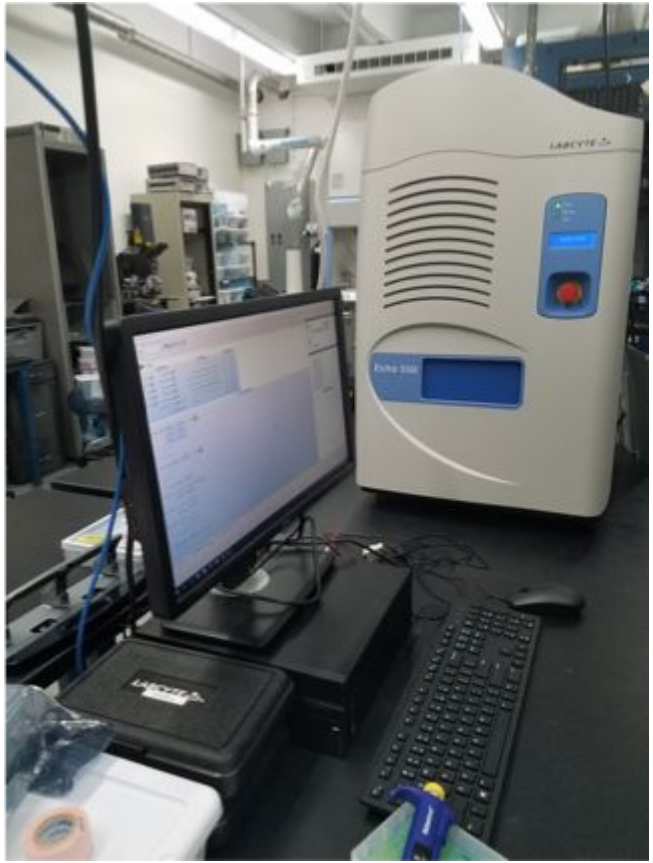
COMBINATORIAL LIBRARY SYNTHESIS

BROWN



ACOUSTIC LIQUID HANDLING

BROWN



Labcyte Echo 550

- 384- and 1536-well plates
- Robotic any-to-any well transfers
- No pipette tips, contactless
- 2.5 nanoliter droplets
- Transfers 200 droplets per second
- Capable of >100,000 transfers per day

FT-ICR MASS SPECTROMETRY

BROWN

Bruker Daltonics SolariX XR 7T FT-ICR

Will rely on a Fourier-Transform Ion Cyclotron Resonance (FT-ICR) Mass Spectrometer for Detection:

- Determines Mass Based Upon Lorentz's Law: F = qv x B
- Mass Resolution: 10^6
- Mass Range: 100-10000 Da
- Loading: 10^6 to 10^7 Charges
- Minimum Number of Ions for Detection: ~100
- ~10^4 to 10^5 compounds per loading

https://www.bruker.com/products/mass-spectrometry-and-separations/ftms/solarix/technical-details.html.

OVERALL EXPERIMENTAL WORKFLOW

BROWN

Manual Chemistry

Input Data

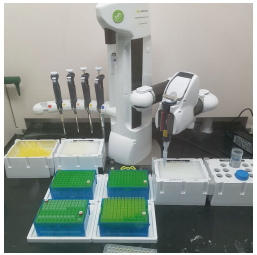
Output Data

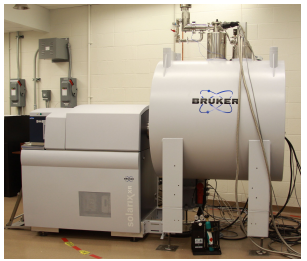
Software

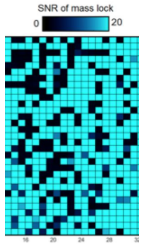
Select Reagents	CAD for Library Synthesis	CAD for Data Mapping and Spotting	Data Analysis

Automated Experiments

Prepare Reagents	Robotic Library Synthesis	Automated Mixture Creation	Acoustic MALDI Spotting	MALDI FT-ICR MS



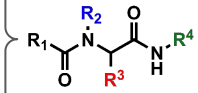




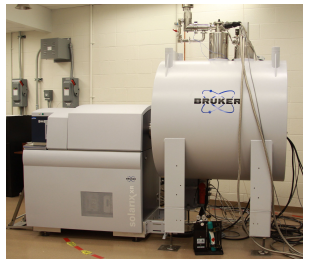
MASS SPEC OF INDIVIDUAL COMPOUNDS

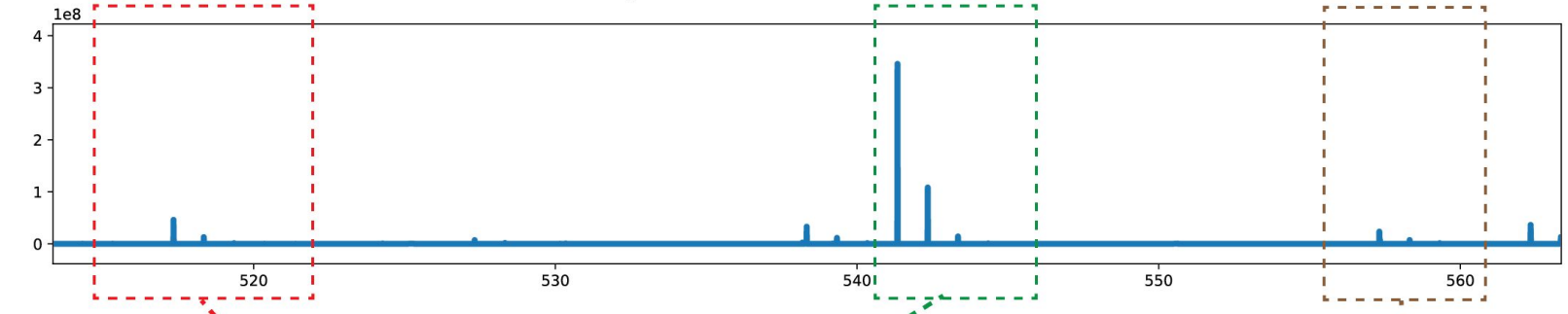
Library 1
20180227_102Ugi_individual

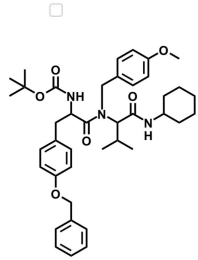
[{'cid_aldehyde': 6561,
'cid_amine': 7504,
'cid_carboxylic_acid': 1581890,
'cid_isocyanide': 23577,
'composition': {'C': 28, 'H': 46, 'N': 4, 'O': 5},
'location': 'F8',
'mass': 518.3468]

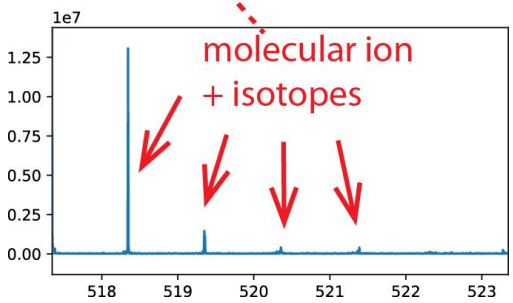


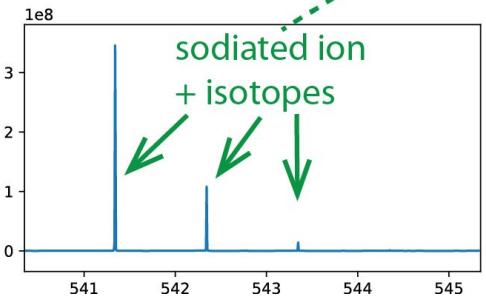
Bruker SolariX
FT-ICR MS

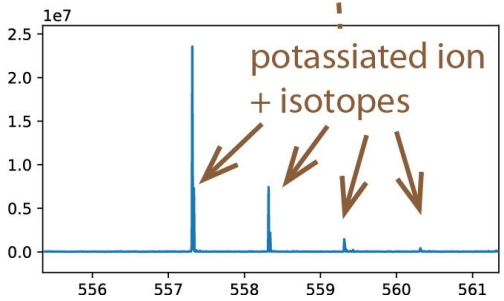








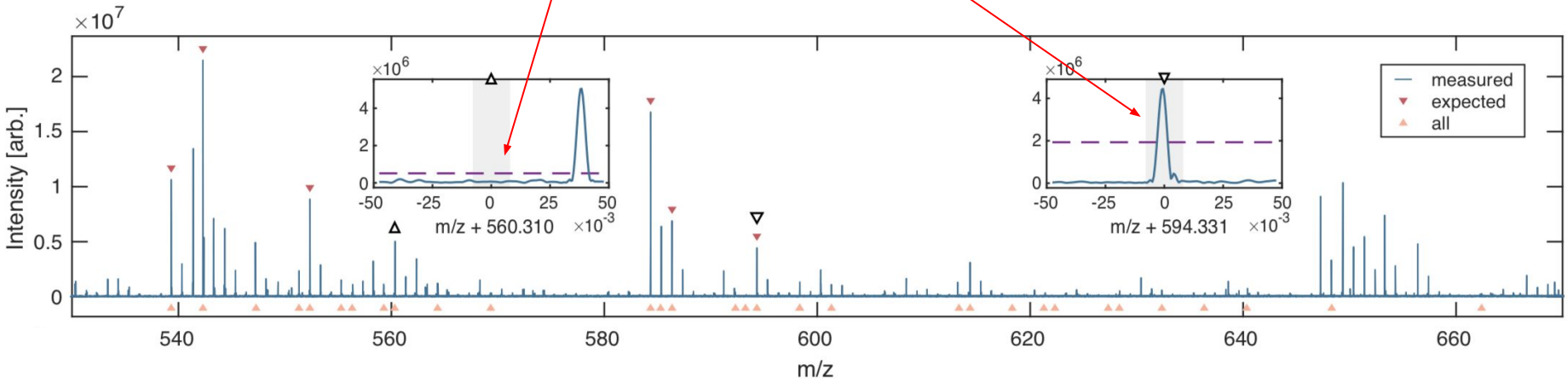




MIXTURE MASS SPECTRA

BROWN

32-Ugi data: 11001000 00010100 10000000 00000000



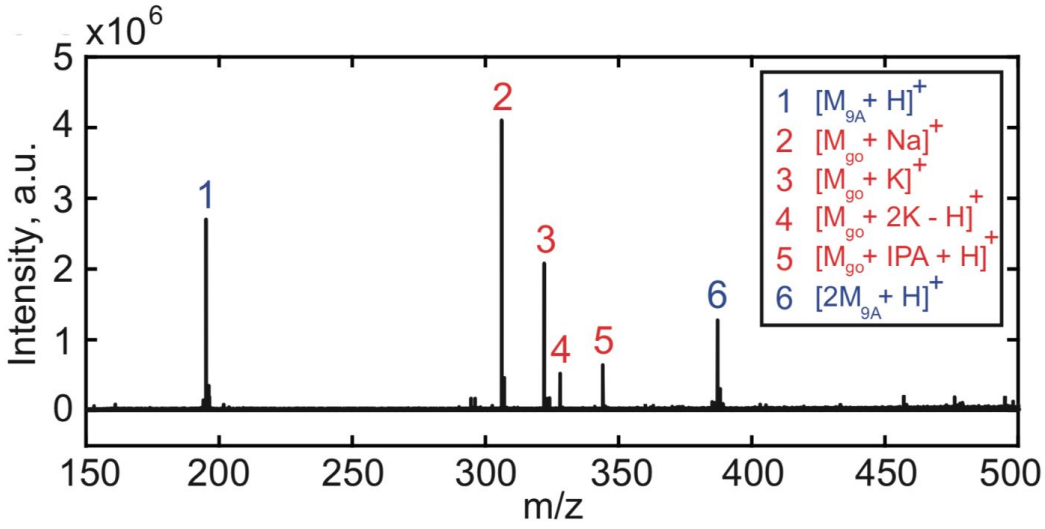
There are only 0-32 library compounds, but there are thousands of MS features.

