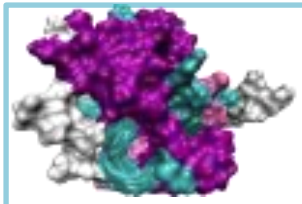
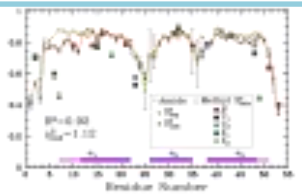
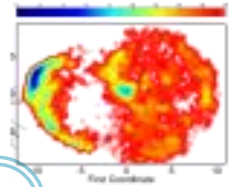
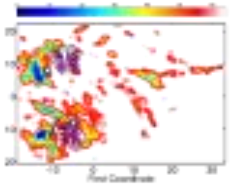
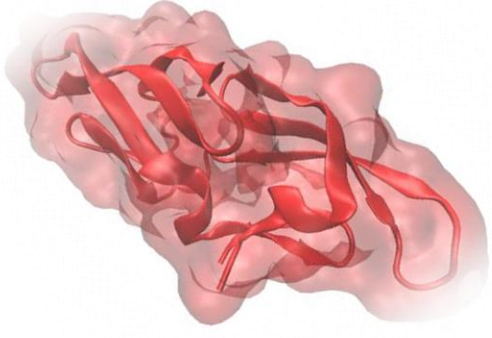



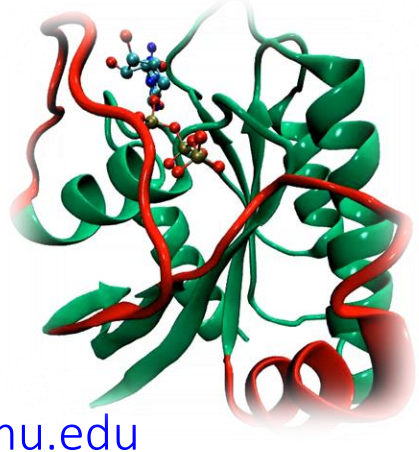




AI/ML-Driven Scientific Advances: A Personal Journey, Lessons, and Outlook

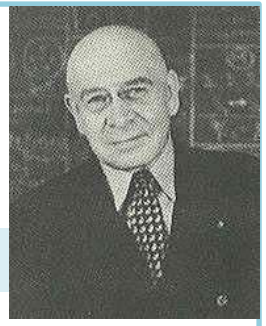


Amarda Shehu
Department of Computer Science
George Mason University
Fairfax, Virginia



http://cs.gmu.edu/~ashehu amarda@gmu.edu

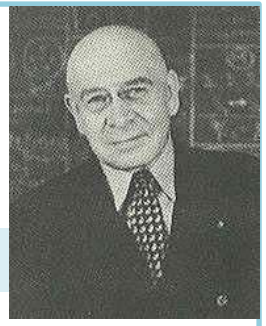
The Map is not the Territory



Alfred Korzibsky
1879-1950

- Developed field of “general semantics”
- Thought deeply about connection between human knowledge and language, and the observed reality versus the observer in defining human knowledge
- Argued that no one can have direct access to reality – knowledge is filtered through the brain's responses to reality
- Best known dictum: "The map is not the territory"

The Map is not the Territory



Alfred Korzibsky
1879-1950

Science and Sanity (1933)

Greek, pre-scientific (idealism)
Observer-centric; observed reality does not matter

Classical scientific (materialistic)
Observed reality-centric; observer does not matter

Modern scientific (pre-digital) [no to small data]
Human knowledge depends on *both* the observed reality and the observer

Small - to medium-size data (shallow Machine Learning) … - 2010

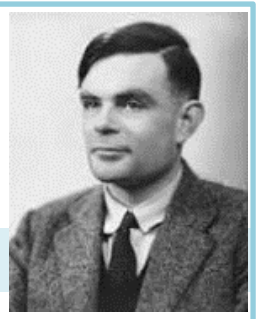
Modern digital (data-centric) [big data → Largely Deep Learning]
Human knowledge entirely derived from the data
Observed reality and the observer do not matter
2010 - …

?

Model-based Research
"Classical AI"
Explicit Knowledge

Data Analytic Modeling
"Machine Learning"
Tacit Knowledge

Start of Personal Journey



Sir Alan Turing
1912-1954

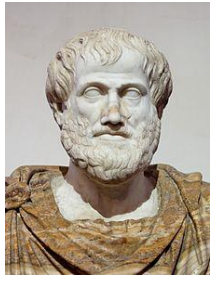
*"The purpose [..] is to discuss a possible mechanism by which [..] genes [..] **may determine the anatomical structure** of the resulting organism. The theory [..] suggests that [..] **well-known physical laws are sufficient** to account for many of the facts."*
Turing, AM. (1952) Chemical basis of morphogenesis. *Phil Trans Royal Soc London. Series B, Biol Sciences* 237(641):37-72.

How does the form of matter determine function?

↓

Guiding Principle

Societal & Health Problems → Outstanding Scientific Challenges → Fundamental AI Advances



"It is the mark of an instructed mind to rest satisfied with the degree of precision which the nature of the subjects permits and not seek an exactness where only an approximation of the truth is possible." Aristotle 319 BC

↓

Computational Focus

Build or Learn
Function-encoding Representation of Form

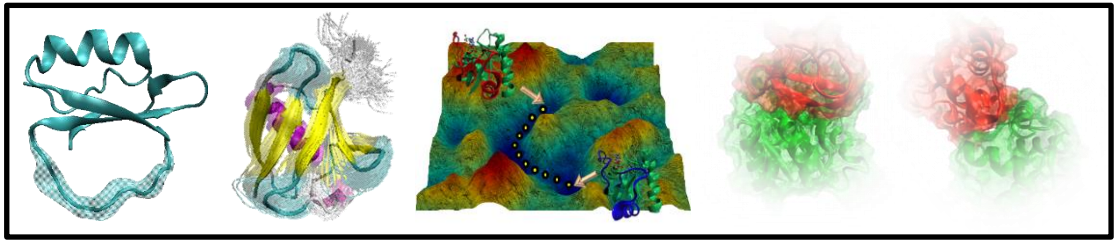
Personal Journey

No to very little data – explicit knowledge

2002-2016

Classic AI: stochastic search- optimization (geometry, kinematics, inverse kinematics, motion planning)

-- molecular structural biology

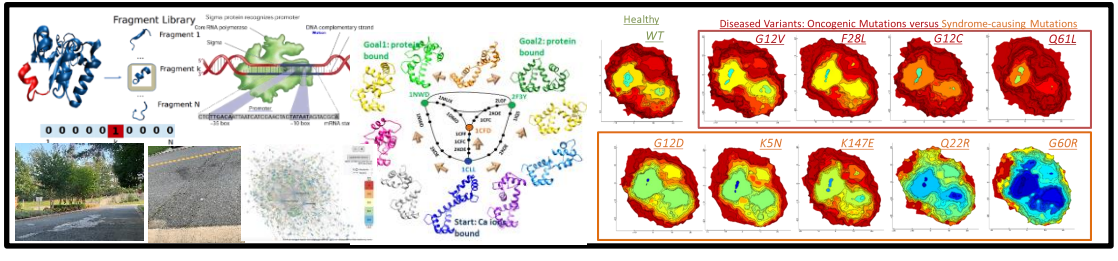


Some data –explicit and tacit knowledge

2010-

Hybrid Models (data-driven AI, knowledge-guided shallow ML, shallow ML + AI)

-- sequence/structural biology, social media user modeling, industrial monitoring, urban planning

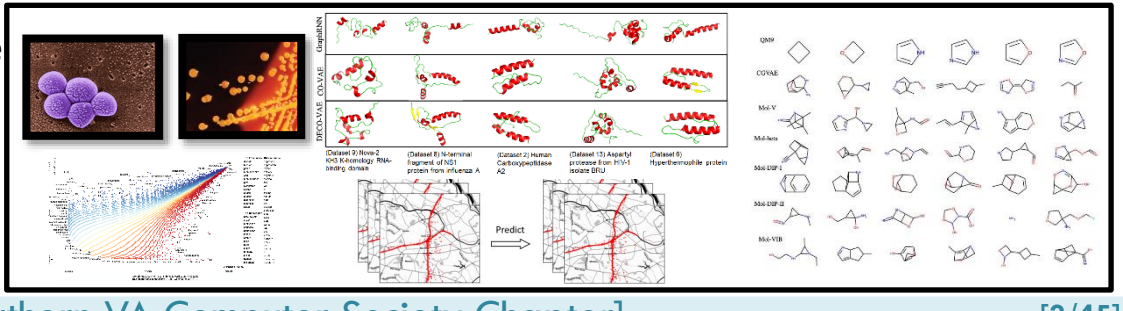


Lots of data – AI romance w/ tacit knowledge

2018-

Deep Learning, NLP, Deep generative models

-- sequence/structural biology, mental health, traffic forecasting, AI for Policy



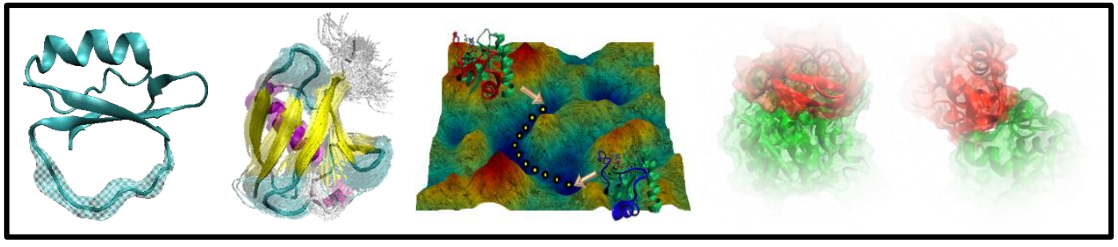
Personal Journey

No to very little data – explicit knowledge

2002-2016

Classic AI: stochastic search- optimization (geometry, kinematics, inverse kinematics, motion planning)

-- molecular structural biology



Some data –explicit and tacit knowledge

2010-

Hybrid Models (data-driven AI, knowledge-guided shallow ML, shallow ML + AI)

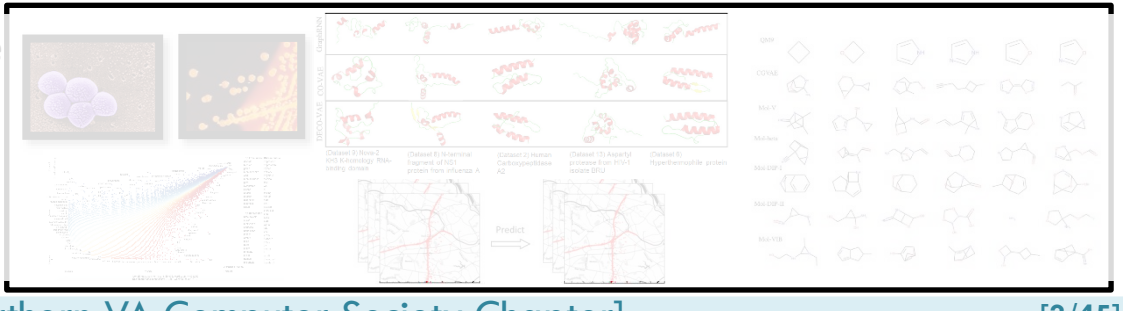
-- sequence/structural biology, social media user modeling, industrial monitoring, urban planning

Lots of data – AI romance w/ tacit knowledge

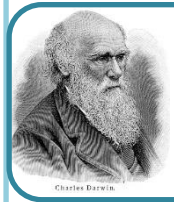
2018-

Deep Learning, NLP, Deep generative models

-- sequence/structural biology, mental health, traffic forecasting, AI for Policy



Chapter I – AI for Dynamics of Complex Systems
Explicit Knowledge at Full Force



On the Origin of Species

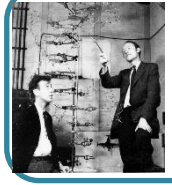
Charles Darwin

1859

1869

Discovery of DNA

Friedrich Miescher



Double helical structure of DNA

Watson & Crick

1953

1950s

Protein Folding and the Thermodynamic

Hypothesis, 1950-1962

Anfinsen

Perutz & Kendrew

Crystal structure of hemoglobin

1957 (6Å) & 1959 (2Å)

1957

1970s

© Nobel Media AB. Photo: A. Mahmoud
Martin Karplus
Prize share: 1/3

© Nobel Media AB. Photo: A. Mahmoud
Michael Levitt
Prize share: 1/3

© Nobel Media AB. Photo: A. Mahmoud
Arieh Warshel
Prize share: 1/3

MD simulation of protein structure and dynamics

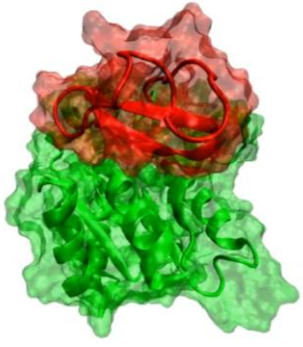
Karplus, McCammon, Levitt, Warshel, Scheraga

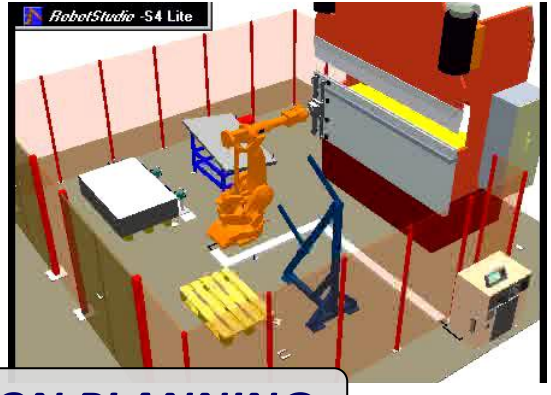
Golden Age

- C++ problems
- Hybridization of sub-domains, domains, and disciplines

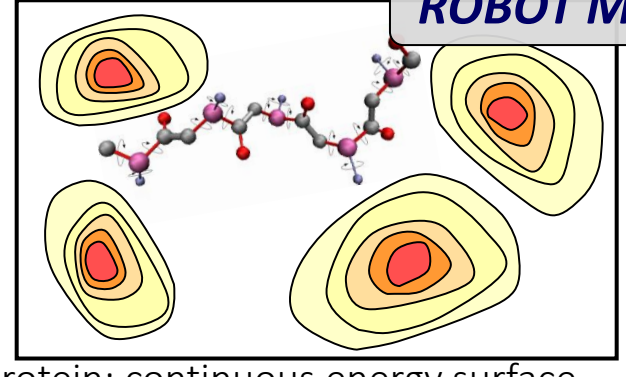
2004-2016

Robot Motions and Protein Motions: Leverage Analogies

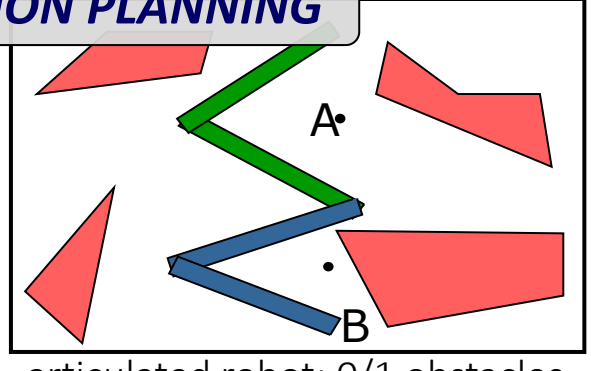




ROBOT MOTION PLANNING



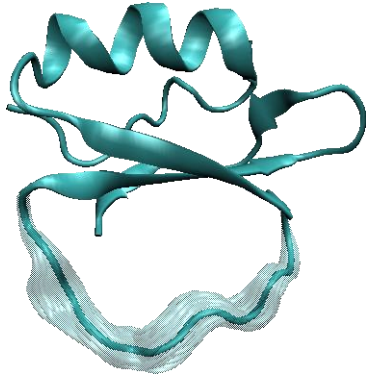
protein: continuous energy surface



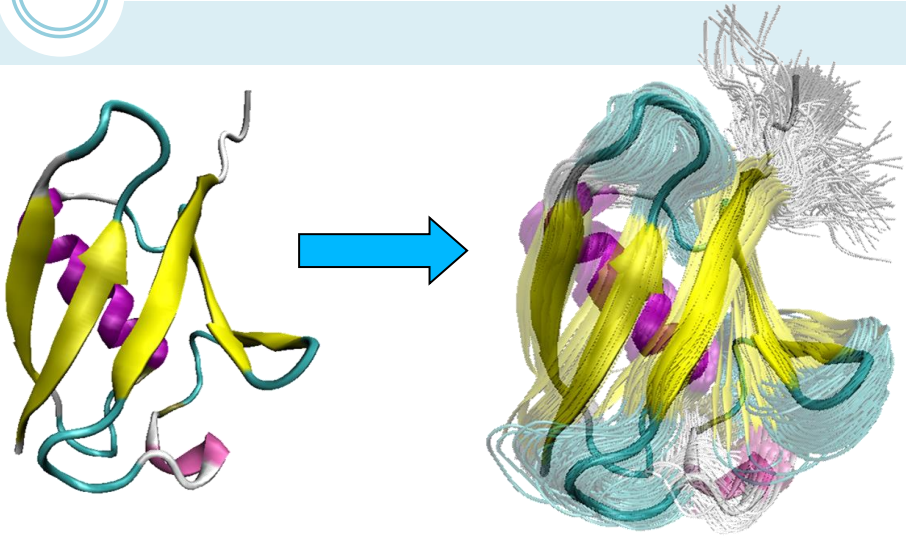
articulated robot: 0/1 obstacles

Prior to the Data Revolution

[It was the best of times. It was the worst of times]



Modeling of equilibrium flexibility of specific, highly-mobile segments

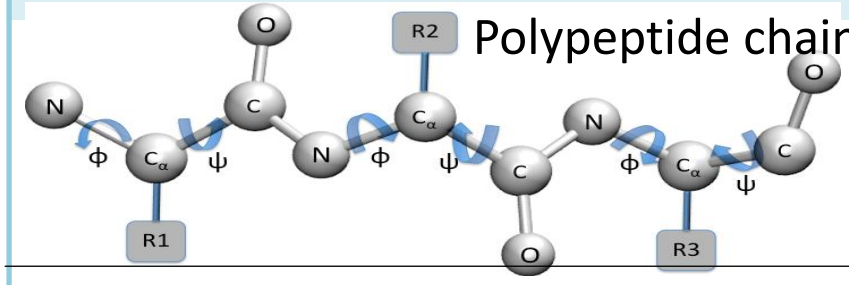


Modeling of equilibrium flexibility of entire protein chain

Goal: Partial or full characterization of protein flexibility by combining fast molecular kinematics (inspired from robotics/geometry of articulated objects) with physics-based treatments (molecular mechanics)

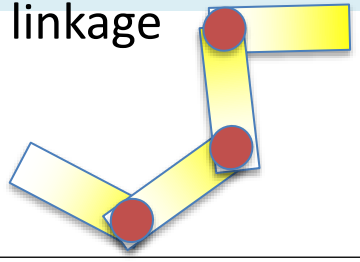
Context: Rich but incomplete knowledge from computer science, biophysics, chemistry, statistical mechanics → ripe for ingenuity, model/algorithmic design and novelty

Tree-based and Roadmap-based Algorithms Inspired from Robot Motion Planning



Polypeptide chain

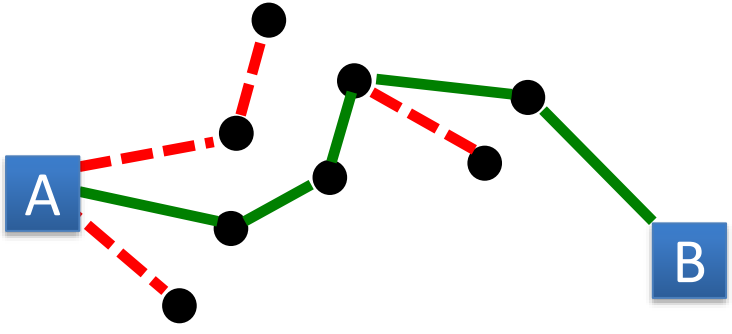
Articulated linkage



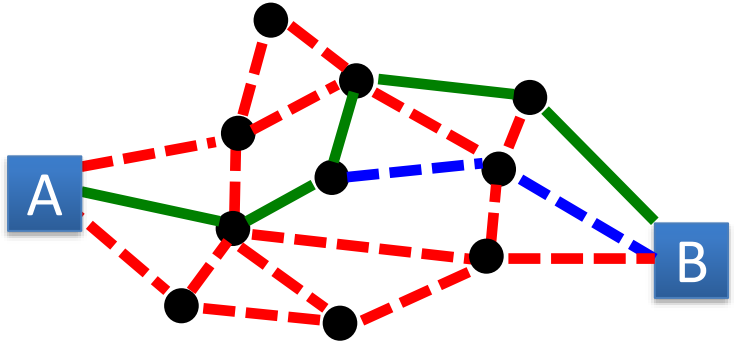
Tree-based

Roadmap-based

Grow tree in state space
(local view of state space)

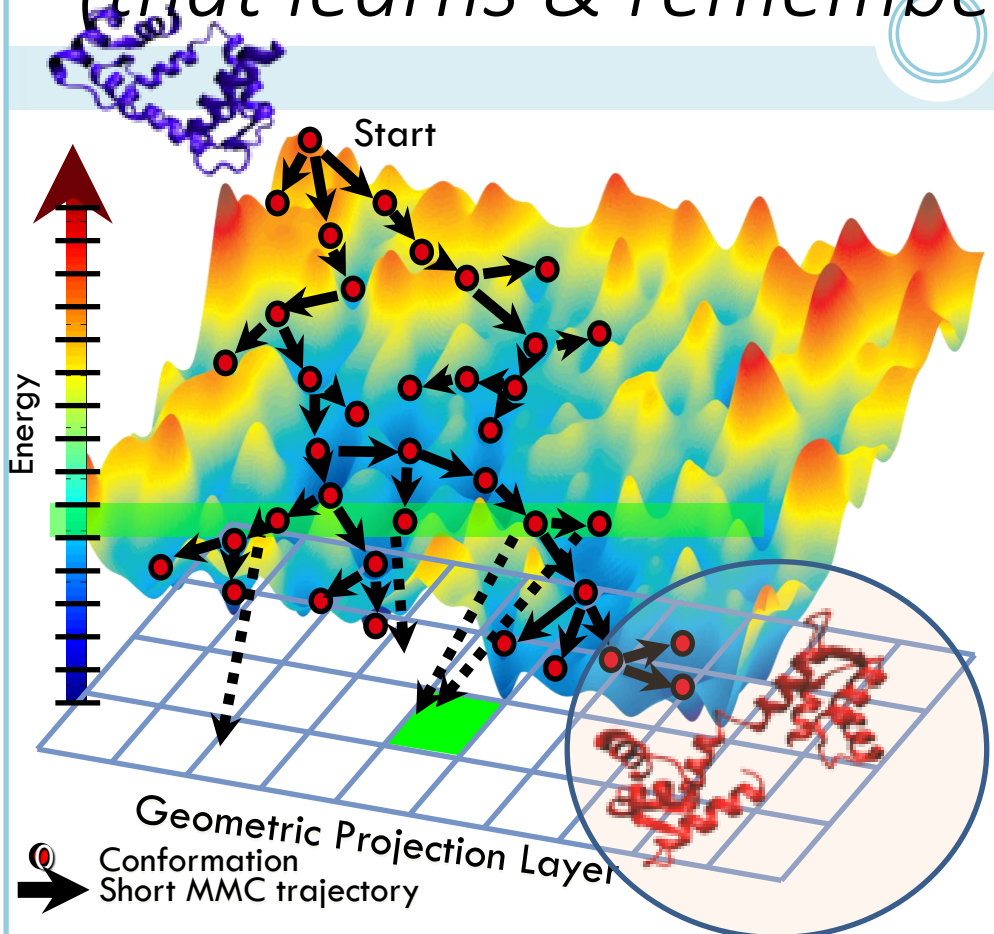


Build roadmap in state space
(non-local view of state space)



Adaptive Tree-based Search

(that learns & remembers where it has been)



Key idea:

Smart use of discretizations/structurization of search space and energy/cost surface to *adaptively* steer search tree towards constraint-satisfying regions

In structure modeling:

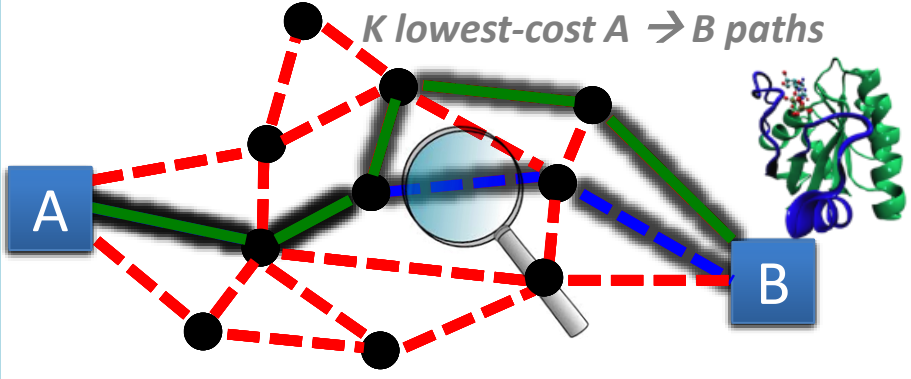
Low-energy AND geometrically-diverse conformations

→ projection layers over energy surface and conformation space

Molloy and Shehu. Elucidating the Ensemble of Functionally-relevant Transitions in Protein Systems with a Robotics-inspired Method. BMC Structural Biology J 13(Suppl1):S8, 2013.

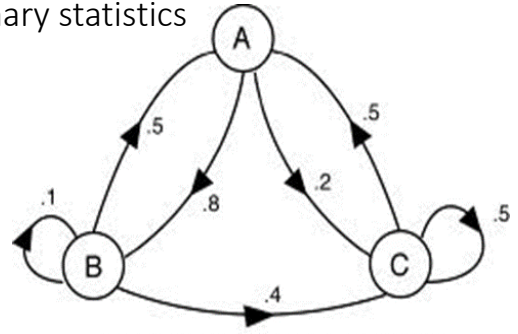
Roadmap-based Algorithms for Hundred-Dimensional (Biomolecular) State Spaces

Roadmap to obtain ensemble of K lowest-cost A → B paths



H-Ras	Transition	Exp Nr. Of Edges
WT	Off → On	3.4×10^{8}
	On → Off	3.9×10^{10}
Q61L	Off → On	1.9×10^{12}
	On → Off	3.8×10^{14}

- Markov State Models (MSMs) as discrete kinetics models that additionally permit calculation of summary statistics



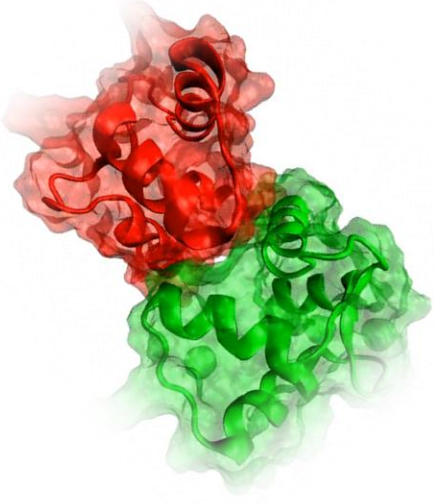
- Expected nr. of edges from a vertex v_i to any v_j in A $\quad t_i = 1 + \sum_{v_{j \in A}} P_{ij} \cdot 0 + \sum_{v_j \notin A} P_{ij} \cdot t_j$

Positive correlation between expected nr. of edges and physical transition time measured in wet laboratory

Molloy, Clausen, and Shehu. A Stochastic Roadmap Method to Model Protein Structural Transitions. Robotica 34(08):1705-1733, 2016 (featured on cover).