MULTI-PEAK STATISTICAL ANALYSIS

BROWN

For each compound, we see multiple MS peaks:

- ^{13}C isotopes
- M+H, M+Na, M+K, and others….
- Clusters of 2M, 3M, 4M
- Matrix peaks
- Instrument harmonics

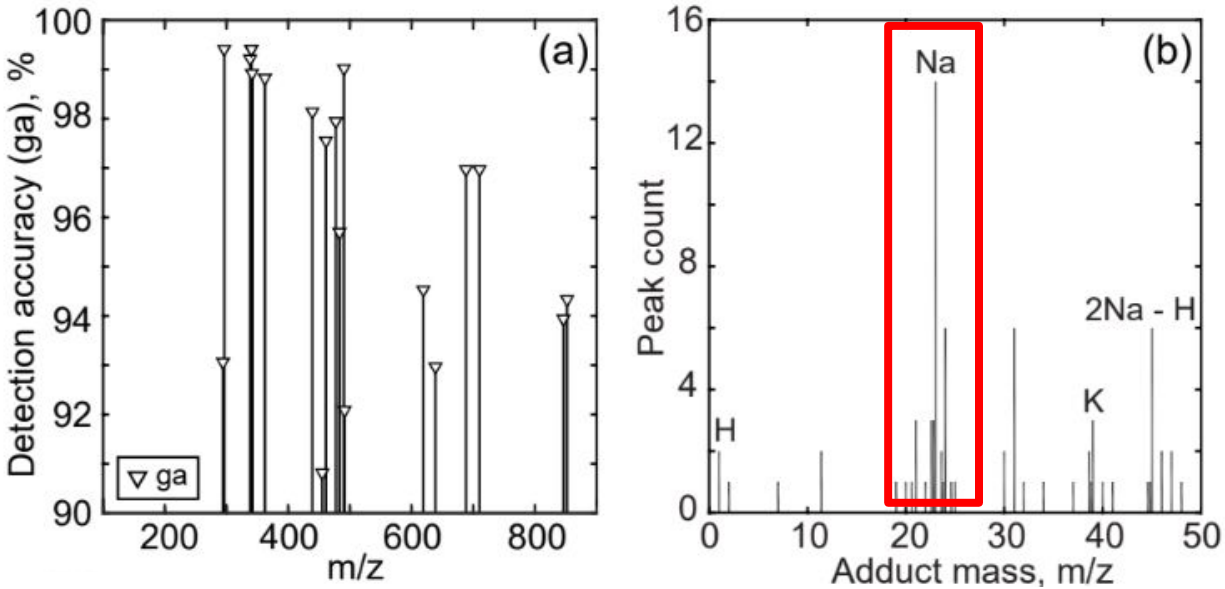


These peaks are generally consistent for one sample and one measurement condition, but their amplitudes are difficult to predict ahead of time => Learn Spectra, Don't Trace Peaks!

E. Kennedy *et al.*, *PLoS One* (2019).

MULTI-PEAK STATISTICAL ANALYSIS

We can rank peaks by discrimination power, and train a classifier using the top N peaks - improves accuracy, generality.



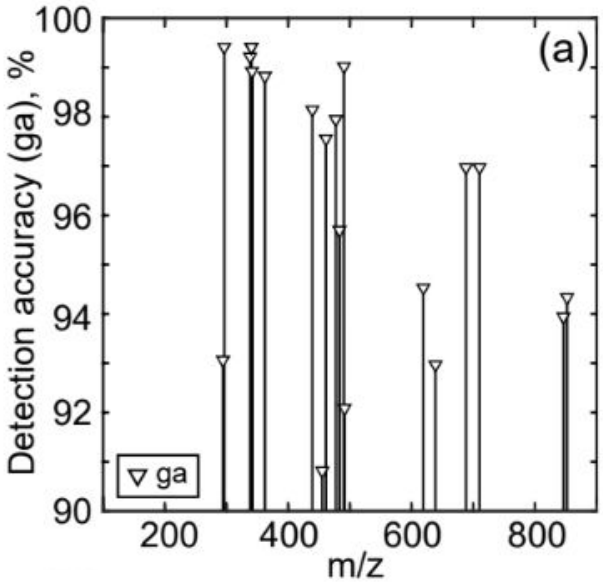
Identify Peaks with Greatest Detection Accuracy

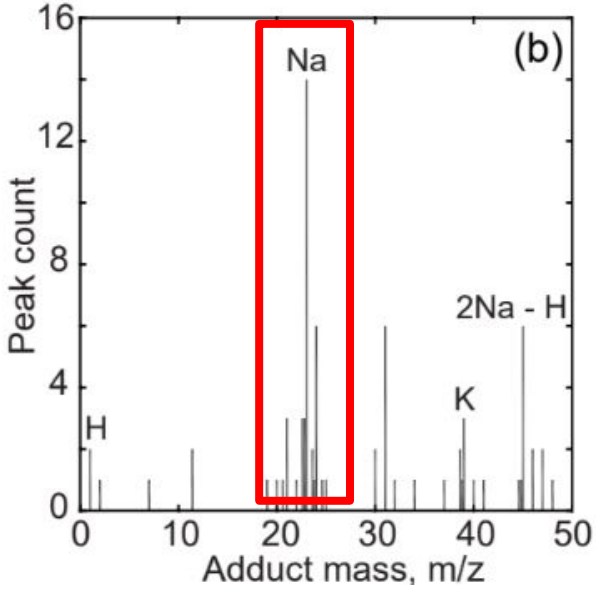
E. Kennedy et al., *PLoS One* (2019).

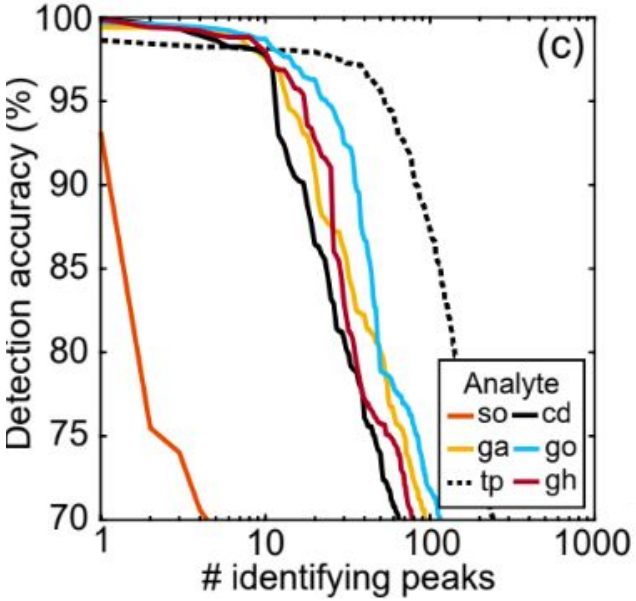
MULTI-PEAK STATISTICAL ANALYSIS

BROWN

We can rank peaks by discrimination power, and train a classifier using the top N peaks - improves accuracy, generality.







Identify Peaks with Greatest Detection Accuracy

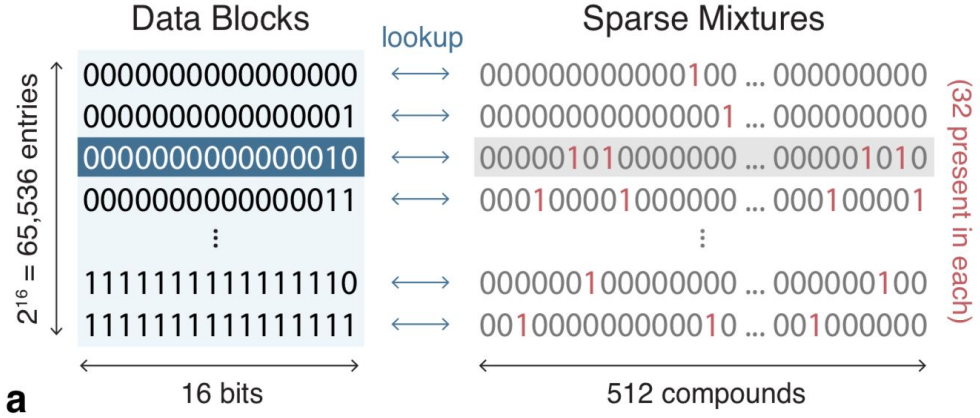
Use Peaks Above Threshold

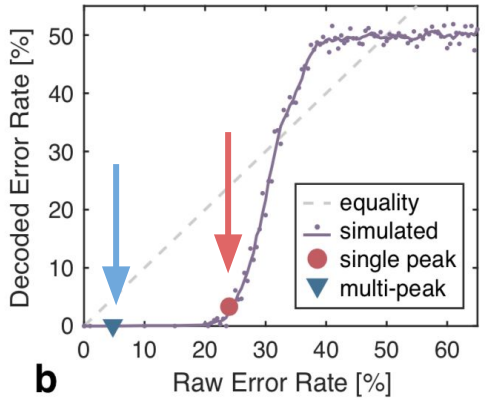
E. Kennedy et al., *PLoS One*, 2019.

HARNESSING SPARSITY

BROWN

Mapping each data block to one mixture out of many, many more increases accuracy.





23,336 pixels

Sparse data-to-mixture mapping

- 16 bits per mixture
- 512 molecules present, *but only use 32*

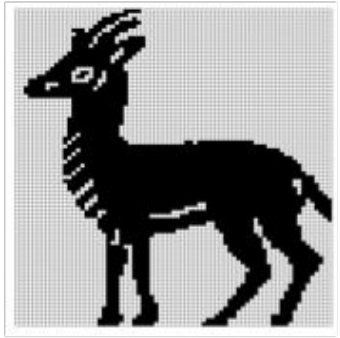
99.89% accuracy

Arcadia *et al.*, *Nature Comm.* (2020).