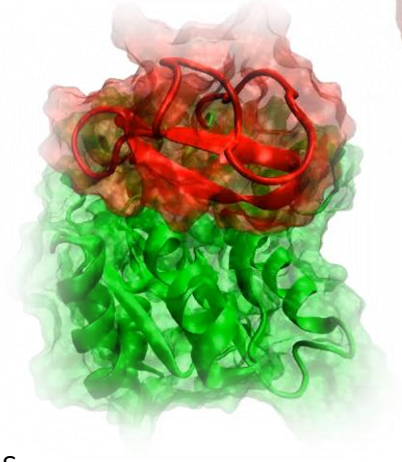
Feasible (Robotics-inspired) Models of Dynamics via Adaptive Search

Calmodulin

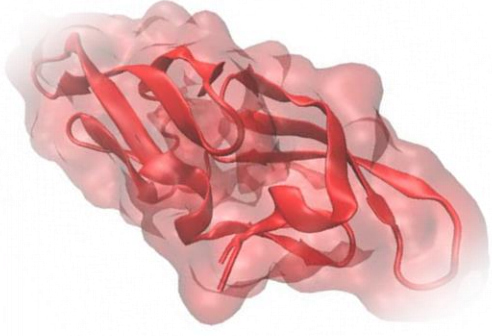


> 13Å open-closed motions
accommodating different binding partners
regulating cascade of signals in living cell

Adenylate Kinase

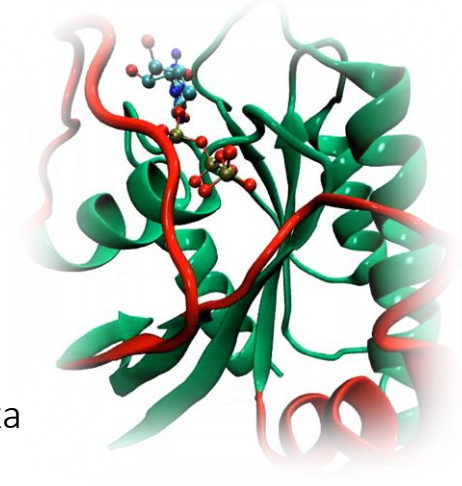


Cyanovirin-N



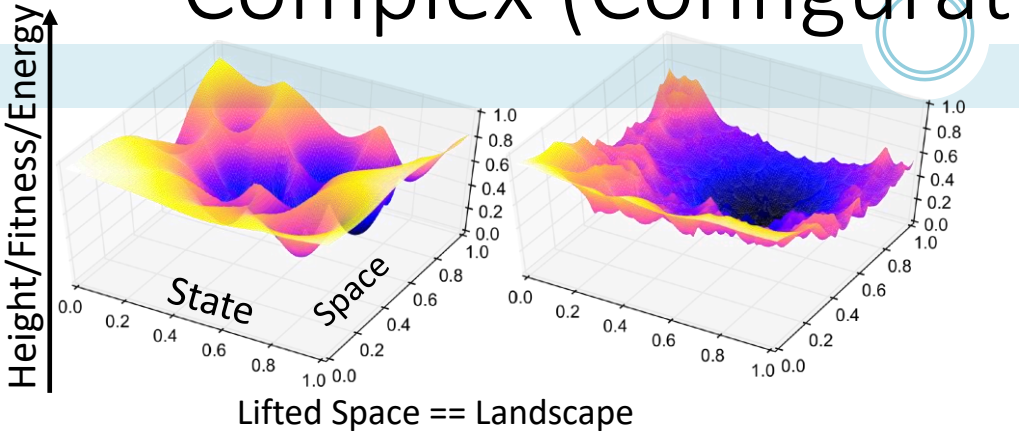
> 16Å motion
potent virucidal protein
against HIV-I and influenza

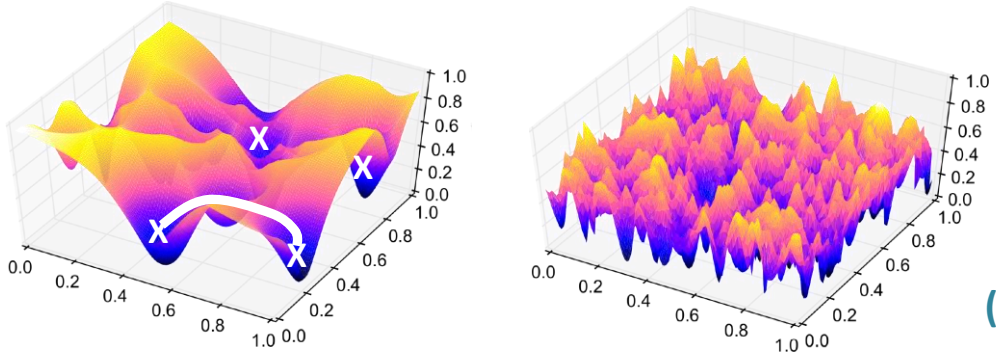
H-Ras



2.5Å on <- -> off switching
regulating cell growth

Navigation on and Exploration of Complex (Configuration) Landscapes





State-to-state navigation

Optimal A → B path(s)

Local view of landscape

State space exploration

Find novel states

Global view of landscape

(Protein) Structure(s) → Dynamics → Function

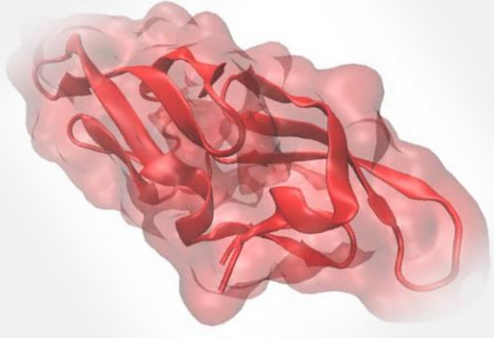
Optimization Problems → Configuration Landscapes → Stochastic Optimization Algorithm/Framework

Connections between robotics-inspired optimization and evolutionary algorithms

What About State Space Exploration?

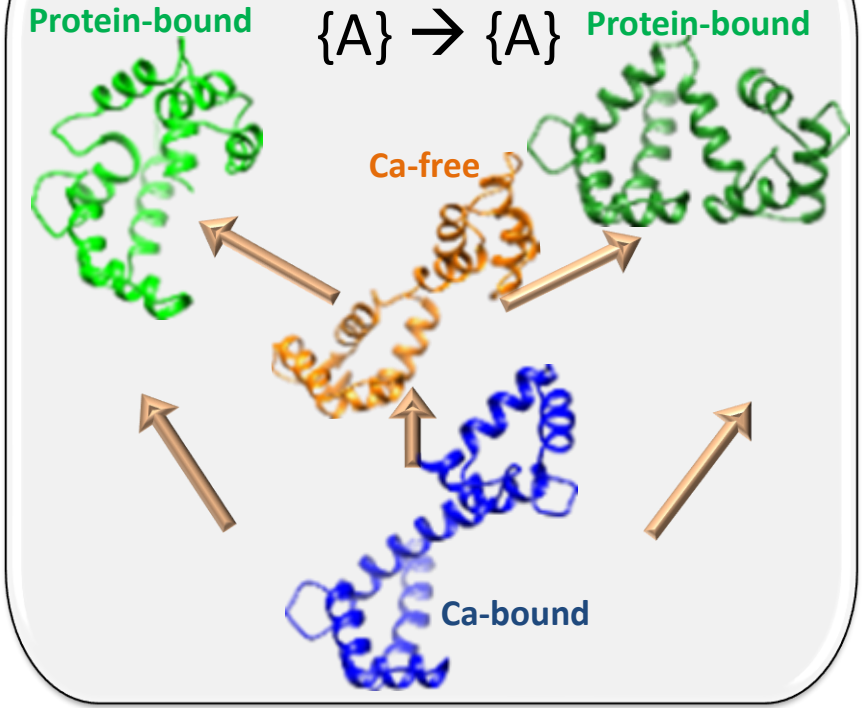
State-to-state Navigation

A → B



State-space Exploration

{A} → {A}



Personal Journey

No to very little data – explicit knowledge

2002-2016

Classic AI: stochastic search- optimization (geometry, kinematics, inverse kinematics, motion planning)

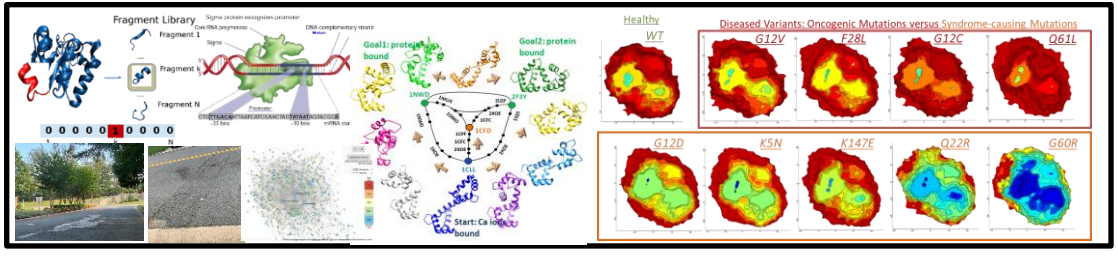
-- molecular structural biology

Some data –explicit and tacit knowledge

2010-

Hybrid Models (data-driven AI, knowledge-guided shallow ML, shallow ML + AI)

-- sequence/structural biology, social media user modeling, industrial monitoring, urban planning

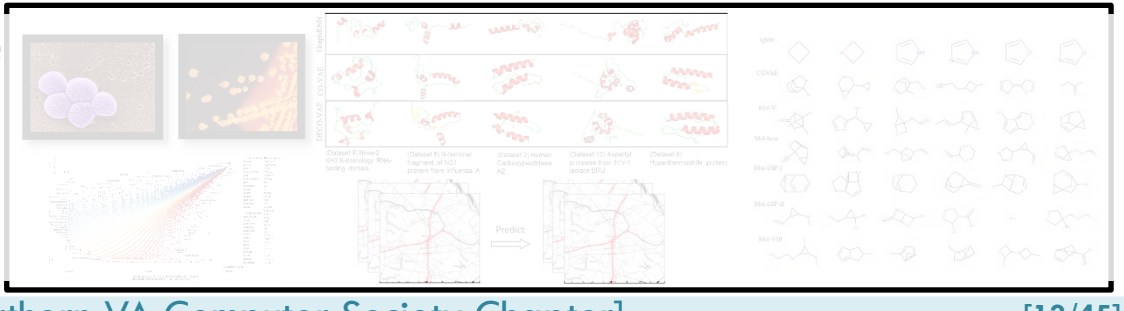


Lots of data – AI romance w/ tacit knowledge

2018-

Deep Learning, NLP, Deep generative models

-- sequence/structural biology, mental health, traffic forecasting, AI for Policy



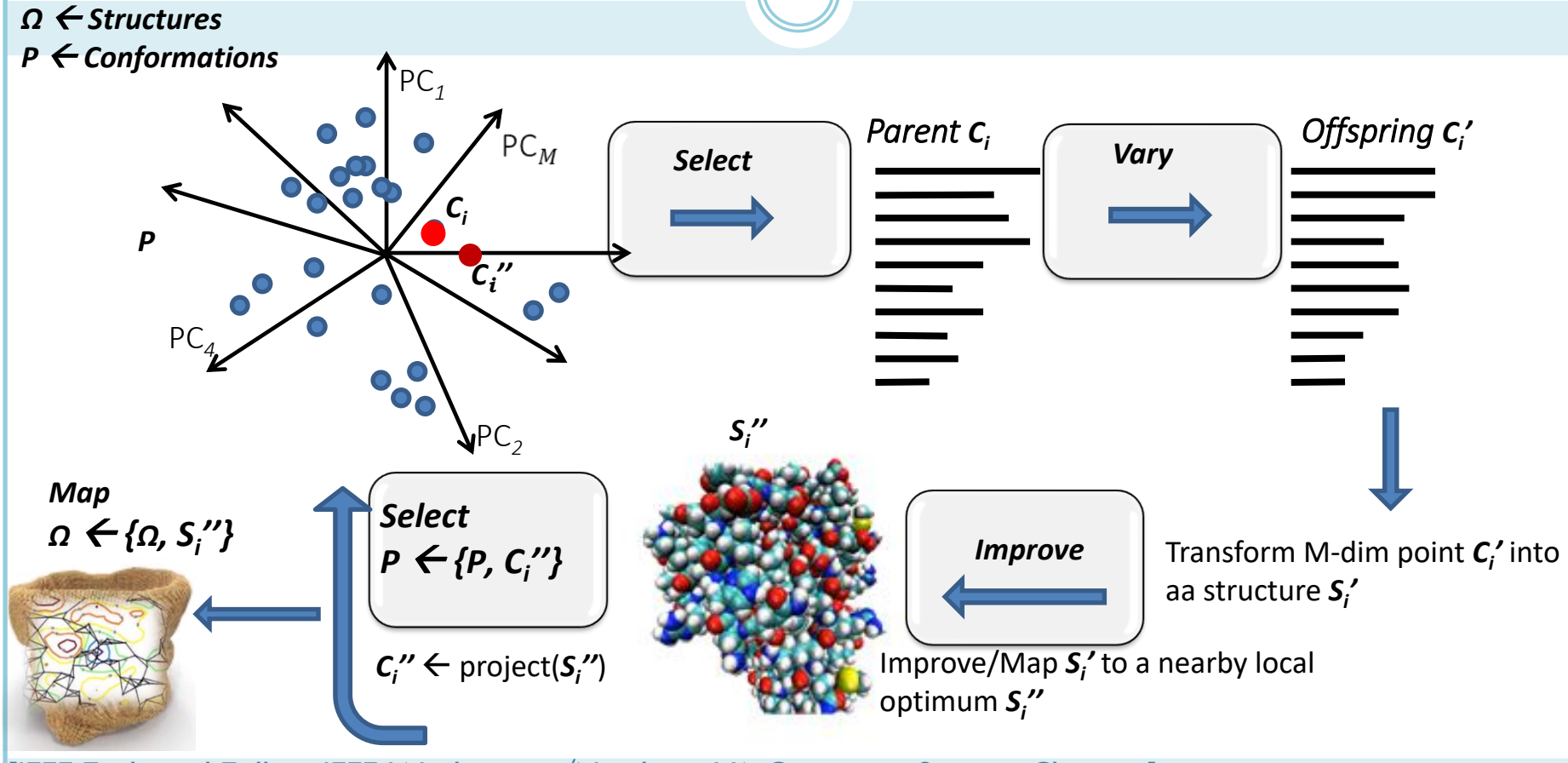
Chapter II.a – AI Leveraging Small Data Connecting Dynamics to (Dys)Function

- <u>Small data:</u> For many proteins, we have now accumulated (~20-100ish) structures "caught" at various conditions, with different binding partners, in naturally-occurring and mutated variants (sequences)
- <u>Knowledge guidance:</u> Conformational selection/population shift principle: external and internal perturbations only affect population probabilities and not the configuration/state space
  - Structures do not emerge, they are there
  - Change in conditions makes some structures more likely than others

Experimentally-determined structures of WT and diseased variants are ***known points*** in the state space!

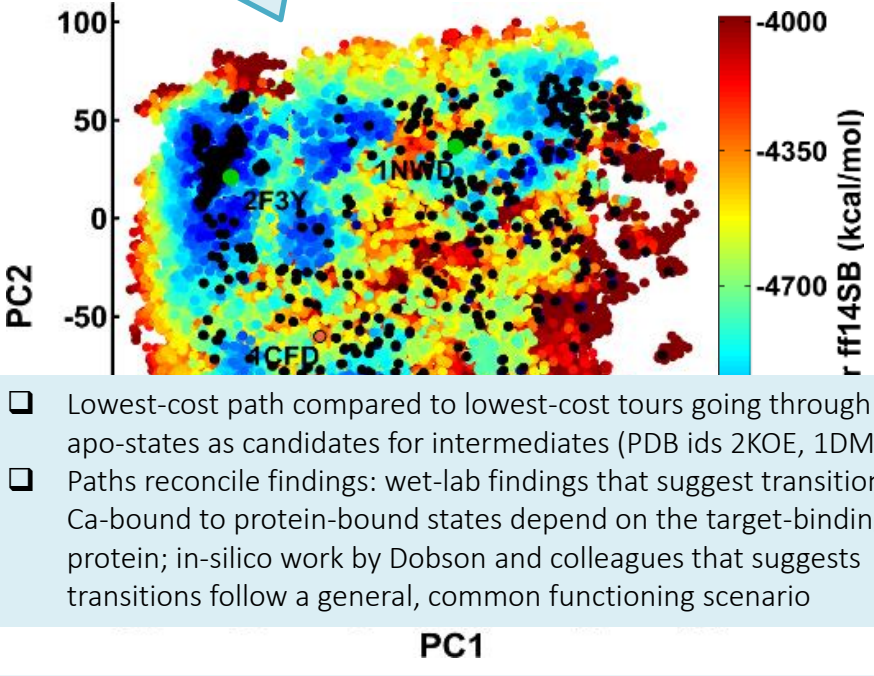
Leverage them to <u>define</u> and <u>initialize</u> variable space

Knowledge-guided and Data-driven AI: EA Sampling of Protein Energy Landscape

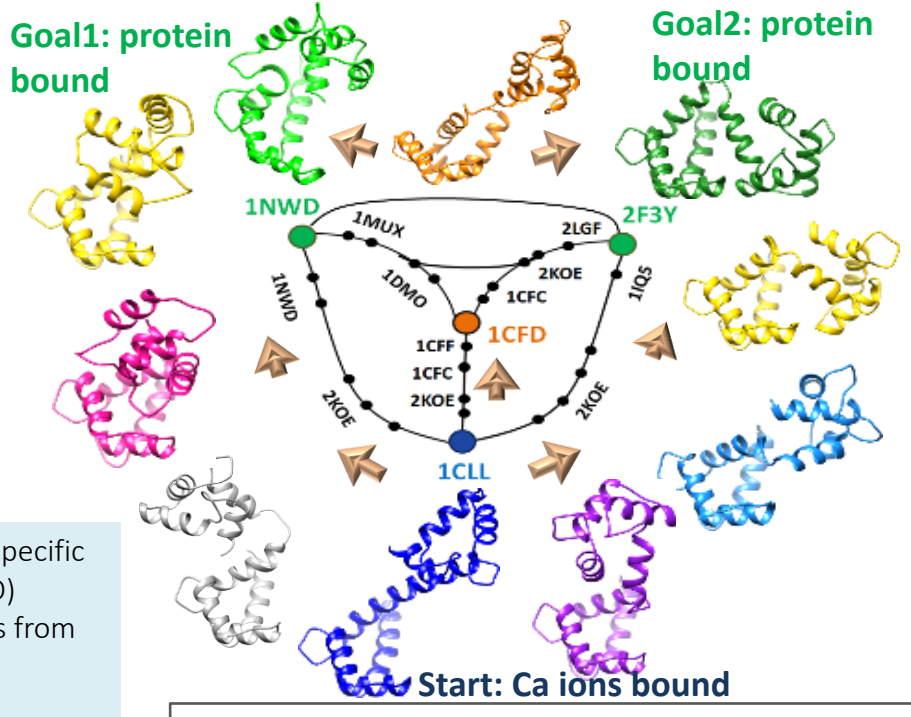


Mapping Energy Landscape of Calmodulin

First-ever view of CaM energy landscape



- Lowest-cost path compared to lowest-cost tours going through specific apo-states as candidates for intermediates (PDB ids 2KOE, 1DMO)
- Paths reconcile findings: wet-lab findings that suggest transitions from Ca-bound to protein-bound states depend on the target-binding protein; in-silico work by Dobson and colleagues that suggests transitions follow a general, common functioning scenario



Maximova, Plaku, and Shehu. Structure-guided Protein Transition Modeling with a Probabilistic Roadmap Algorithm. IEEE/ACM Trans Comp Biol and Bioinf (TCBB)15:(6), 1783-1796, 2018.

Sample-based Representation of Protein (and Peptide) Energy Landscape

First-ever view of H-Ras energy landscape

WT GTP bound, ON (1QRA)
WT canonical structure (1CTQ)

WT GTP GAP-bound, 1WQ1
WT, A59G **SOS-bound**, 1NV[W,X,U] - the same mechanism (as for GAP-bound), involving **G12, Q61** in catalysis

Q61L, G12R, G12V-A59T, P38E, **Mutations, no effectors** (721P, 421P, 521P, 221P)

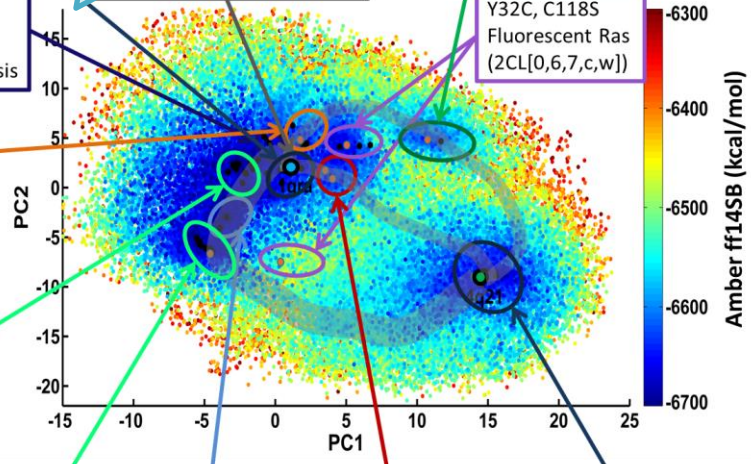
Allosteric switch, shift ON
- ❑ G12V + Raf kinase (3OI[W,U])
- ❑ Q70E + Ca compounds (4DL[W,T,S])
- ❑ WT soaked in Ca compounds (acetate, chlorine …) (3LB[H,I,N])
- ❑ WT (3K8Y)

Allosteric switch, shift OFF
- ❑ G12V + Raf kinase (3OIV)
- ❑ Q70E + Ca compounds (4DL[R,V,X,Z])
- ❑ Q61L,I,K,V + Raf (2RG[A,B,C,D])
- ❑ No mutations (3RS[0,2,3,5])

…DS (1LFD), (6Q21.B) …

…GDP (1Q21, 2Q21)

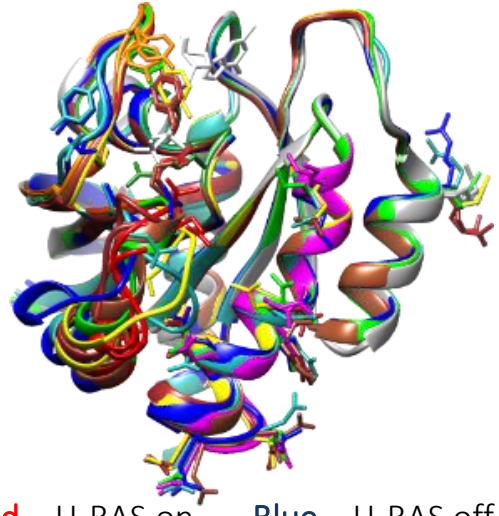
Y32C, C118S Fluorescent Ras (2CL[0,6,7,c,w])



WT (6Q21.A, (3L8Y with cyclen)

A59G - 1LF0
WT - 6Q21.D

WT GDP bound, OFF (4Q21)

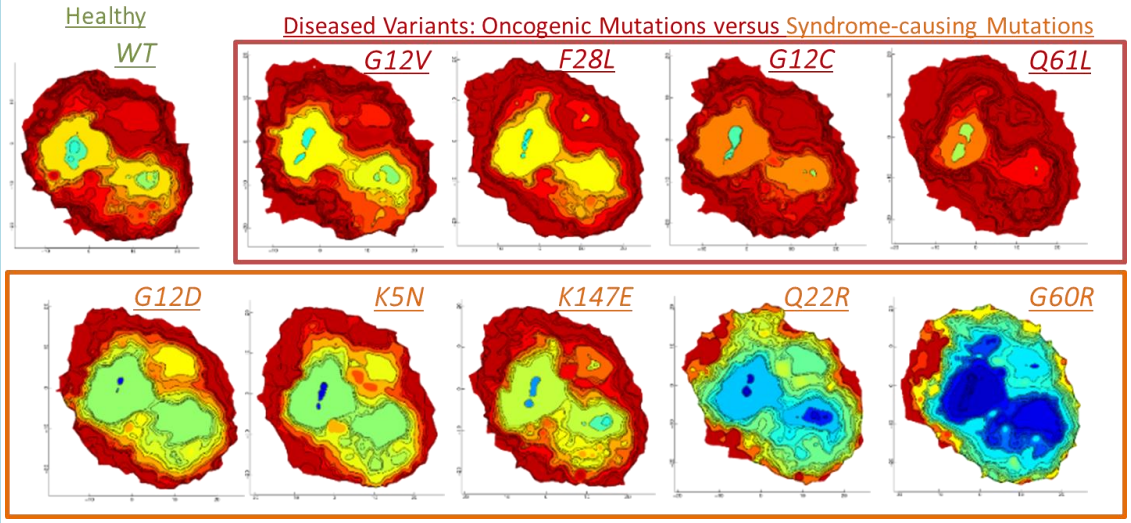


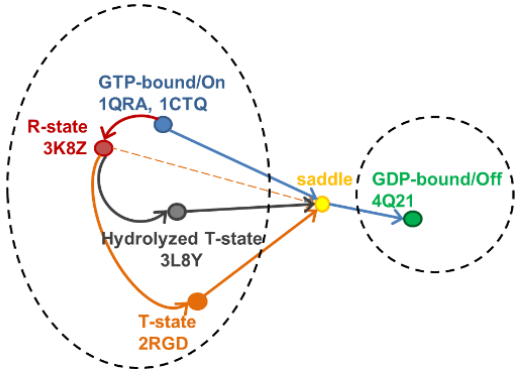
Red – H-RAS on Blue – H-RAS off
Light blue - 1Q21 (GDP bound)
Shades of Brown – allosteric switch
Yellow – 1LF0 (A59G)
Green, Orange – 412P, 721P (G12R, Q61L)

Clausen, Ma, Nussinov, and Shehu. Mapping the Conformation Space of Wildtype and Mutant H-Ras with a Memetic, Cellular, and Multiscale Evolutionary Algorithm. PLoS Computational Biology 11(9): e1004470, 2015.

Predicting Phenotypical Impact of Mutations

Level-set based analysis allows identification of basins and saddles and reconstruction of landscape from hundreds of thousands of multi-dimensional (sampled) points corresponding to protein structures





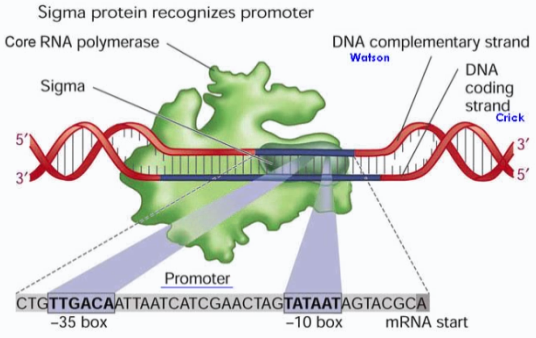
Spatial and energetic distances of basins/states of interest be extracted as *landscape descriptors*/features

Variations of each landscape-extracted descriptor (across variants) correlated to variations of biochemical parameters of various activities measured in wet laboratory

Qiao, Akhter, Fang, Maximova, Plaku, and Shehu. From Mutations to Mechanisms and Dysfunction via Computation and Mining of Protein Energy Landscapes. BMC Genomics 19 (Suppl7):671, 2018.

Chapter II.b – AI Leveraging Small Data

Knowledge-guided AI + Shallow ML, Shallow ML + AI



Kamath, De Jong, and Shehu. Effective Automated Feature Construction and Selection for Classification of Biological Sequences. PLoS One, 9(7): e99982, 2014.

Kamath, Compton, Islamaj Dogan, De Jong, and Shehu. An Evolutionary Algorithm Approach for Feature Generation from Sequence Data and its Application to DNA Splice-Site Prediction. IEEE Trans Comp Biol and Bioinf 2012, 9(5):1387-1398.

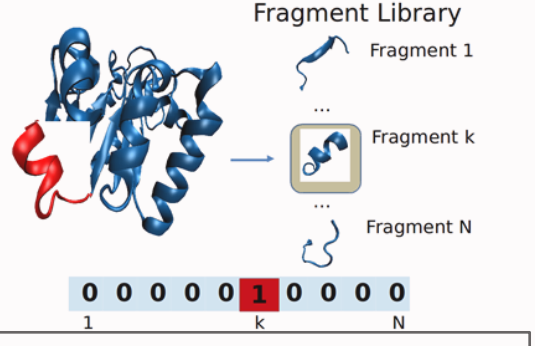
Kamath, Shehu, and De Jong. A Two-Stage Evolutionary Approach for Effective Classification of Hypersensitive DNA Sequences. J Bioinf and Comp Biol 2011, 9(3): 399-413.

Kamath, De Jong, and Shehu. An Evolutionary-based Approach for Feature Generation: Eukaryotic Promoter Recognition. IEEE Congress on Evol Comput, New Orleans, 2011, 277-284.