READ/WRITE OF DIGITAL IMAGES

BROWN

83 x 74 = 6142 bits

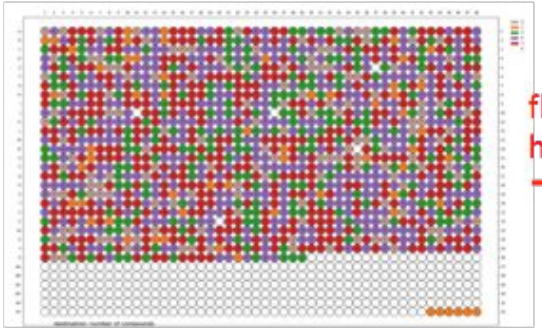


vectorize

(Including Mapping Sparsity, Coding, Training Regressors, etc.)

$$\underset{1 \times 6142}{[\text{image data vector}]} \underset{\text{known permutation matrix}}{[\quad]} = \underset{1 \times 6142}{[\text{chemical data vector}]}$$

Data vector mapped to 1229 mixtures with a library of 5 chemicals (1229 x 5 = 6145)

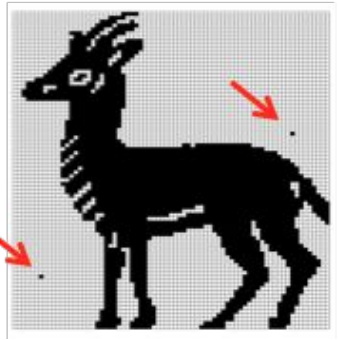


fluid handling

MALDI plate

FT-ICR MS, peak finding, inverse permutation

Recovered image
6140/6142 correct
= 99.97%



Arcadia *et al.*, *Nature Comm.* (2020).

MOLECULAR DATA STORAGE GALLERY

BROWN

768 bits

6,142 bits

8,904 bits

17,424 bits

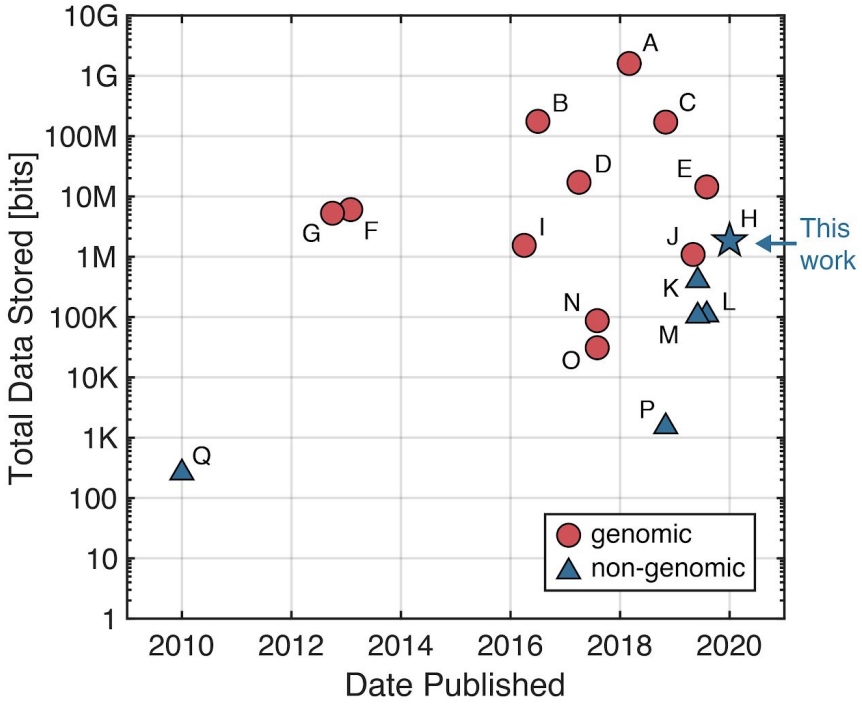
97,969 bits

125,630 raw bits

257,544 raw bits

Arcadia *et al.*, *Nature Comm.* (2020).

COMPARISON WITH OTHER TECHNIQUES

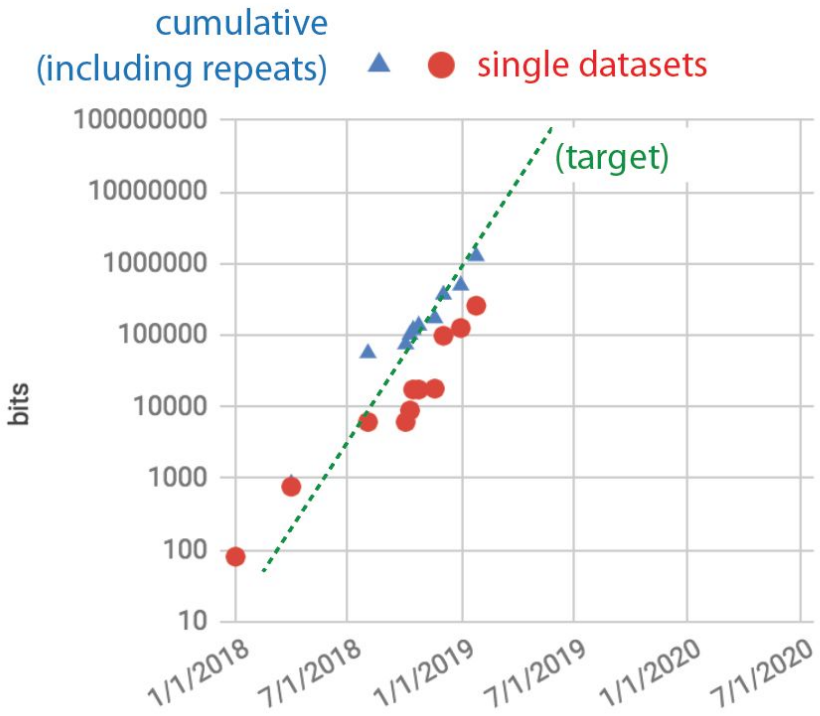


ID	1st Author	Published	Total [bits]	Chemistry
A	Organick	2018 (Feb.)	1,601,866,400	DNA
B	Blawat	2016 (Jun.)	176,000,000	DNA
C	Anavy	2018 (Oct.)	171,040,180	DNA
D	Erlich	2017 (Mar.)	17,174,528	DNA
E	Lopez	2019 (Jul.)	14,392,000	DNA
F	Goldman	2013 (Jan.)	6,056,408	DNA
G	Church	2012 (Sep.)	5,270,208	DNA
H	Arcadia	2019 (Dec.)*	1,844,527	Ugi products
I	Bornholt	2016 (Mar.)	1,544,000	DNA
J	Choi	2019 (Apr.)	1,090,032	DNA
K	Cafferty	2019 (May.)	402,880	peptides
L	Kennedy	2019 (Jul.)	108,692	metabolites
M	Rosenstein	2019 (May.)	104,111	Ugi products
N	Yazdi	2017 (Jul.)	87,152	DNA
O	Shipman	2017 (Jul.)	31,024	DNA
P	Martens	2018 (Oct.)	1,523	oligomers
Q	Ratner	2009 (Dec.)	261	aromatics

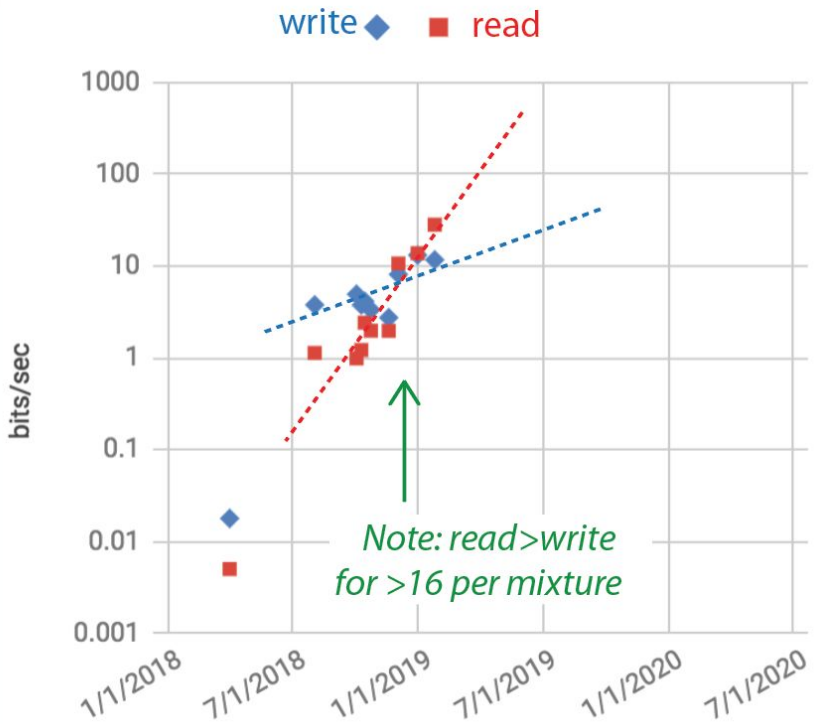
Arcadia *et al.*, *Nature Comm.* (2020).

MOLECULAR STORAGE CAPACITY GROWTH

Raw Chemical Information Storage
(without redundancy / error correction)



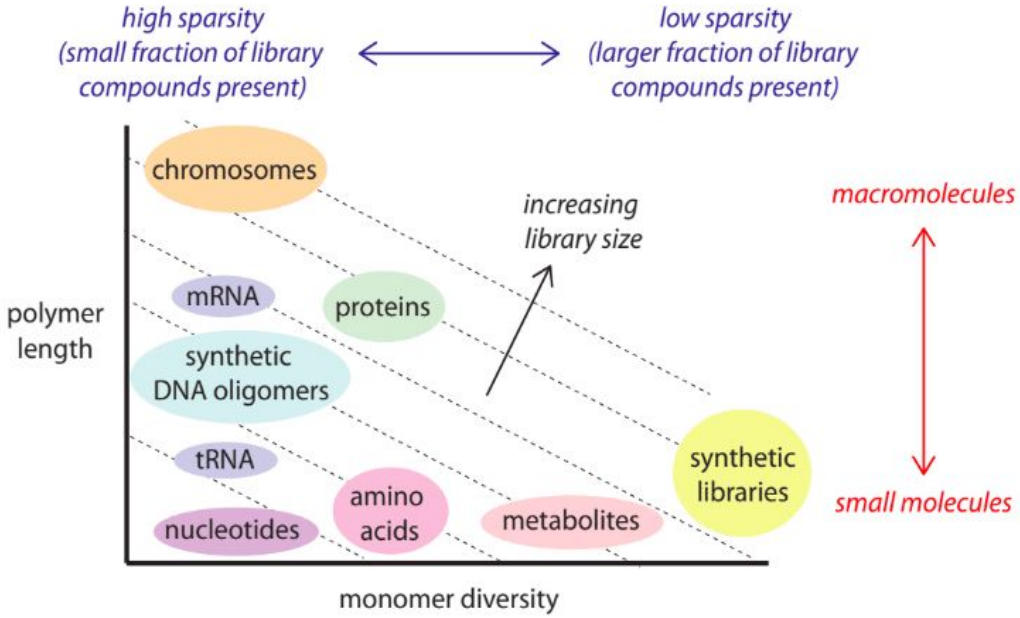
Read/Write Speeds



THE LARGER THEORETICAL PICTURE

BROWN

Storing information in disordered mixtures of small molecules is but one option out of a spectrum of potential storage techniques that exploit polymer length and monomer diversity in different ways.