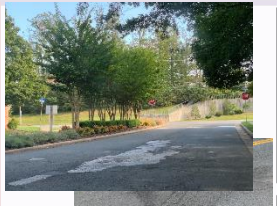
Methicillin-resistant *Staphylococcus aureus*, Carbapenem-resistant Enterobacteriaceae --multi-drug resistant bacteria

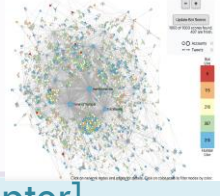
Veltri, Kamath, and Shehu. Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming. IEEE/ACM Trans Comp Biol and Bioinf, 14(2): 300-313, 2017.

Veltri, Kamath, and Shehu. A Novel Method to Improve Recognition of Antimicrobial Peptides through Distal Sequence-based Features. IEEE Intl Conf on Bioinf and Biomed, Belfast, UK, 2014, pg. 371-378 (Best Student Paper Award).



Molloy, Van, Barbarà, and Shehu. Exploring Representations of Protein Structure for Automated Remote Homology Detection and Mapping of Protein Structure Space. BMC Bioinformatics 15 (Suppl 8):S4, 2014.



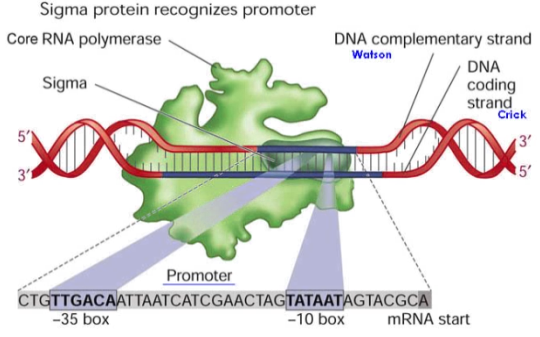


Kamranfar, Lattanzi, Shehu, and Stoffels. Pavement Distress Recognition via Wavelet-based Clustering of Smartphone Accelerometer Data. Journal of Computing in Civil Engineering, 2022.

Kamranfar, Lattanzi, and Shehu. Meta-Learning for Industrial System Monitoring via Multi-objective Optimization. Intl Conf on Data Science, Las Vegas, 2020.

Rajabi, Shehu, and Purohit. User Behavioral Modeling for Fake Information Mitigation on Social Web. SBP-BRiMS, Washington, D.C. 2019, 234-244.

Focus on Representation
Capture Function Constraints on Sequence

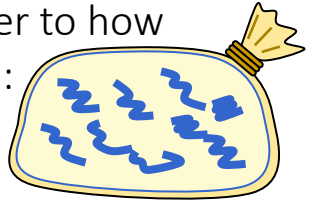


- ***Key insight***: encode implicit constraints in linear representation
  - Non-local/distal constraints imposed by function
  - Capture them as features

Example of *a* biological signature: Motif 'TTGACA' at some position i AND 'TATAAT' at some position j

- Rich signatures:
  - Compositional
  - Positional
  - Correlational
  - …

- Explicit, *interpretable* answer to how sequence encodes function:



(Sequence) Form → Function: Explicit, Interpretable Signatures

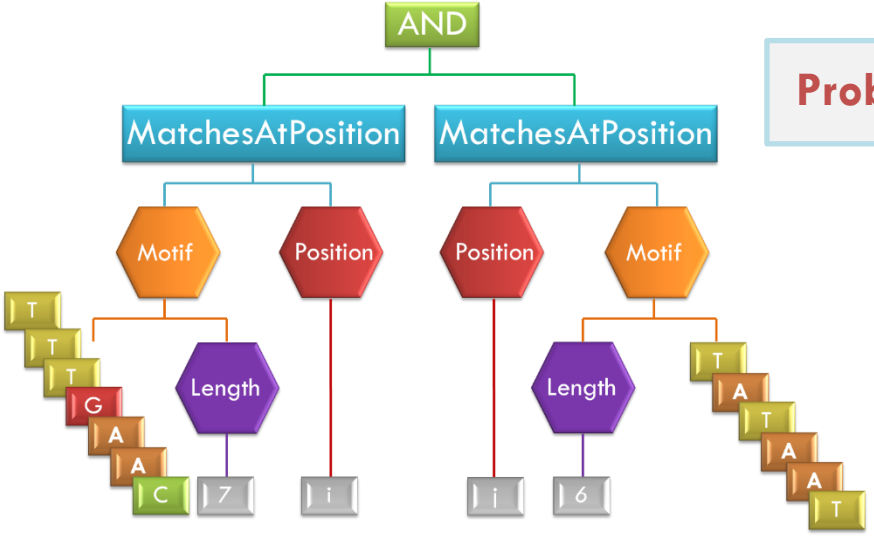
- Contribution #1
  - Richer representation for local and non-local constraints
    - Structured representation in predicate logic:
      - Boolean combinations over basic building blocks

Grammar Domain

Name	Args	Return Type	Constraints
AND	2 non-terminals	Boolean	
OR	2 non-terminals	Boolean	
NOT	2 non-terminals	Boolean	
Matches	Motif	Boolean	
MatchesAtPosition	Motif, Position	Boolean	
Motif	ERC-chars	Motif	
Position	ERC-int	Integer	{1, . . . , 162}
Length	ERC-int	Integer	{2, . . . , 6}
ERC-char		Character	{A, C, G, T}

Operators

Termi



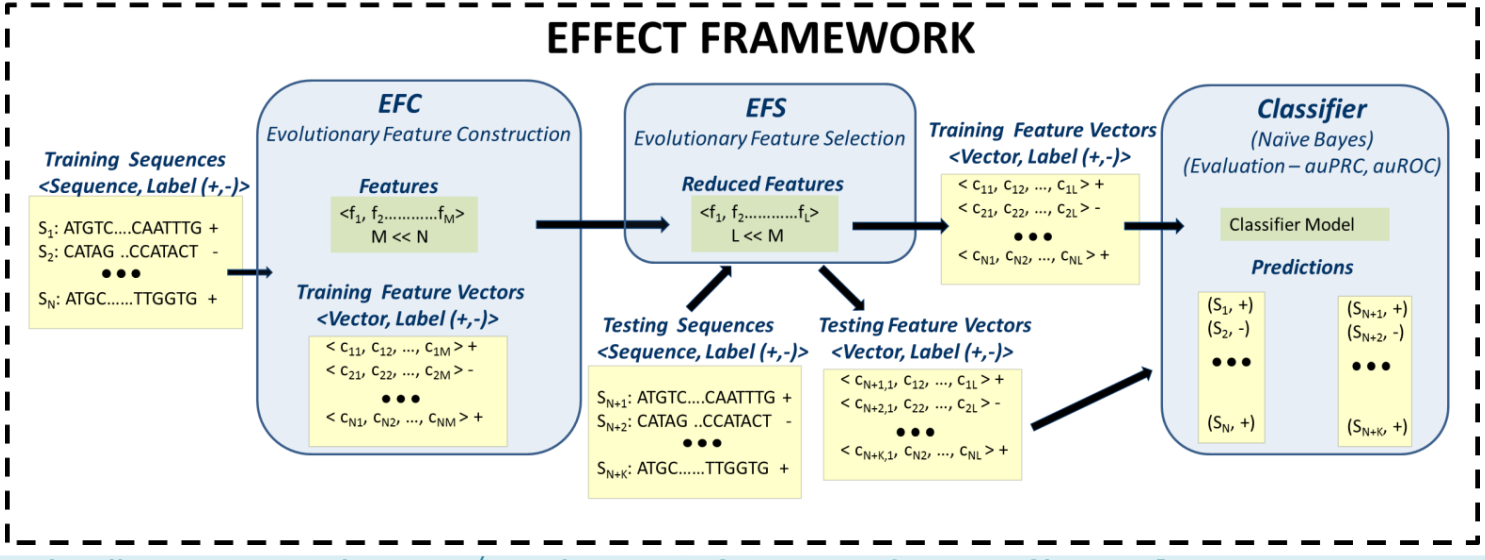
Problem: Exponential explosion of feature space!

- Contribution #2
  - Genetic Programming for feature space exploration
  - Surrogate fitness function instead of wrapper classification model
  - Complete treatment in classification context

(Sequence) Form → Function : Explicit, Interpretable Signatures

Automated generation of complex yet interpretable features that satisfy local and non-local (implicit) constraints posed by function

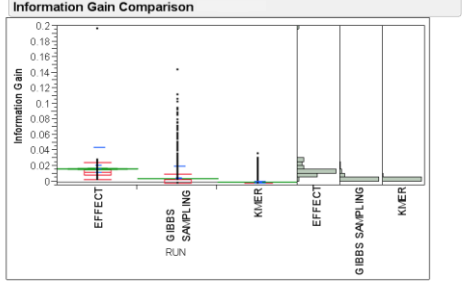
Open-source



Superior Performance on Hard Problems:

Recognition of HSS Sites, Promoter Sites, and More

Algorithm	auROC	auPRC
Feature-based		
K-mer	82.20	82.6
Gibbs Sampling	79.3	50.3
EFFECT	**89.7**	**89.2**
Statistical-based		
PWM-HMM	70.8	47.8
BayesNetwork	72.5	49.5
HomogenousHMM	82.02	71.5
WAM-HMM	80.05	70.0
MSP	85.5	72.9
Kernel-based		
WeightedPosition	80.01	62.3
WeightedPositionShift	80.93	64.9



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
EFFECT	45	0.017658	0.027906	0.00416	0.00927	0.02604
GIBBS SAMPLING	1030	0.005664	0.015545	0.00048	0.00471	0.00661
KMER	65535	0.000282	0.001252	4.89e-6	0.00027	0.00029

Interpretable features validated and advancing knowledge

Known signals in a typical pre-mRNA include the branch site, the pyrimidine-rich region, splice site consensus signals, and

Feature and Kernel Evolution for Improved Classification via SVM

Many A/C-rich motifs, such as CACACA, GCCCAA, CATTCA, CCTACA, found and hypothesized in experiment

Kamath, De Jong, and Shehu. Effective Automated Feature Construction and Selection for Classification of Biological Sequences. PLoS One, 9(7): e99982, 2014.

```
(OR
 (MP (Motif (5) AGGCG) @ 84)
  (OR
   (MP (Motif (3) TCG) @ 47)
    (OR
     (OR
      (OR
       (OR
        (MP (Motif (2) GT) @ 85)
         (OR
          (MP (Motif (3) AGC) @ 79)
          (MP (Motif (2) AG)  @ 84) ) )
        (MP (Motif (3) GAG) @ 83) )
       (MP (Motif (2) TC) @ 47) )
      (MP (Motif (3) AGG) @ 84) ) ) )
```

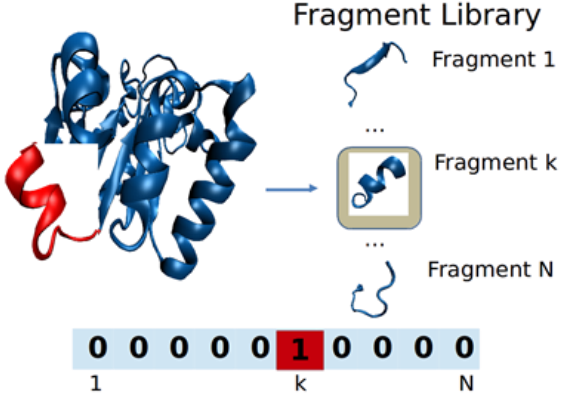
(a)

```
(OR
 (AND
  (Match (Motif (3) AGC)
  (MP (Motif (2) GT) @ 85)
 )
 (OR
  (OR
   (MP (Motif (2) GT) @ 85)
   (MP (Motif (2) AG) @ 84) )
  (AND
   (MP (Motif (2) CC) @ 65)
   (MP (Motif (6) TAACCG) @ 151) ) ) )
```

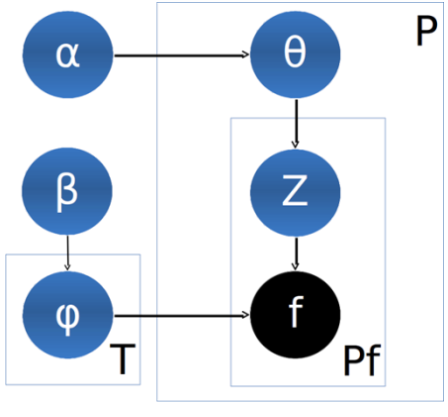
Shallow ML Pt I: Focus on Features (Protein Structure) Form → Function

Remote (protein) homologs:
Low sequence identity but high structure → function similarity

- How to extract functional information from structure?
- Key insight: Add spatial information to building blocks
- Objective: reduced yet informative representation of structure



- Analogies with text mining
- Topic-based representation via Latent Dirichlet Allocation (LDA)
- Reduction: 400→10 dimensions!

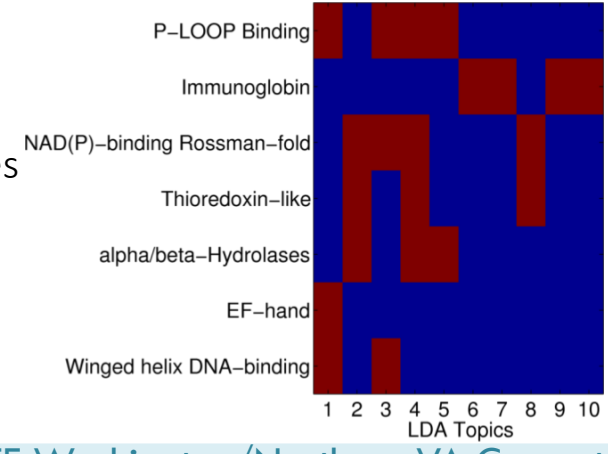


Superfamily Recognition with Support Vector Machines over Learned Topics

- Prediction of superfamily membership

	Fragbag Representation				Topic-Based Representation			
SCOP Superfamily	Accuracy (%)	TPR	FPR	AUC	Accuracy (%)	TPR	FPR	AUC
P-Loop Binding	96.4	0.98	0.05	0.95	84.3	0.97	0.29	0.84
Immunoglobin	100.0	1.00	0.00	1.000	99.9	0.99	0.0	1.0
NAD(P)-binding Rossman Fold	98.7	0.99	0.02	0.99	90.9	0.94	0.13	0.91
Thioredoxin-like	98.8	0.98	0.01	0.99	80.2	0.92	0.32	0.80
alpha/beta Hydrolases	99.1	1.00	0.02	0.99	92.7	0.95	0.10	0.93
EF-hand	100.0	1.00	0.00	1.000	98.8	0.99	0.01	0.99
Winged helix DNA-binding	98.7	0.98	0.01	0.99	84.4	0.79	0.11	0.84

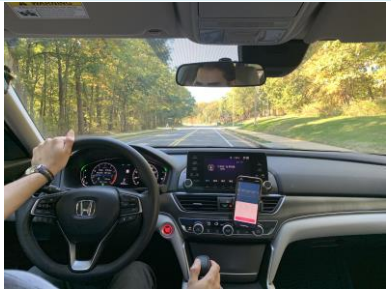
- Topic signatures across superfamilies



- Other contributions:
  - Detection of remote homologs
  - Organization of protein structure space that preserves function co-localization

Molloy, Van, Barbarà, and Shehu. Exploring Representations of Protein Structure for Automated Remote Homology Detection and Mapping of Protein Structure Space. BMC Bioinformatics 15 (Suppl 8):S4, 2014.

Predicting Pavement Distress from Passively-collected Smartphone Data

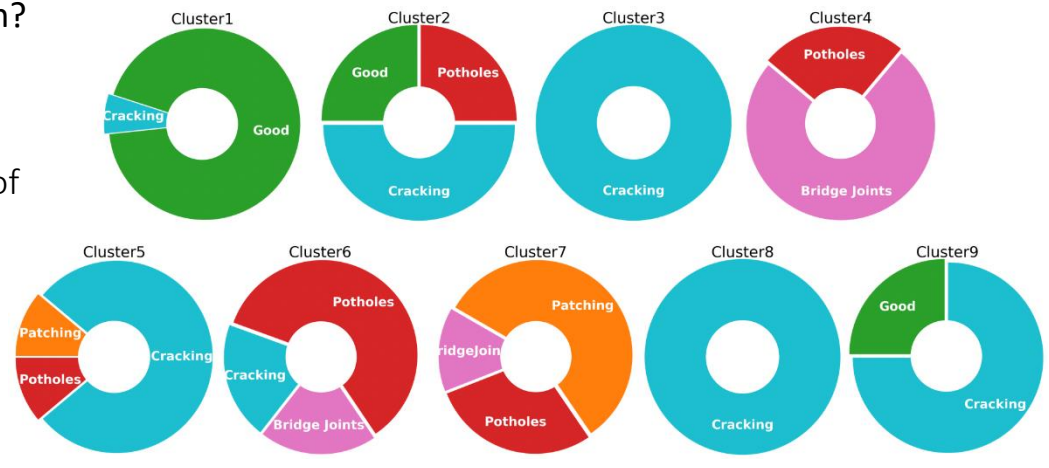


- 2018 Honda Accord vehicle and iPhone XS smartphone to collect accelerometer Z data on road segments in NOVA
- Human labels: sparse and noisy

Bridge joints, Cracking, Potholes, Patching, vs. Normal

Example of challenges:
how to distinguish "normal" patch from utility patch?

- Unsupervised learning over wavelet-based features to group data into clusters
- Internal multi-objective Pareto-based selection of unsupervised strategy
- Evaluation informed by present labels shows coarse distinctions can be made



Kamranfar, Lattanzi, Shehu, and Stoffels. Pavement Distress Recognition via Wavelet-based Clustering of Smartphone Accelerometer Data. J of Computing in Civil Engineering, 2022.

Personal Journey

No to very little data – explicit knowledge

2002-2016

Classic AI: stochastic search- optimization (geometry, kinematics, inverse kinematics, motion planning)
-- molecular structural biology

Some data –explicit and tacit knowledge

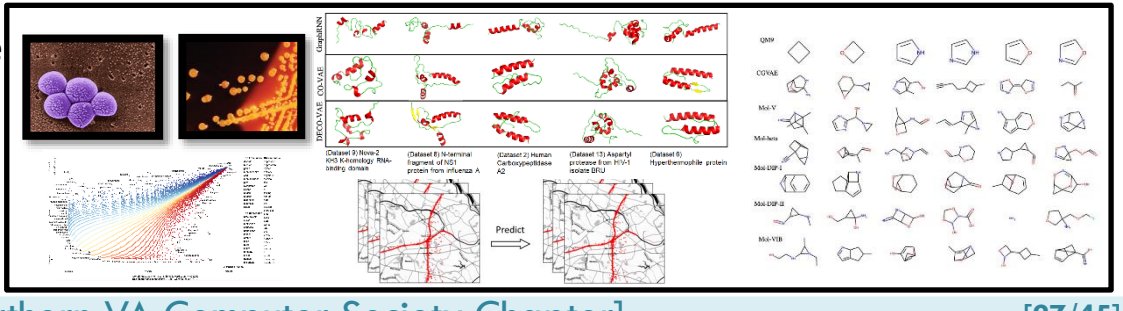
2010-

Hybrid Models (data-driven AI, knowledge-guided shallow ML, shallow ML + AI)
-- sequence/structural biology, social media user modeling, industrial monitoring, urban planning

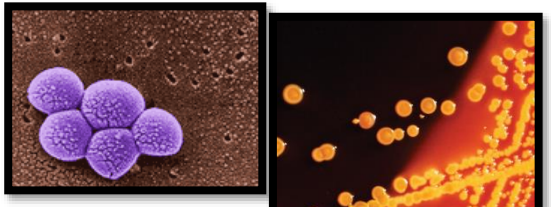
Lots of data – AI romance w/ tacit knowledge

2018-

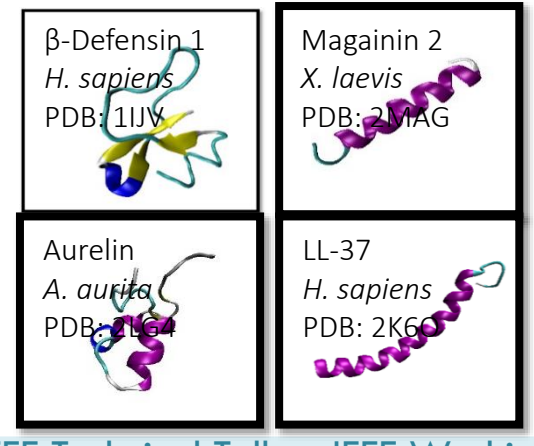
Deep Learning, NLP, Deep generative models
-- sequence/structural biology, mental health, traffic forecasting, AI for Policy



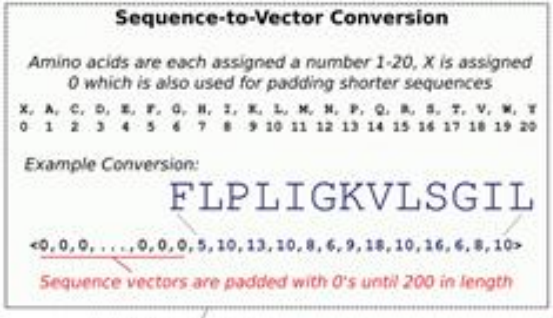
Our First Foray into DL: Sequence → Antimicrobial Activity

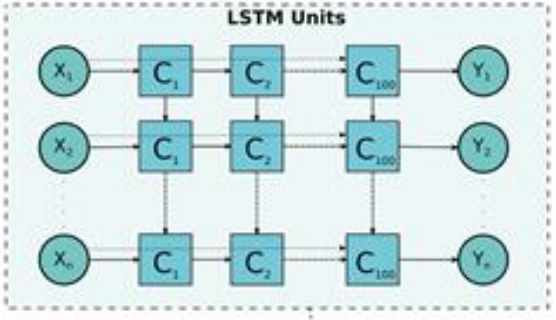


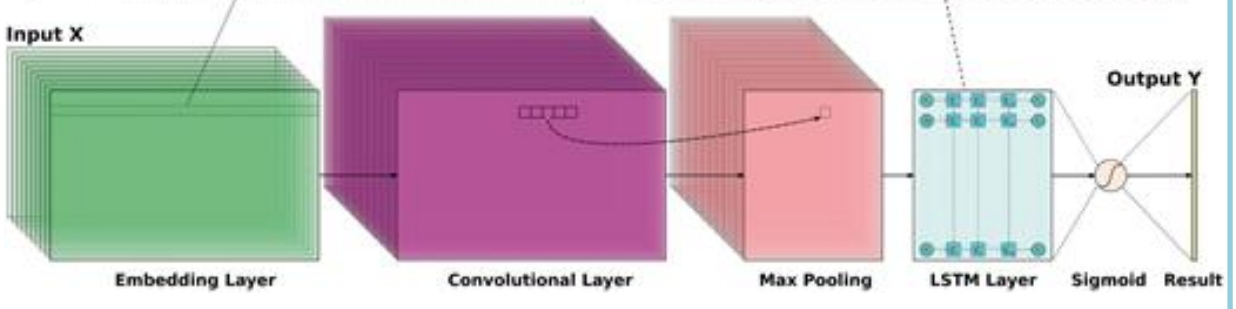
Methicillin-resistant *Staphylococcus aureus* (MRSA), Carbapenem-resistant Enterobacteriaceae, etc. Increasing numbers of multi-drug resistant bacteria



- Convolutional layers allow incorporating local effects from amino-acid neighbors in the peptide chain
- Recurrent layer handles arbitrary-length peptides
- Outperforms all existing ML models (including our own 2017 work)







Sequence → Function via Deep Learning:

Sacrificed Interpretability

Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming

Daniel Veltri, Uday Kamath, and Amarda Shehu

Abstract—Growing bacterial resistance to antibiotics is spurring research on utilizing naturally-occurring antimicrobial peptides (AMPs) as templates for novel drug design. While experimentalists mainly focus on systematic point mutations to measure the effect on antibacterial activity, the computational community seeks to understand what determines such activity in a machine learning setting. The latter seeks to identify the biological signals or features that govern activity. In this paper, we advance research in this direction through a novel method that constructs and selects complex sequence-based features which capture information about distal patterns within a peptide. Comparative analysis with state-of-the-art methods in AMP recognition reveals our method is not only among the top performers, but it also provides transparent summarizations of antibacterial activity at the sequence level. Moreover, this paper demonstrates for the first time the capability not only to recognize that a peptide is an AMP or not but also to predict its target selectivity based on models of activity against only Gram-positive, only Gram-negative, or both types of bacteria. The work described in this paper is a step forward in computational research seeking to facilitate AMP design or modification in the wet laboratory.

Index Terms—Antimicrobial peptide recognition, Gram-positive, Gram-negative, feature construction, feature selection, evolutionary computing, genetic programming, evolutionary algorithms, machine learning

1 Introduction

The U.S. Center for Disease Control estimates that more than two million people in the U.S. are diagnosed with antibiotic-resistant infections every year. With some suggesting an era of untreatable infections has arrived [1], there is renewed focus on pursuing novel antibacterials [2]. The discovery of anti-pathogen peptides in the innate immune system of many organisms has been met with great enthusiasm. The effectiveness of these antimicrobial peptides (AMPs) in killing even resistant bacteria has spurred significant research in the last two decades on characterizing AMPs and understanding how they can be effectively employed to combat even multi-drug resistant bacteria [3].

Experimental and computational studies devoted to answering the open question of what governs antibacterial activity in AMPs have generally proceeded orthogonally. In the experimental community, the focus has been largely on template-based studies (where known AMPs are modified and tested against bacterial cultures in the wet laboratory) and systematic virtual screenings of peptide libraries [3]. Such studies, though narrow in scope, have advanced knowledge by elucidating what biological properties correlate with antibacterial activity. For instance, studies of interactions with bacterial membranes rule out the employment of a universal sequence motif and instead have led to fundamental determinants or features, such as residue composition, charge, length, secondary structure, hydrophobicity, and amphipathic character [4]. Though laborious and on a case-by-case setting, wet-lab studies are expected to reveal more features that contribute to antibacterial activity [3].

Computational research has focused on AMP recognition as a means of understanding what features relate to activity. Techniques from machine learning are applied, seeking to test the predictive power of a given set of features in the context of supervised classification. Methods of choice include support vector machines (SVM), hidden Markov models (HMMs), artificial neural networks (ANN) and logistic regression (LR) [5], [6], [7], [8], [9], [10], [11]. Features vary, from those elucidated by wet-lab studies which characterize the entirety or part of a peptide, to simple ones based on amino acid composition [7], [8], and to averaged whole-peptide physicochemical profiles built on known amino acid properties [9]. Recently, wet-lab studies have begun to use some of these classifiers with limited success as an initial screening mechanism for new AMP sequences [12].

As Table 1 summarizes, the recognition accuracy of machine learning methods ranges from the upper 70 to the lower 90 percent. Direct comparisons are difficult due to the use of different training and testing datasets. Some high performers fall short on more recent challenging datasets [11]. The consensus is that performance has stagnated, and the community is shifting its attention to constructing effective features [13]. This is non-trivial, not

• *D. Veltri is in the School of Systems Biology, George Mason University, Fairfax, VA 22030. E-mail: dveltri@gmu.edu.*
• *U. Kamath is with OntoLabs, 20929 Ironmount Terrace, Ashburn, VA 20147. E-mail: ukamath@gmu.edu.*
• *A. Shehu is with the Department of Computer Science, George Mason University, Fairfax, VA 22030. E-mail: amarda@gmu.edu.*

Manuscript received 3 Mar. 2015; accepted 15 July 2015. Date of publication 29 July 2015; date of current version 22 Mar. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2462364

Sequence analysis

Deep learning improves an … recognition

Daniel Veltri[1,2,*], Uday Kamath[3] and Am…

[1]Bioinformatics and Computational Biosciences Branch, Office … National Institute of Allergy and Infectious Diseases, U.S. Nati… [2]Medical Science & Computing, LLC, Rockville, MD 20852, U… [3]Department of Computer Science and [4]Department of Bioe… 22030, USA and [5]School of Systems Biology, George Mason Un…

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on December 14, 2017; revised on March 6, 2018; editorial decision on … and published on March 24, 2018

Abstract

Motivation: Bacterial resistance to antibiotics is a growing concern. Antimicrobial peptides (AMPs), natural components of innate immunity, are popular targets for developing new drugs. Machine learning methods are now commonly adopted by wet-laboratory researchers to screen for promising candidates.

Results: In this work, we utilize deep learning to recognize antimicrobial activity. We propose a neural network model with convolutional and recurrent layers that leverage primary sequence composition. Results show that the proposed model outperforms state-of-the-art classification models on a comprehensive dataset. By utilizing the embedding weights, we also present a reduced-alphabet representation and show that reasonable AMP recognition can be maintained using nine amino acid types.