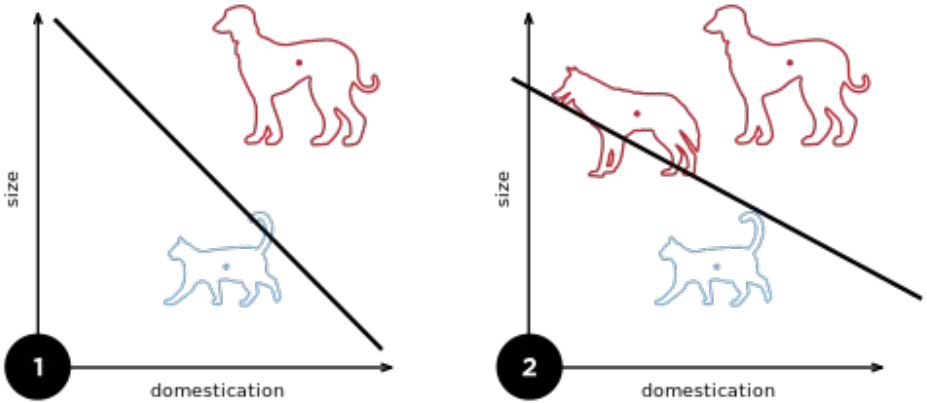
Rosenstein *et al.*, *Trans. NanoBioSciences* (2020).

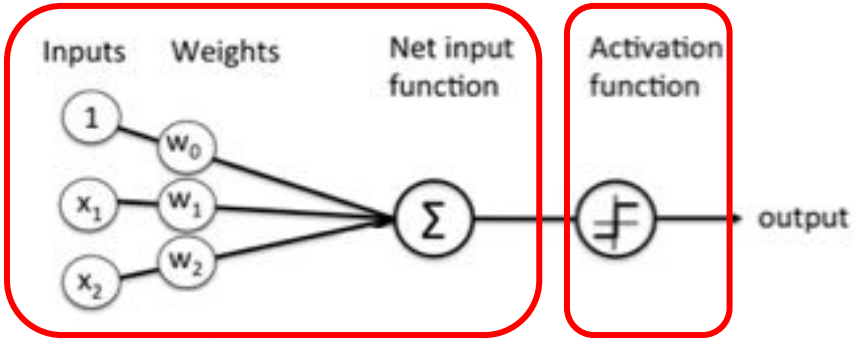
COMPUTATION

PERCEPTRONS

$$z = \sum_{i=1}^{N} w_i \cdot x_i + b$$

Perceptron Classifying Domesticated Animals





Multiply Accumulate

Nonlinear Thresholding

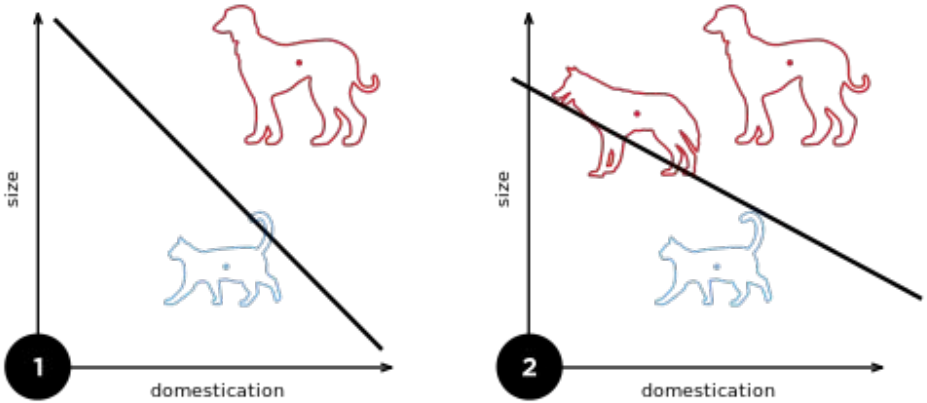
- **Goal:** To devise a molecular perceptron that can serve as the backbone for signal processing (e.g., filtering) and pattern recognition (e.g., image classification)

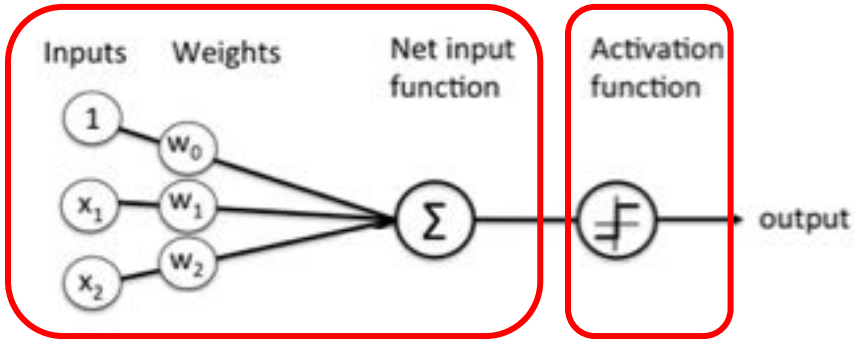
C. Arcadia *et al.*, ICRC (2018).

PERCEPTRONS

$$z = \sum_{i=1}^{N} w_i \cdot x_i + b$$

Perceptron Classifying Domesticated Animals





Multiply Accumulate

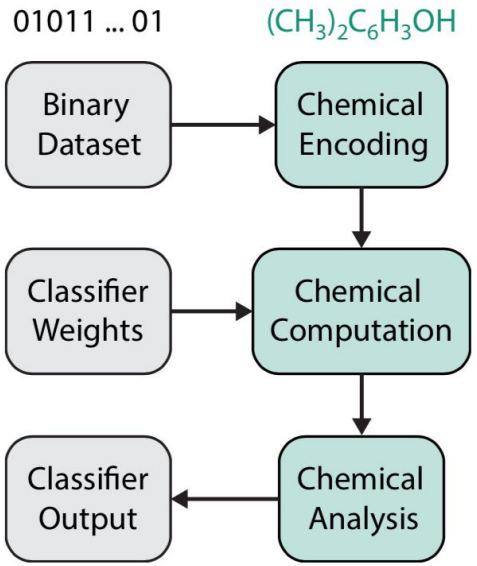
Nonlinear Thresholding

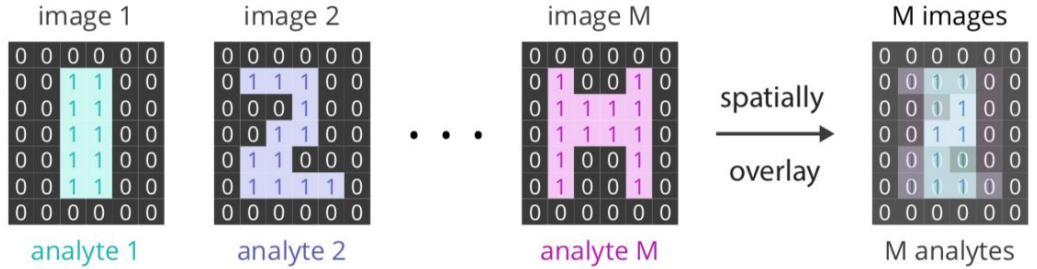
Chemical Multiply-Accumulate Operation

$$z = \sum_{i=1}^{N} w_i \cdot x_i + b \longrightarrow C_f = \sum_{i=1}^{N} \frac{V_i}{V_f} \cdot C_i$$

C. Arcadia *et al.*, ICRC (2018).

PARALLEL MOLECULAR COMPUTING





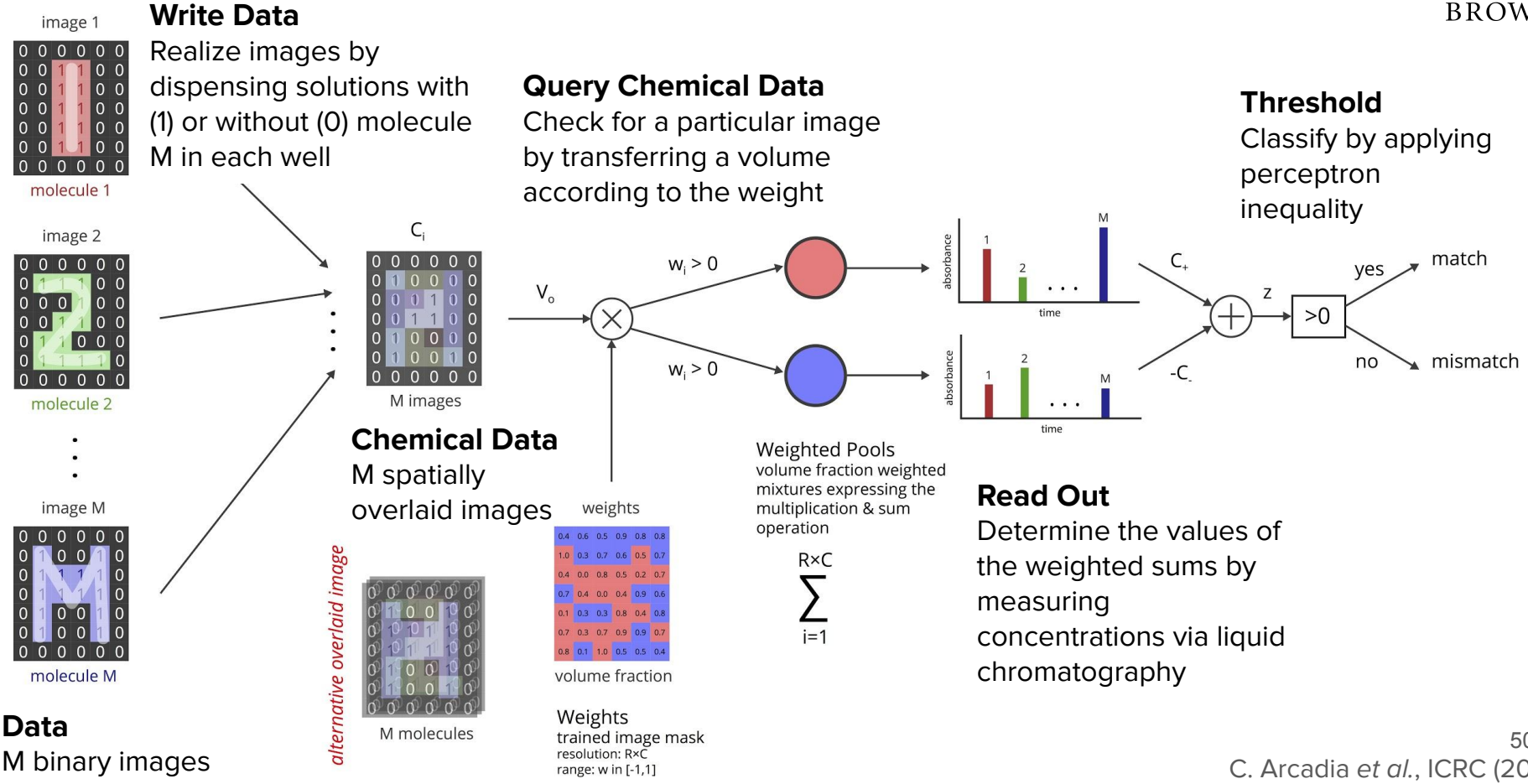
Here we take advantage of co-existing chemicals for parallelized image classification:

- Each pixel location maps to 1 well
- 1 image maps to 1 chemical
- Multiple datasets are overlaid (chemically mixed)