Availability and implementation: Models and datasets are made freely available through the Antimicrobial Peptide Scanner vr.2 web server at www.ampscanner.com.

Contact: amarda@gmu.edu (for general inquiries) or dan.veltri@gmail.com (for web server information)

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

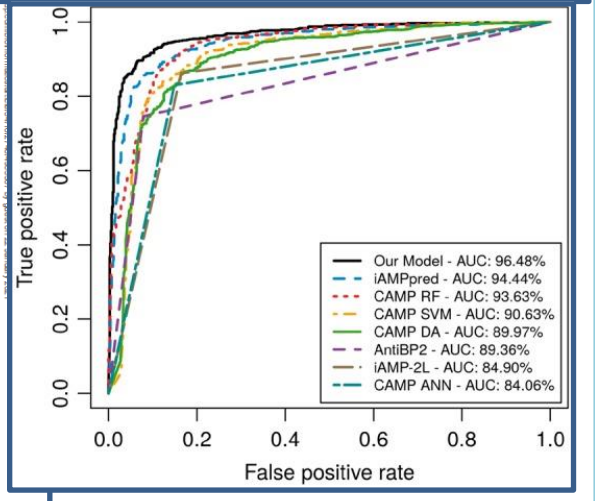
Antimicrobial resistance remains a serious problem for humans and livestock around the world, as more drugs lose sensitivity to the pathogens they were designed to eliminate (Price *et al.*, 2012; U.S. Department of Health and Human Services, 2013; World Health Organization, 2014). Over the past few decades, natural antimicrobial peptides (AMPs) have been an active area of research and have shown a lowered likelihood for bacteria to form resistance compared to many conventional drugs (Boman, 2003; Zelezetsky *et al.*, 2006). AMPs are short innate immunity peptides that fall into a number of diverse sequence families (e.g. cathelicidins, defensins, cecropins, etc.) and kill their targets through various mechanisms, such as cell membrane damage, DNA interference or signaling for adaptive immune responses, as reviewed by Wimley and Hristova (2011). While we focus in this work exclusively on peptides that kill Gram-positive and/or Gram-negative bacteria, we note some AMPs have also been shown effective against a variety of fungal and viral pathogens (Wang, 2010).

To aid wet-laboratory researchers in novel AMP discovery, a variety of computational approaches are proposed for AMP recognition. Many incorporate machine learning algorithms or statistical analysis techniques, such as artificial neural networks (ANN) (Lata *et al.*, 2010; Thomas *et al.*, 2009; Torrent *et al.*, 2011), discriminant analysis (DA) (Thomas *et al.*, 2009), fuzzy *k*-nearest neighbor (Xiao *et al.*, 2013), hidden Markov models (Fjell *et al.*, 2009), logistic regression (Veltri *et al.*, 2017; Randou *et al.*, 2013), random forests (RF) (Thomas *et al.*, 2009; Veltri, 2015) and support vector machines (SVM) (Lata *et al.*, 2010; Lee *et al.*, 2016; Meher *et al.*, 2017; Thomas *et al.*, 2009; Torrent *et al.*, 2011).

Table 2. Performance comparison on the AMP dataset testing partition

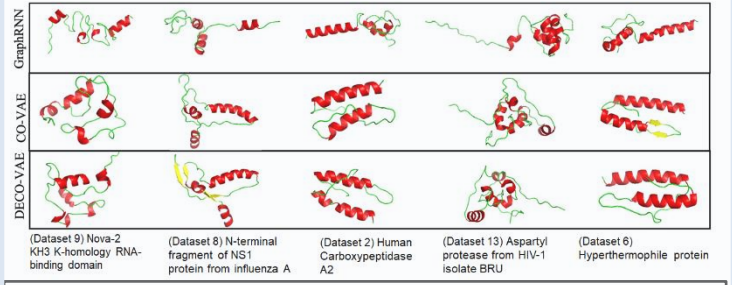
Method	SENS(%)	SPEC(%)	ACC(%)	MCC	auROC(%)
AntiBP2	87.91	90.80	89.37	0.7876	89.36
CAMP-ANN	82.98	85.09	84.04	0.6809	84.06
CAMP-DA	87.08	80.76	83.92	0.6797	89.97
CAMP-RF	**92.70**	82.44	87.57	0.7554	93.63
CAMP-SVM	88.90	79.92	84.41	0.6910	90.63
iAMP-2L	83.99	85.86	84.90	0.6983	84.90
iAMPpred	89.33	87.22	88.27	0.7656	94.44
gkmSVM	88.34	90.59	89.46	0.7895	94.98
Our DNN	89.89	**92.13**	**91.01**	**0.8204**	**96.48**
DNN reduced amino acid	88.66 (±4.06)	90.47 (±3.05)	89.57 (±0.94)	0.7938 (±0.02)	96.13 (±0.32)
DNN random amino acid	81.00 (±5.95)	81.64 (±7.73)	81.32 (±3.19)	0.6310 (±0.06)	89.55 (±2.55)
gkmSVM reduced amino acid	87.92	87.64	87.78	0.7556	94.16
gkmSVM random amino acid	80.02 (±3.77)	78.13 (±3.22)	79.07 (±3.16)	0.5819 (±0.06)	86.68 (±3.17)

Note: Recognition performance on the testing dataset is shown for state-of-the-art methods (listed in column 1) on the metrics listed in columns 2–6. Best performance on a metric is marked in bold. Our DNN model is shown in row 9. The four bottom rows show performance of the DNN model and the gkmSVM model on the DNN-reduced versus random alphabets.



Let's Play: See Ma, no Hands!

Generate Protein Structures with Deep Models



Guo, Du, Tadepalli, Zhao, and Shehu. Generating Tertiary Protein Structures via Interpretable Graph Variational Autoencoders. Bioinformatics Advances 1(1): vbab036, 2021

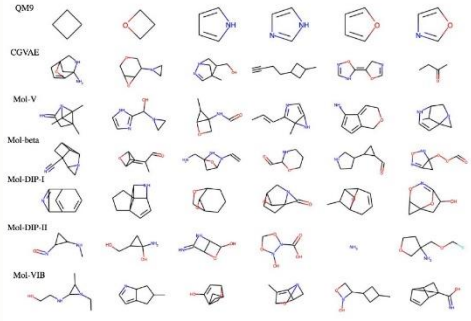
Rahman, Du, Zhao, and Shehu. Generative Adversarial Learning of Protein Tertiary Structures. Molecules 26(5): 1209, 2021.

Rahman, Du, and Shehu. Graph Representation Learning for Protein Conformation Sampling. IEEE Intl Conf on Comput Adv in Bio and Medical Sciences (ICCABS) 2021, Virtual, 2021.

Alam and Shehu. Generating Physically-Realistic Tertiary Protein Structures with Deep Latent Variable Models Learning Over Experimentally-available Structures. IEEE Intl Conf on Bioinf and Biomedicine Workshops: Computational Structural Biology Workshop, Virtual, 2021, pg. 1-8.

Alam and Shehu. Variational Autoencoders for Protein Structure Prediction. ACM Conf of Bioinf and Comput Biol, Virtual, 2020, pg. 1-10.

Generate Small Molecules with Desired Properties

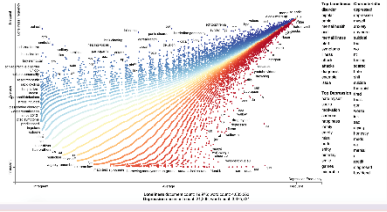


Du, Guo, Shehu, and Zhao. Disentangled Representation Learning for Interpretable Molecule Generation. Bioinformatics (under review)

Du, Guo, Shehu, and Zhao. Interpretable Molecular Graph Generation via Monotonic Constraints. SIAM Conf of Data Mining, Virtual, 2022.

Du, Wang, Alam, Lu, Guo, Zhao, and Shehu. Deep Latent-Variable Models for Controllable Molecule Generation. IEEE Intl Conf on Bioinf and Biomedicine, Virtual, 2021.

Emotions and mental health from social media text

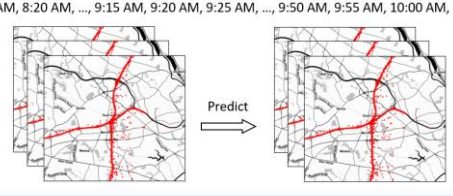


Vajre, Naylor, Kamath, and Shehu. PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis. IEEE Intl Conf on Bioinf and Biomedicine, Virtual, 2021.

Rajabi, Uzuner, and Shehu. Detecting Scarce Emotions Via BERT and Hyperparameter Optimization. Intl Conf on Artificial Neural Networks,2021.

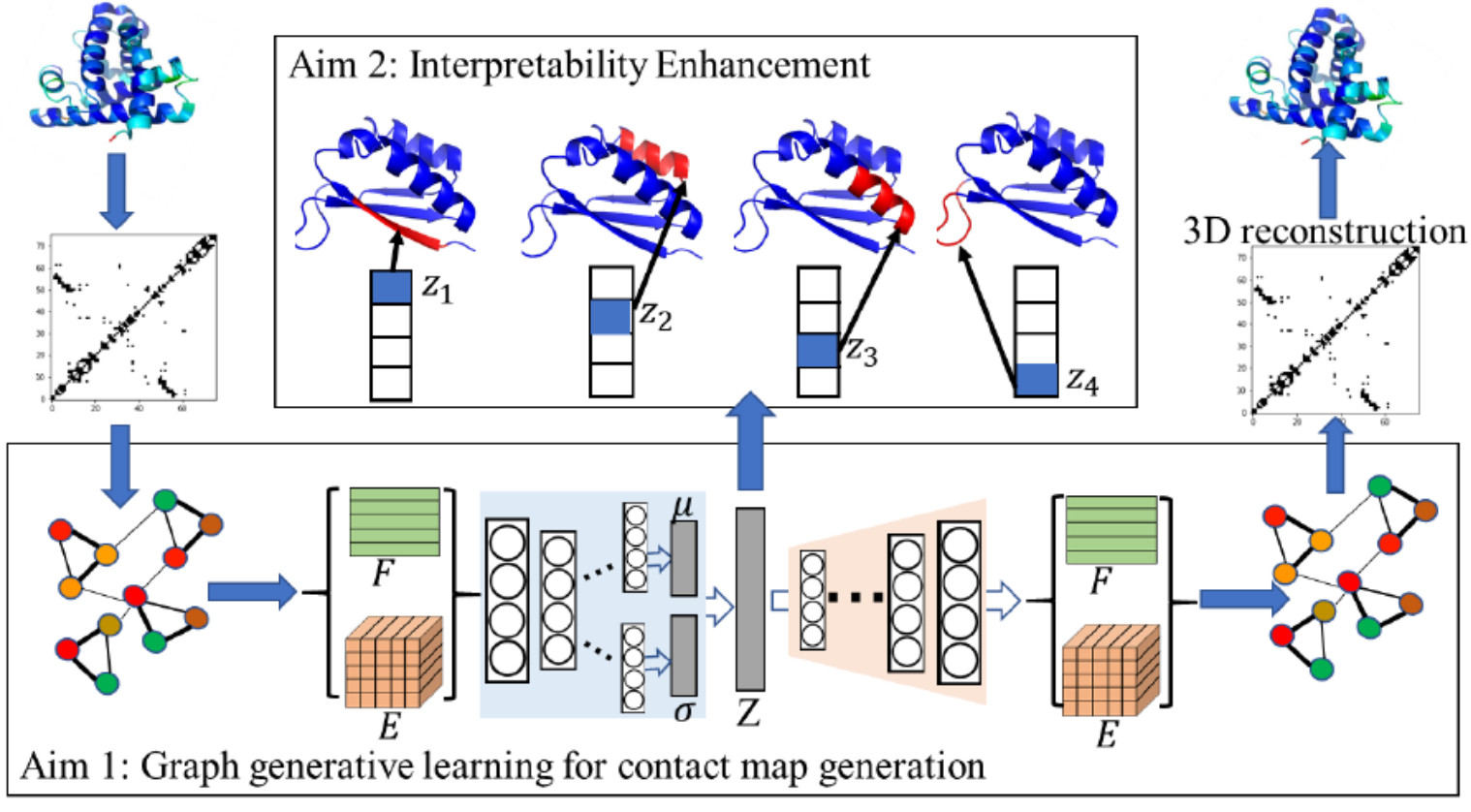
Rajabi, Uzuner, and Shehu. A Multi-channel BiLSTM-CNN Model for Multilabel Emotion Classification of Informal Text. IEEE Intl Conf on Semantic Computing, 2020.

Forecast Traffic Speed

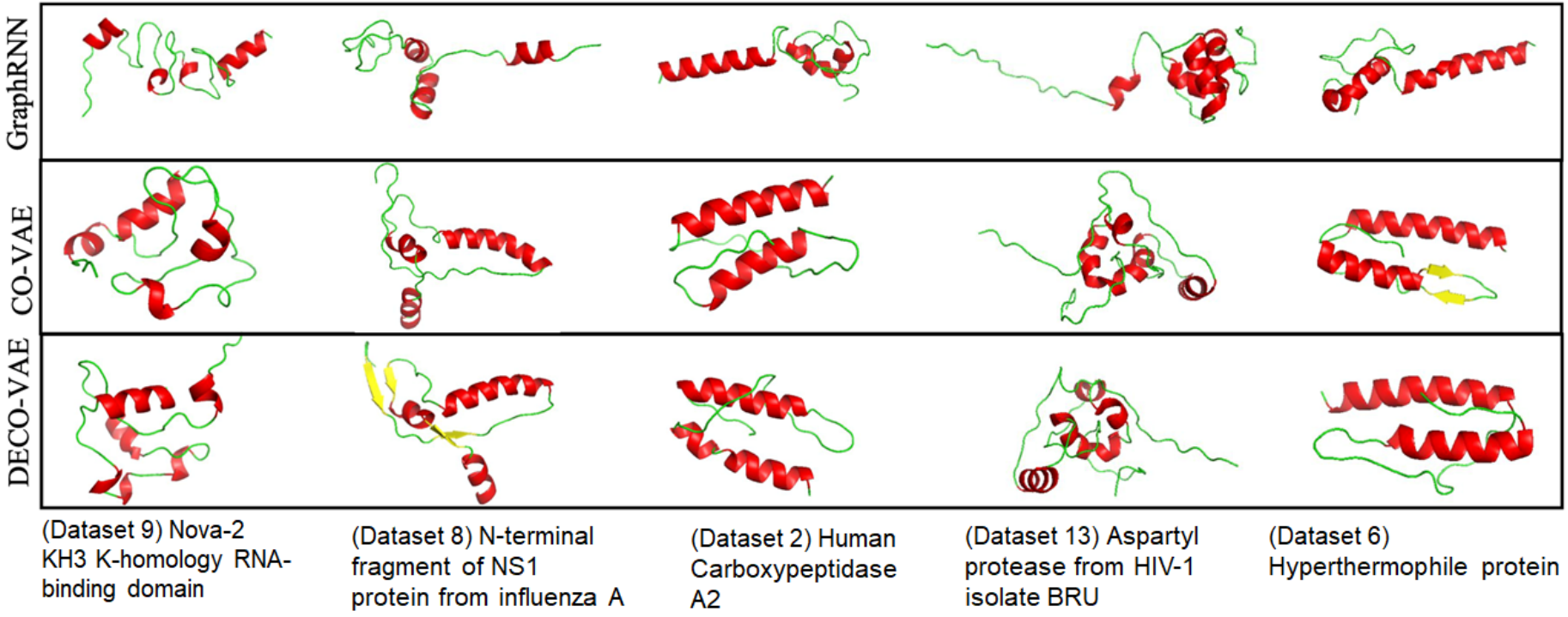


Lu, Kamranfar, Lattanzi, Shehu. Traffic Flow Forecasting with Maintenance Downtime via Multi-Channel Attention-Based Spatio-Temporal Graph Convolutional Networks. IEEE Transactions on Intelligent Transportation Systems (under review).