C. Arcadia *et al.*, ICRC (2018).

MOLECULAR PERCEPTRON CONCEPT

BROWN

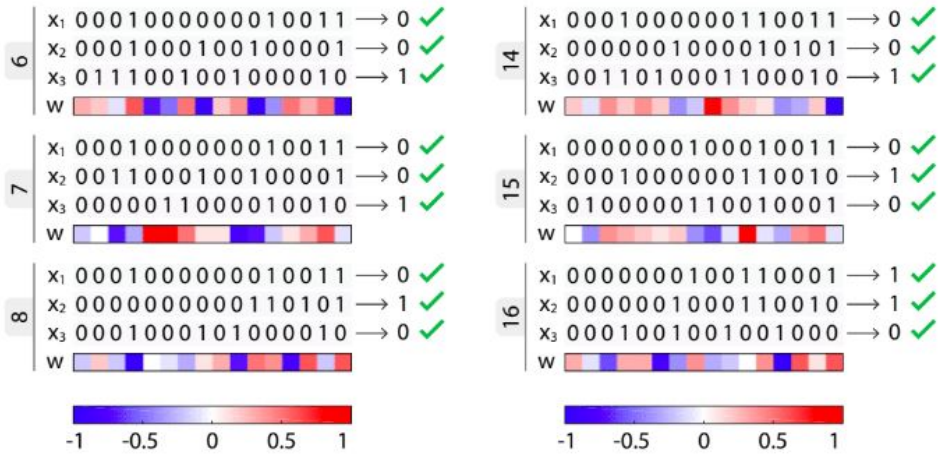


C. Arcadia *et al.*, ICRC (2018).

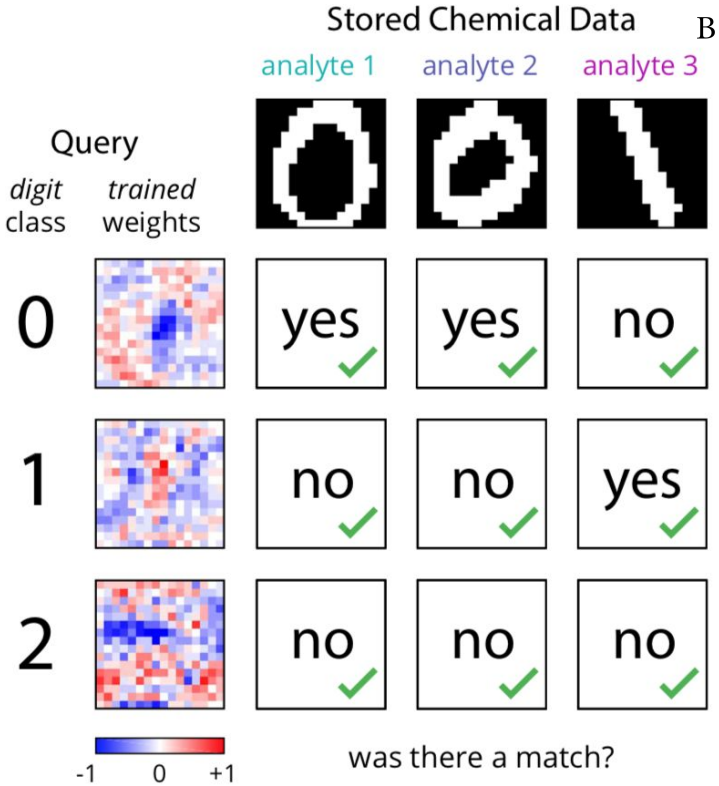
CHEMICAL MNIST DATA CLASSIFICATION

BROWN

16 sets of 3 vectors, each 16 bits
= 48 total vectors



96% Rate of Correct Classification



Used for classification here, but can easily also be applied to
signal processing, linear algebra

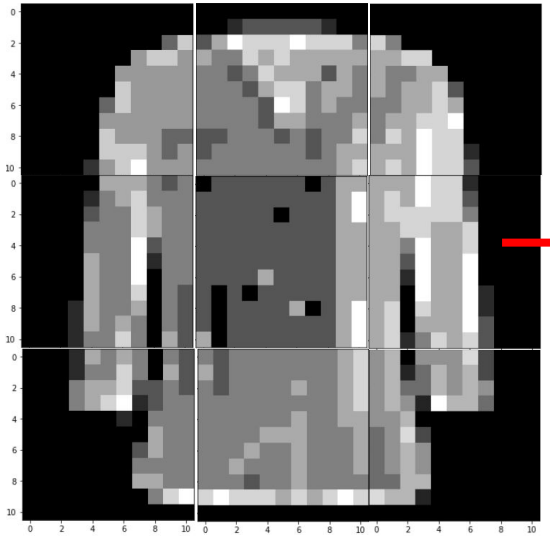
C. Arcadia *et al.*, ICRC (2018).

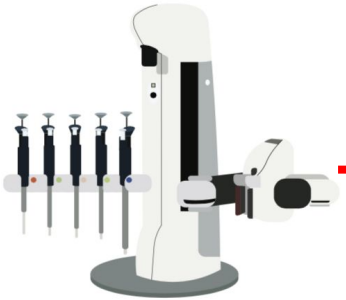
REPRESENTING IMAGES IN CHEMISTRY

BROWN

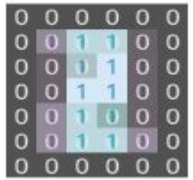
- Can Exploit Molecular Parallelism to Partition Images and Process at Once

Quantized Image After Partitioning and Padding





Andrew Alliance Liquid Handler



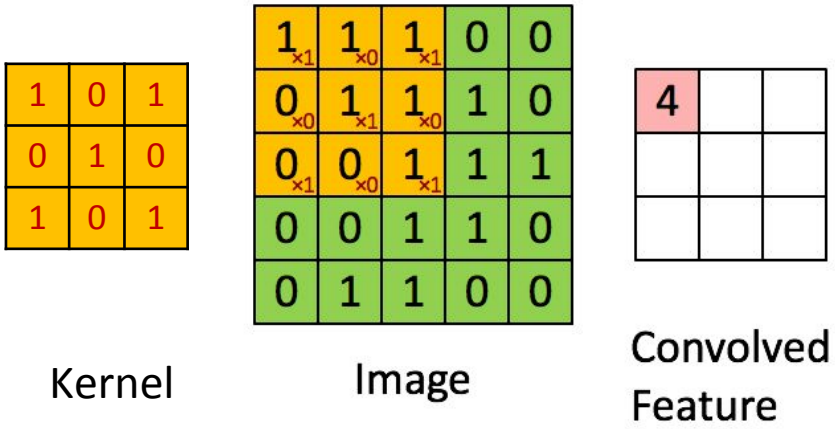
Well Plate

- Partition Image Into 9 Parts, Each Represented By a Unique Phenol
- All Partitions Are Superimposed → Enables Parallel Processing
- **27 Bits Per Well! → 9x Improvement over Earlier Perceptron Work**

FOUNDATIONS OF IMAGE PROCESSING

BROWN

- Convolutions Rest at the Heart of Most Image Processing Techniques
  - Weighted Sum of a Kernel Across an Image
- Requires Addition and Multiplication
  - Multiplication Via Volumetric or Reaction-Based Weighting

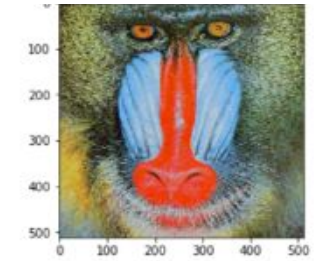


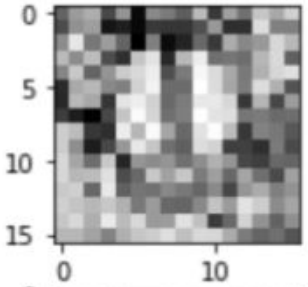
$$P_{ij} = \sum_{k=-R}^{R} \sum_{l=-R}^{R} P_{i+k\ j+l} K_{kl}$$

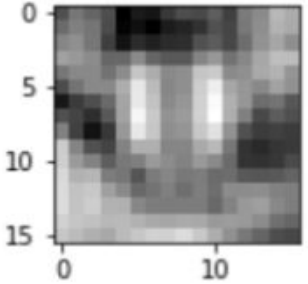
IMAGE PROCESSING: BLURRING EXAMPLE

BROWN

Image Blurring Example

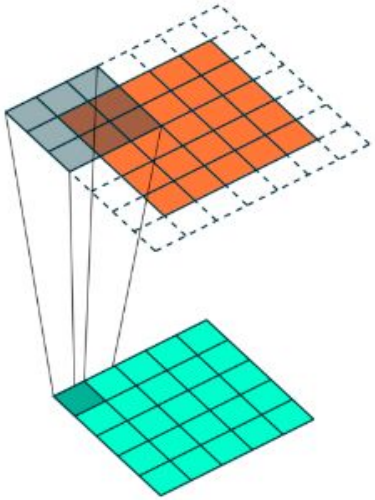






Apply a Kernel Over Sliding Windows of Pixels to Blur

-Attempt with 8x8 image in well plate



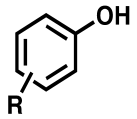
0	0	0	0	0	0	0
0	60	113	56	139	85	0
0	73	121	54	84	128	0
0	131	99	70	129	127	0
0	80	57	115	69	134	0
0	104	126	123	95	130	0
0	0	0	0	0	0	0

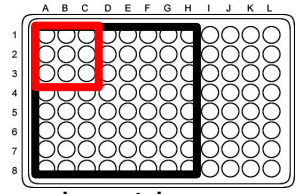
Kernel

0	-1	0
-1	5	-1
0	-1	0

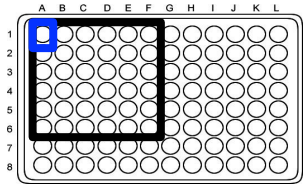
114				

Molecular Implementation Using Phenols and HPLC

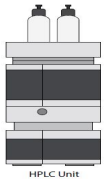




input image



Output image



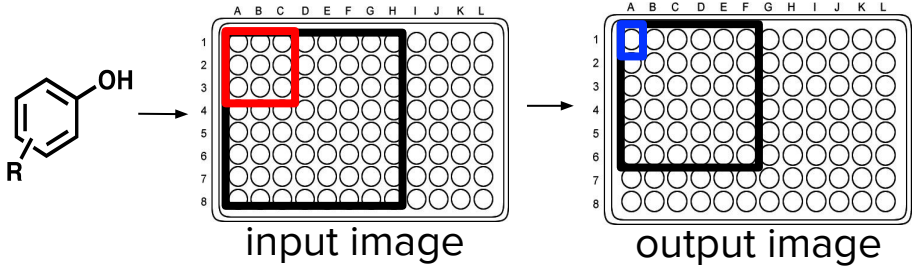
HPLC

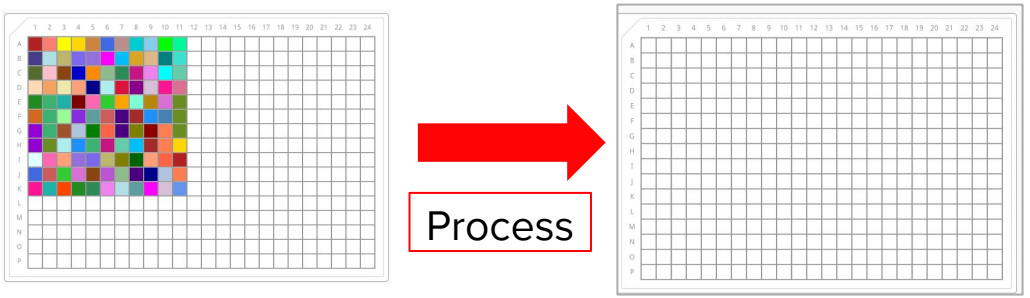
MULTIPLICATION VIA VOLUMETRIC WEIGHTING

BROWN

Kernel :
Gaussian Blur

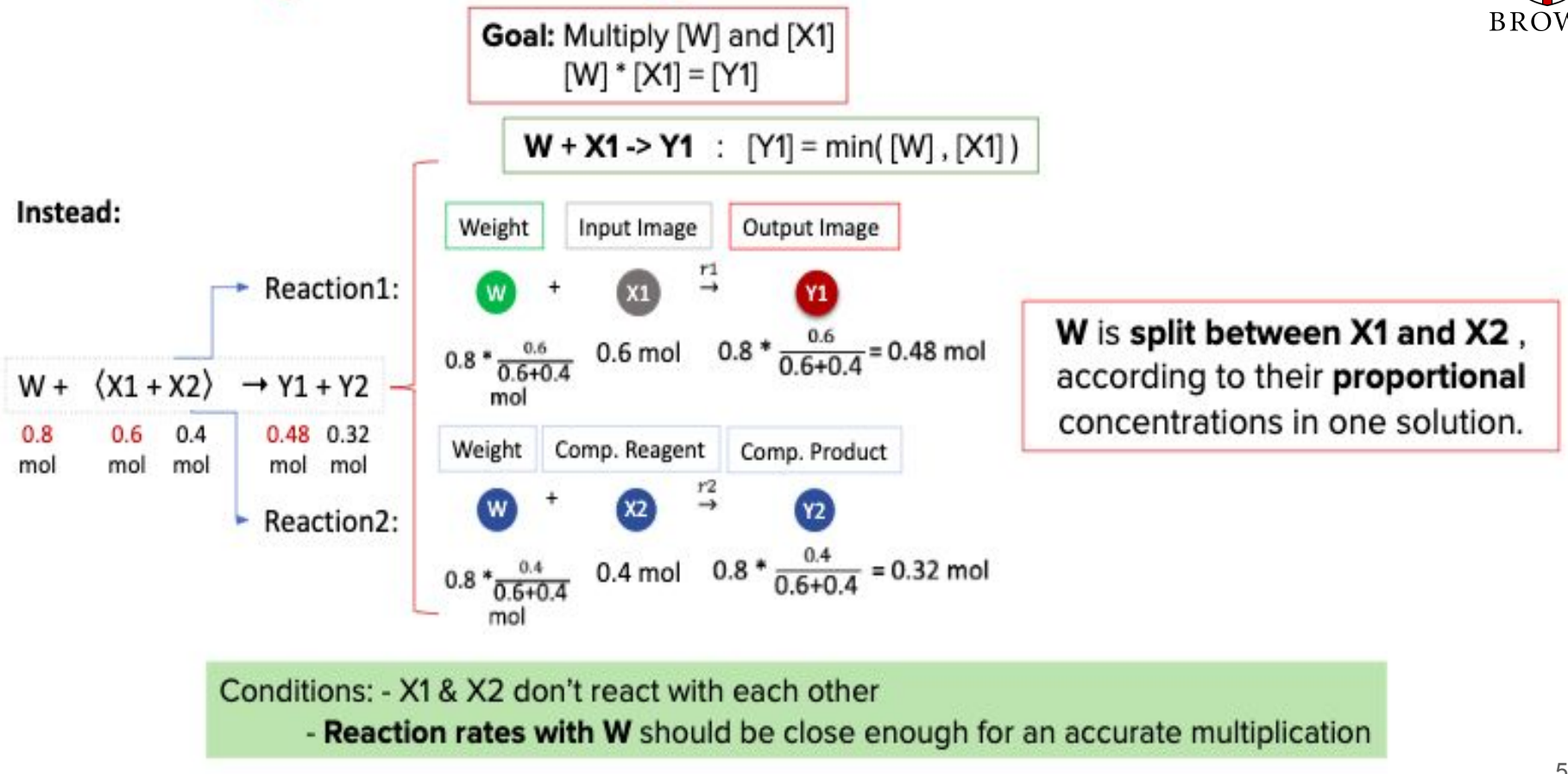
0.0673	0.1248	0.0673
0.1248	0.2314	0.1248
0.0673	0.1248	0.0673





Weights = Volumetric Fractions Of Input Image

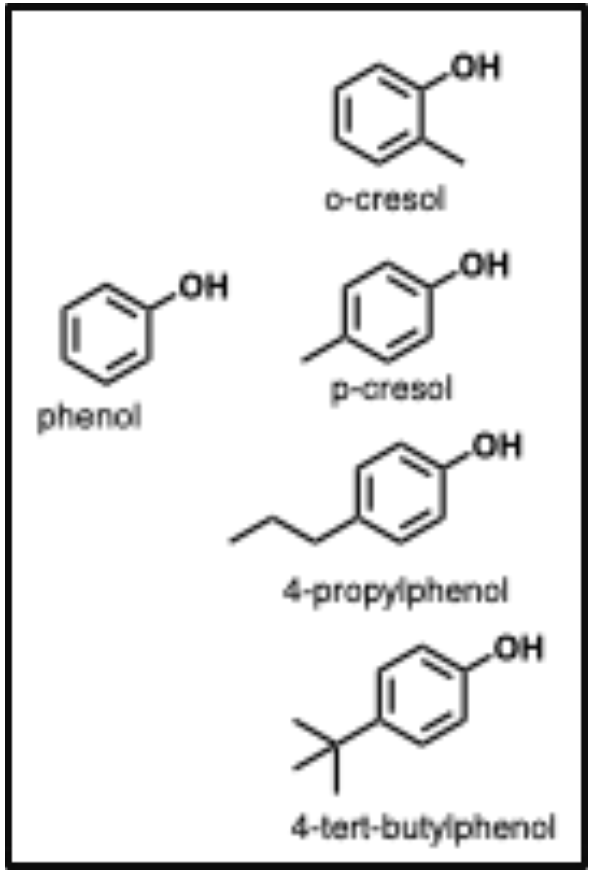
MULTIPLICATION VIA REACTIONS



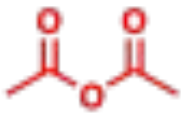
MULTIPLICATION VIA REACTIONS

BROWN

Input Image



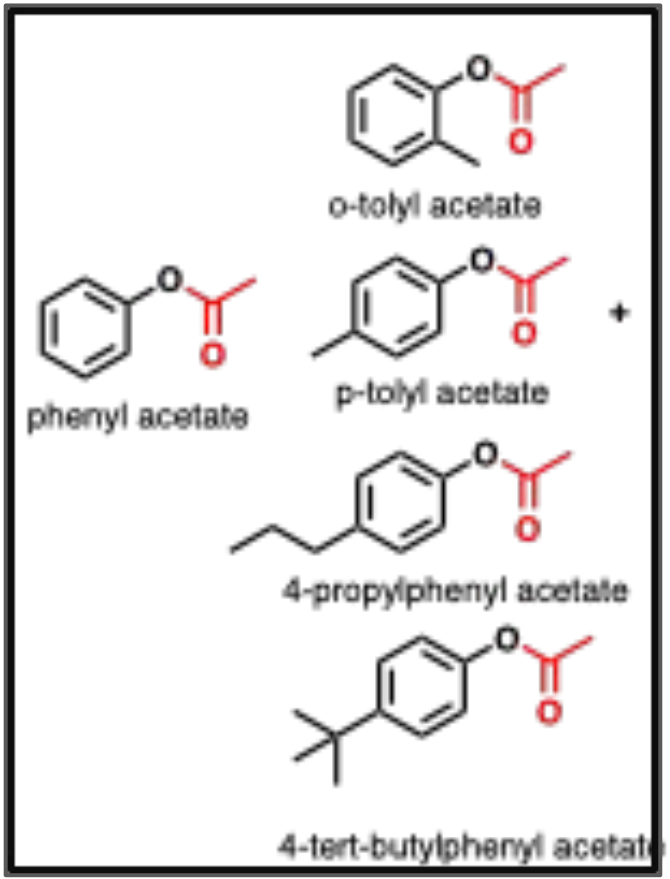
Kernel



EtCN
24hrs, RT

3 different concentrations represent 3 different values of the Gaussian kernel

Output Image

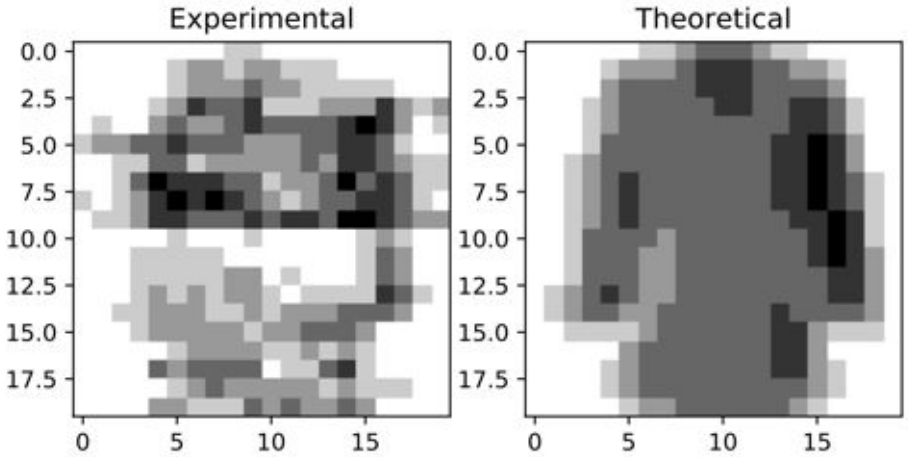


Look for [acetates] in GC-MS

FILTERING IMAGES

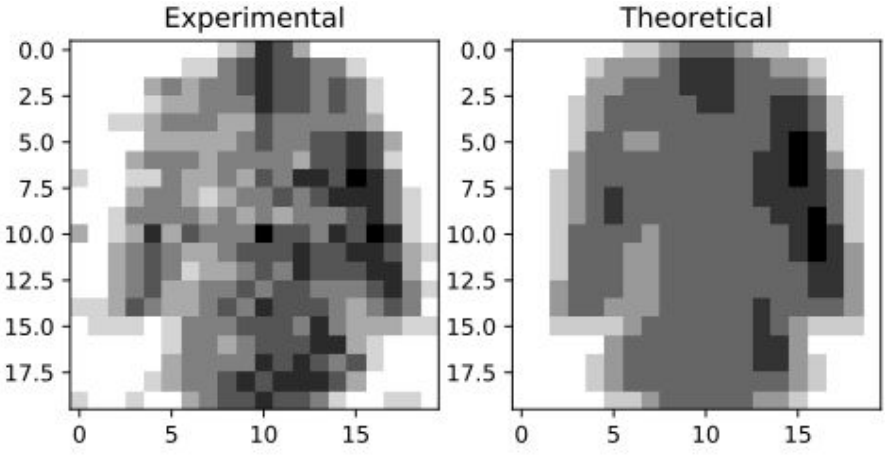
BROWN

Quantized Image Before Error Correction



SSIM = 0.3654

Quantized Image After Error Correction



SSIM = 0.6103

=> With appropriate EC, can chemically process images.

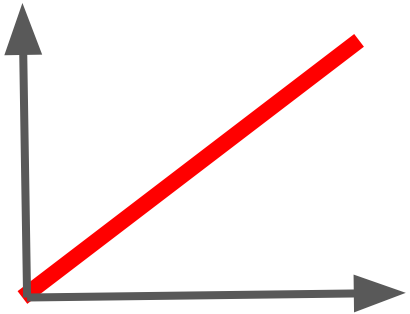
The Structural Similarity Index (SSIM) compares experimental and theoretical blurred images (0 is least similar and 1 is most similar).

PATH TO CHEMICAL, MULTILAYER NETWORKS

BROWN

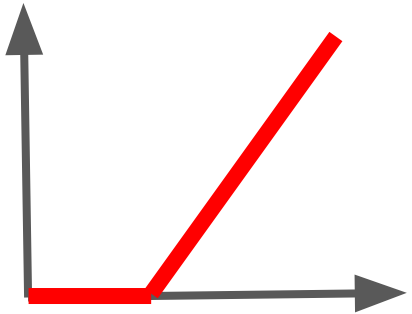
Developing a multilayer neural network requires:

- Activation - a nonlinear function that suppresses some differences and amplifies others
- Properties that can be transferred from layer to layer (can't be abstract)
- The ability to multiply and accumulate over many inputs

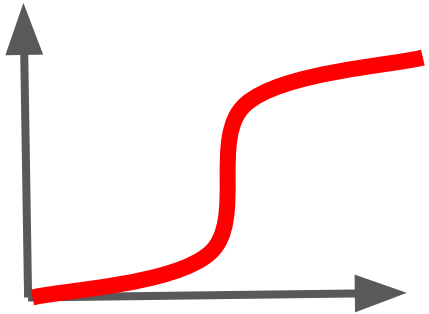


Linear Function

Less Powerful for Classification



ReLU



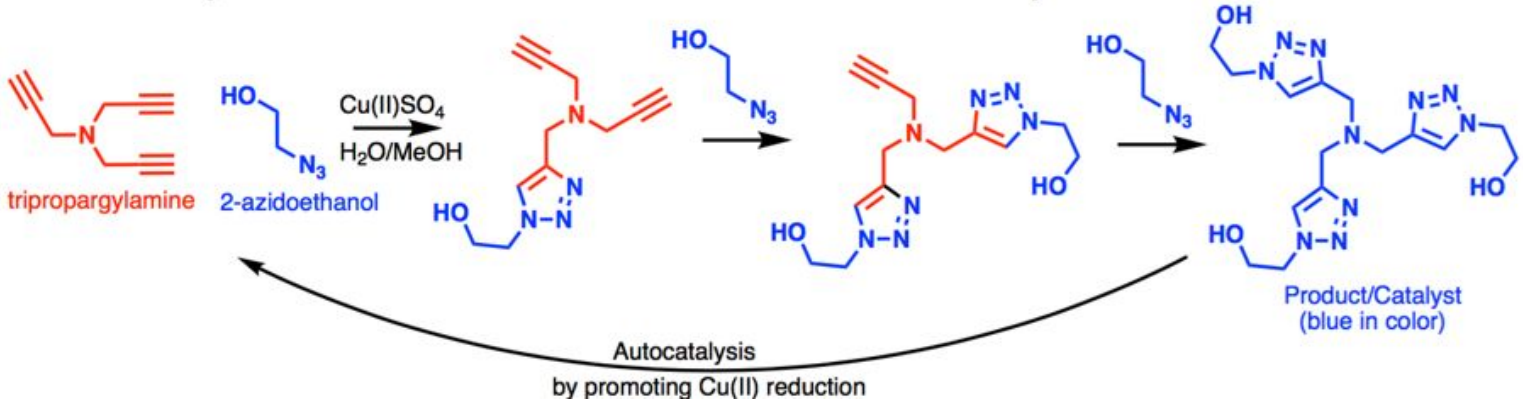
Sigmoid

More powerful for classification

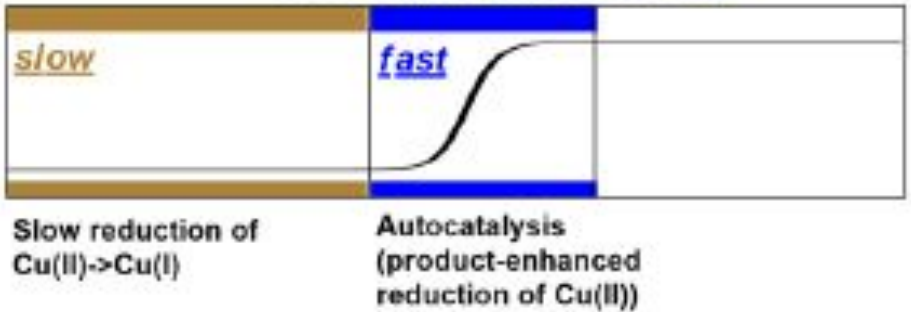
ACTIVATION VIA AUTOCATALYSIS

BROWN

The Copper-Catalyzed Azide-Alkyne Cycloaddition Reaction



A Way To Measure Time!

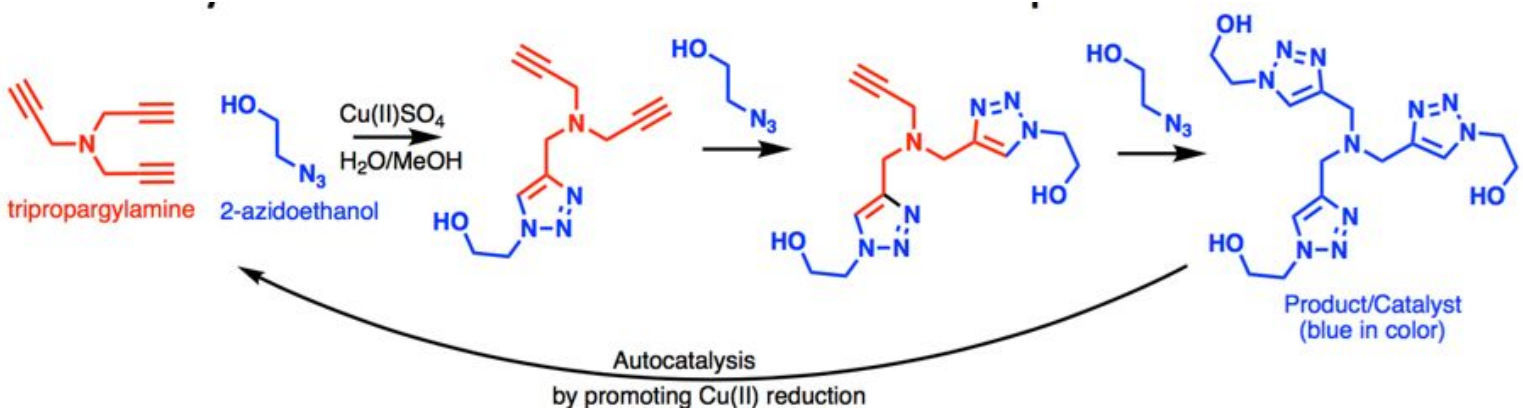


S. Semenov *et al.*, JACS (2018); C. Arcadia *et al.*, In Prep (2020).

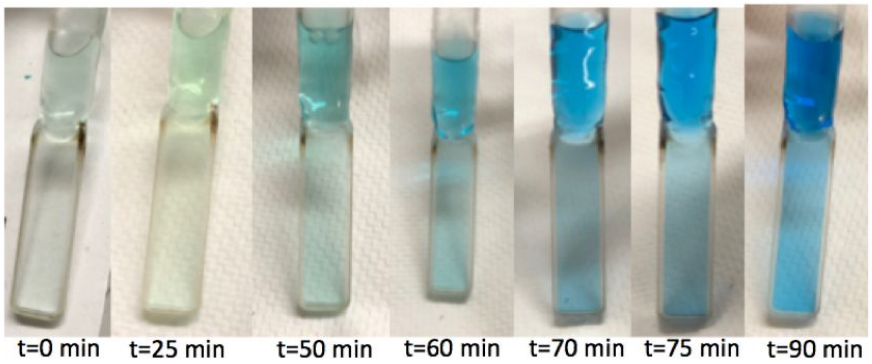
ACTIVATION VIA AUTOCATALYSIS

BROWN

The Copper-Catalyzed Azide-Alkyne Cycloaddition Reaction



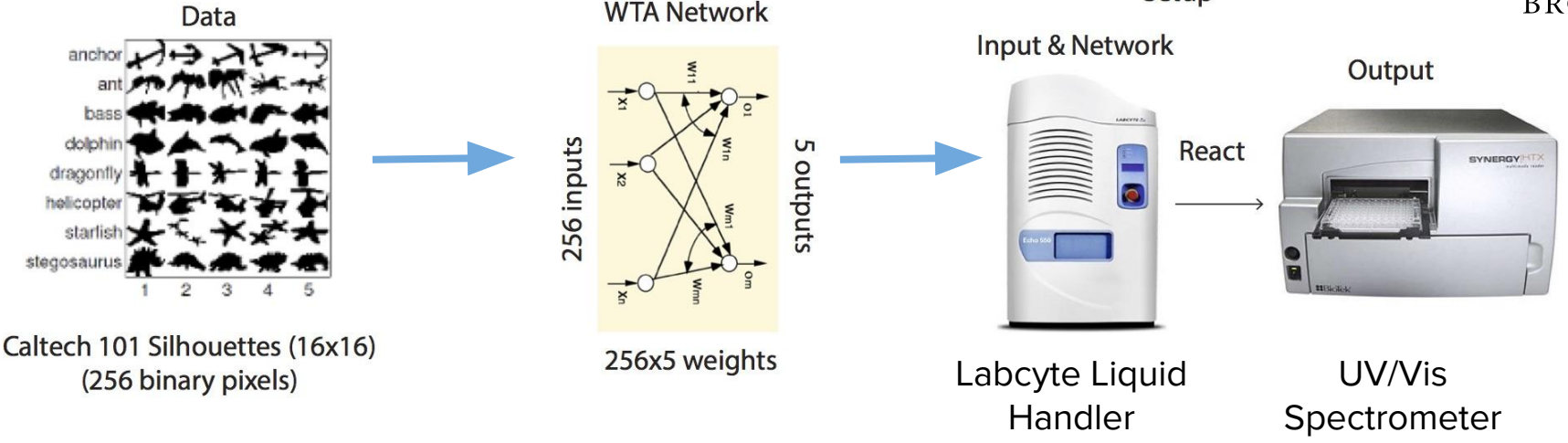
A Way To Measure Time!



Minimally absorbs in the 400 nm range ➔ blue

S. Semenov *et al.*, JACS (2018); C. Arcadia *et al.*, In Prep (2020).

WINNER-TAKE-ALL-NETWORK FLOW

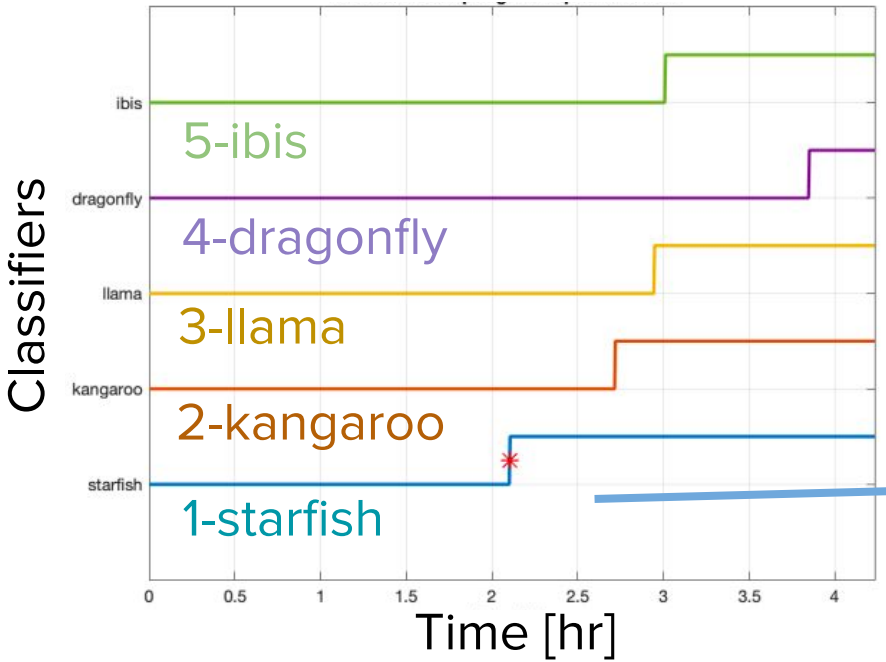


S. Semenov *et al.*, JACS (2018); C. Arcadia *et al.*, In Prep (2020).

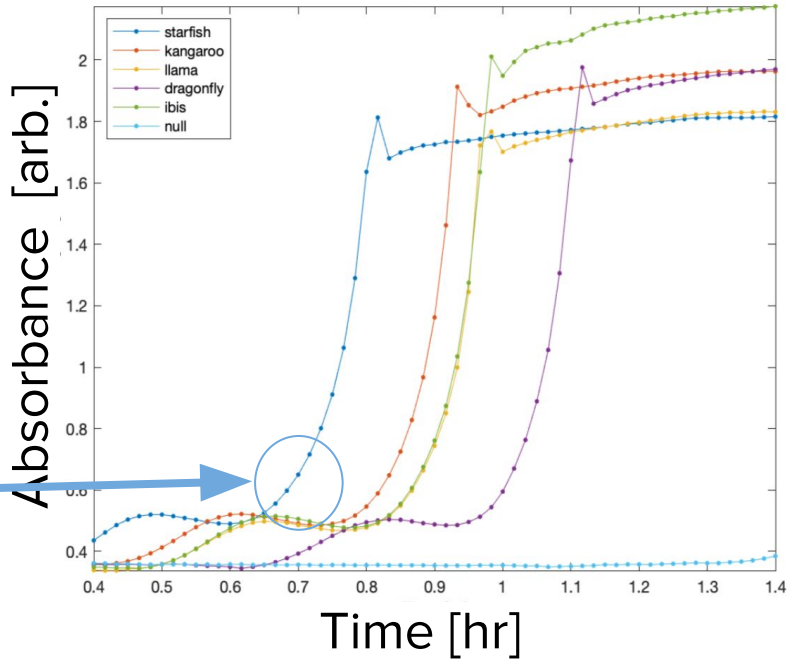
WINNER-TAKE-ALL NETWORK

BROWN

Simulated



Experimental

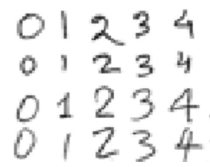


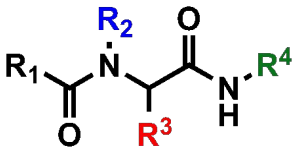
1-starfish, 2-kangaroo, 3-llama, 4-dragonfly, 5-ibis

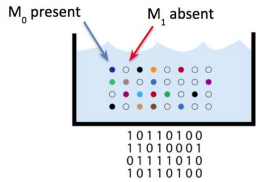
OUTLOOK

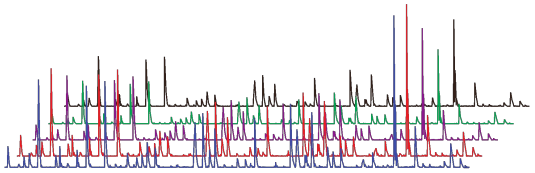
SUMMARY/OPPORTUNITIES

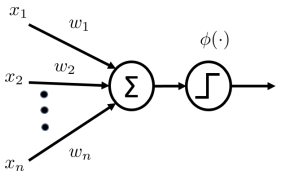
- Developed new ways to store information in molecules which rely on mass - extremely dense, discreet, and stable
- Opens up possibilities for extremely parallel computation at different scalings than conventional computation; this was exploited in our molecular neural networks
- Ultimately, molecules can be used for more complex computing tasks than classification - but how?

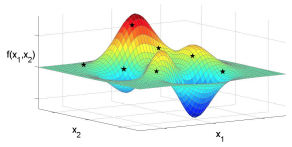






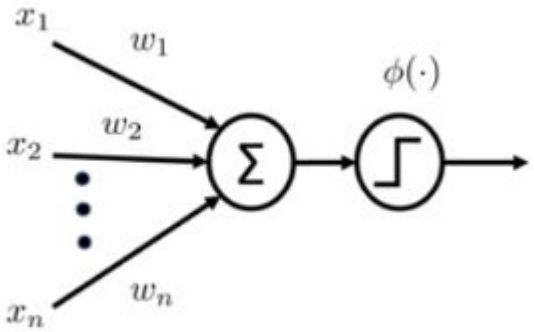




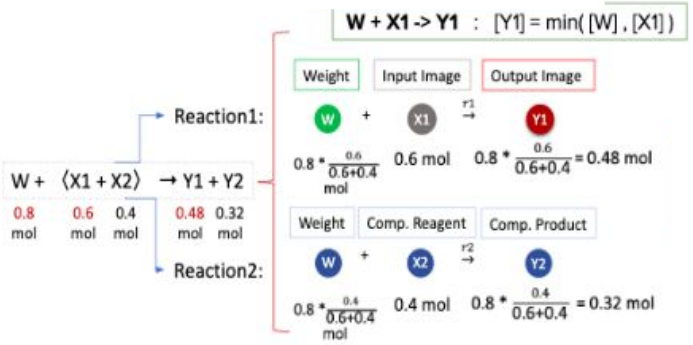


MOTIVATION

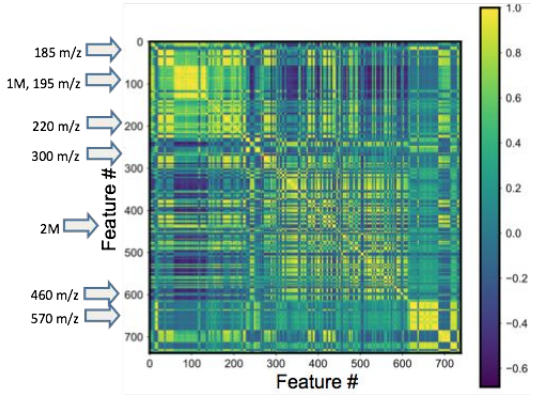
Thus far We Have Guessed “Functional” Reactions



Volumetric and Autocatalytic Perceptrons



Chemical Multiplication



Identifying Logic Gates in Metabolism

Key Goal:

Let’s Not Guess!

Let’s Develop a Systematic Way of Identifying “Functional” Reactions We Can Realize in the Lab

ACKNOWLEDGEMENTS

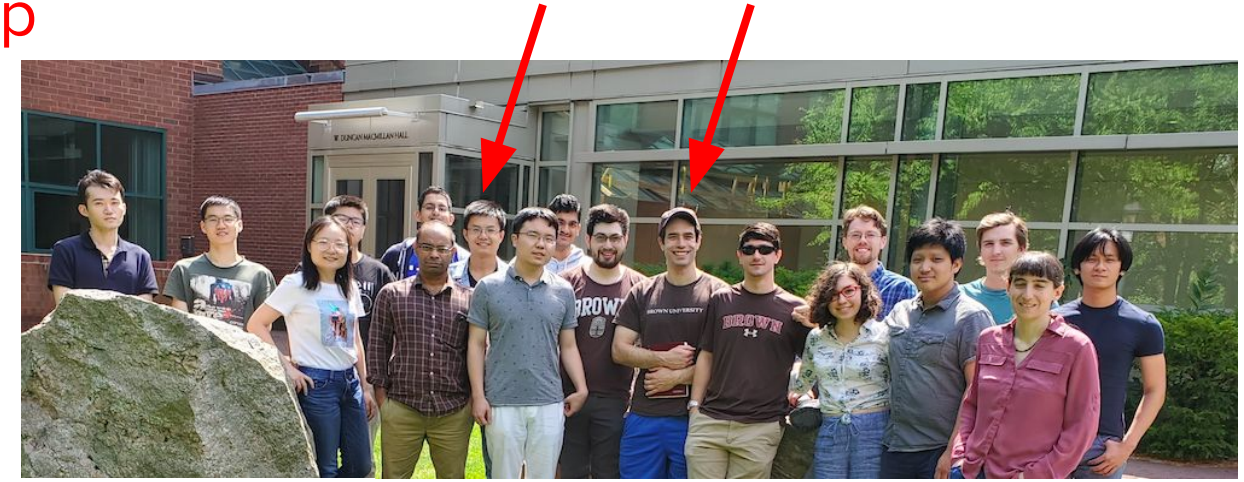
- **Prof. Eunsuk Kim (Chem.)**
- **Prof. Sherief Reda (Eng.)**
- **Prof. Chris Rose (Eng.)**
- **Prof. Jacob Rosenstein (Eng.)**
- **Prof. Brenda Rubenstein (Chem.)**
- **Prof. Jason Sello (Chem.)**
- **Prof. Peter Weber (Chem.)**
- Mr. Chris Arcadia (Eng.)
- Ms. Shui-Ling Chen (Chem.)
- Ms. Amanda Dombrowski (Bio.)
- Ms. Kady Ferguson (Chem.)
- Dr. Joseph Geiser (Chem.)
- Dr. Eamonn Kennedy (Eng.)
- Ms. Asami Odate (Chem.)
- Mr. Aaron Ray (Eng.)
- Mr. Hokchhay Tann (Eng.)

BROWN

DARPA

GROUP ACKNOWLEDGEMENTS

The Rubenstein Group



Rubenstein Group, Summer 2019

Electronic Structure Team

- Dr. Matthew Church
- Mr. Cancan Huang
- Mr. Gopal Iyer
- Dr. Yuan Liu
- Dr. Ravindra Nanguneri
- Mr. Tong Shen

Biophysics Team

- Mr. Haobo Yang
- Mr. Gabriel M. da Silva

Molecular Computing Team

- Mr. Benjamin Foulon
- Mr. Leonard Sprague

RESEARCH CORPORATION for SCIENCE ADVANCEMENT

DARPA

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

NSF

ALFRED P. SLOAN FOUNDATION 1934