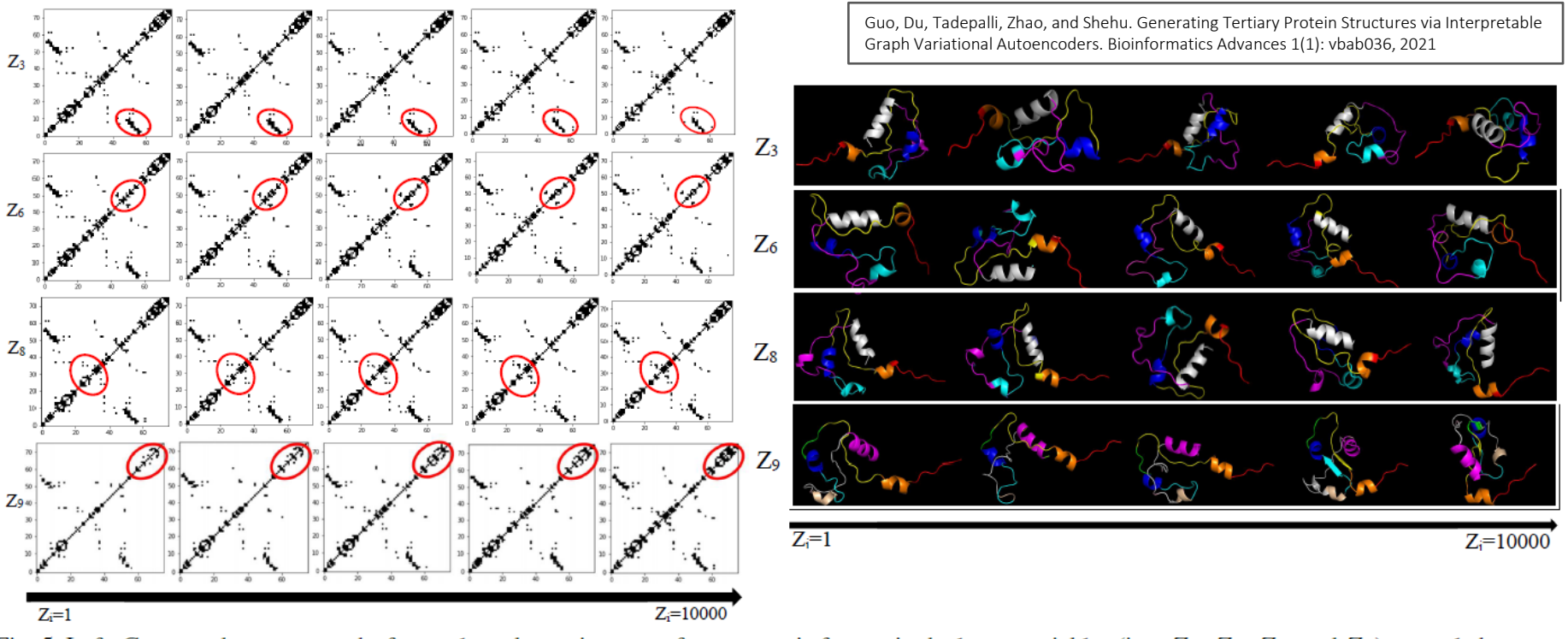
Disentangled Graph-based VAEs for Protein Structure Representation Learning



Physically-realistic Structures, Outperforms GraphVAE, GraphRNN, Graphite, etc.



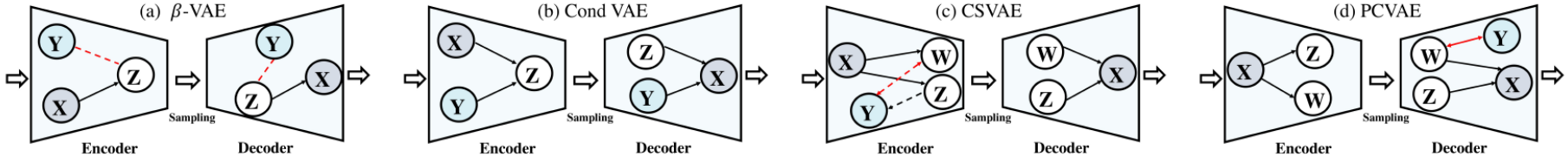
Latent Factors Control Structural Changes



Guo, Du, Tadepalli, Zhao, and Shehu. Generating Tertiary Protein Structures via Interpretable Graph Variational Autoencoders. Bioinformatics Advances 1(1): vbab036, 2021

Fig. 5: Left: Generated contact graphs for a selected protein target; four semantic factors in the latent variables (i.e., Z_3, Z_6, Z_8, and Z_9) control changes in the contact graphs; the value of latent variables changes from 1 to 10000; Right: corresponding reconstructed tertiary structures.

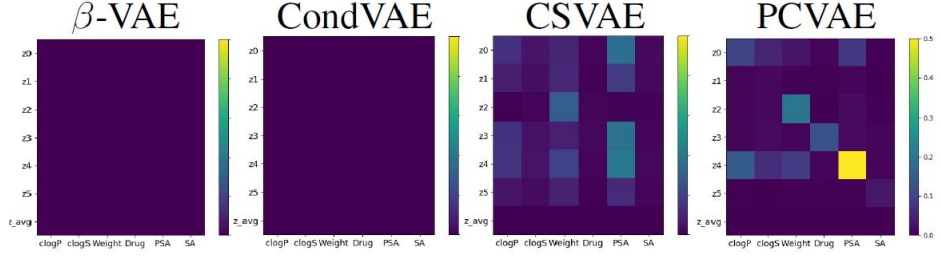
Graph-based VAE Models for Generating Small (Drug-like) Molecules w/ Property Control



Each sub-figure depicts the generative model (decoder) and its model inference (encoder). The enforcement of independence is shown by dotted red arrows, whereas the invertible dependence between two variables is represented by double arrows. Data is denoted by X and Z. W are subsets of latent variables, and Y denotes the molecular properties (cLogP, cLogS, PSA, SA, Weight, and Drug-likeness).

	QM9			ZINC			MOSES		
Model	Validity	Novelty	Uniqueness	Validity	Novelty	Uniqueness	Validity	Novelty	Uniqueness
β-VAE	**100.00%**	98.23%	**99.28%**	**100.00%**	**100.00%**	**99.78%**	**100.00%**	99.92%	**99.88%**
CondVAE	**100.00%**	92.60%	90.00%	**100.00%**	99.98%	98.02%	**100.00%**	99.98%	93.30%
CSVAE	**100.00%**	97.01%	27.41%	**100.00%**	**100.00%**	42.72%	**100.00%**	**100.00%**	54.28%
PCVAE	**100.00%**	97.43%	88.24%	**100.00%**	**100.00%**	99.48%	**100.00%**	99.96%	98.62%

Models generate valid, novel, and unique molecules. ***Validity*** measures the fraction of generated molecules that are chemically valid. ***Novelty*** measures the fraction of generated molecules that are not in the training dataset. ***Uniqueness*** measures the fraction of generated molecules after and before removing duplicates.



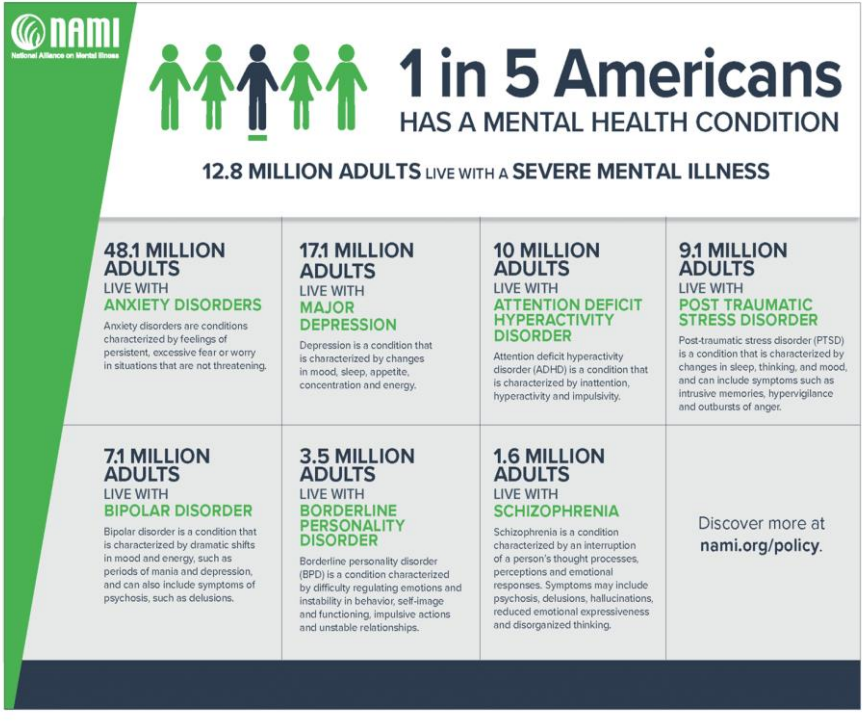
Mutual Information values between each disentangled factor learned by a (QM9-trained) model and properties computed on molecules generated by the model show several models affording better property control.

The models leverage both graph representation learning to learn inherent constraints in the chemical space and inductive bias to connect the chemical and biological space. Promising step in controllable molecule generation in support of cheminformatics, drug discovery, etc.

DL & NLP for Mental Health: Transformer-based Classification Models

- Depression alone is estimated to affect more than 300 million people worldwide, but approximately 35% of adults let their depressive symptoms go untreated.
- Most Interactions happen on social media (Twitter, Reddit, Facebook, Instagram, SnapChat,..)
- Social Media communication can be an indicator of symptoms/signs of mental health
- Natural language processing and machine learning can be used for detection of these symptoms.

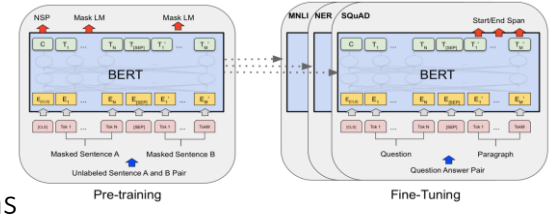
Key Idea: Fine-tune a pre-trained language model (BERT) to learn representations of words related to mental health, and then add classification layer



DL & NLP for Mental Health: Transformer-based Classification Models

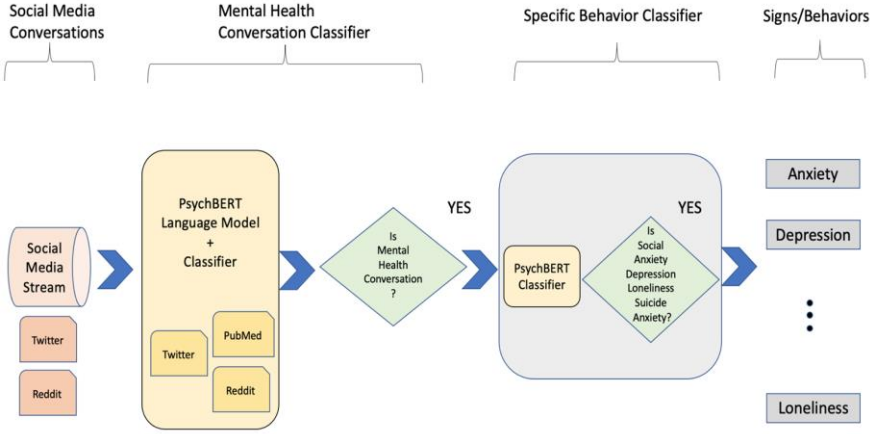
Contextual representations: (SOTA)

- BERT is a multi-layer bidirectional Transformer encoder
- BERT generates word representations by considering both positional and contextual information.



Fine-tune pre-trained BERT over relevant social media text and PUBMED psychology and neuropsychiatry journals

Social Media Sources	Size	Traits
Twitter hashtags	19,943	depression
Twitter hashtags	21,208	social anxiety
Twitter hashtags	19,975	loneliness
Subreddit r/anxiety	11,544	anxiety
Subreddit r/mentalhealth	18,924	mental illness
Twitter hashtags	2,344	depression
Subreddit r/suicidewatch	12,276	suicide
Subreddit r/jokes	30,786	non mental health
Subreddit r/meditation	25,743	non mental health
Subreddit r/parenting	49,684	non mental health



- Stage 1: PsychBERT language model & a binary classifier used to classify and split into mental health/non mental health
- Stage 2: Takes only mental health-classified data as input and uses PsychBERT and a multi-class classifier to separate into 6 classes

DL & NLP for Mental Health: Transformer-based (Interpretable) Model

Mental vs. Non-Mental Health

Category	Classifier	F1
Traditional	Naive Bayes	0.91
	Logistic Regression	0.95
	Decision Tree	0.73
Interpretable	Boosted Rules	0.88
	DL8.5	0.94
	EBM	0.88
Deep Learning Transformers	Kim CNN	0.94
	LSTM	0.95
	PsychBERT	**0.98**

Multi-class Classification

Category	Classifier	F1
Traditional	Naive Bayes	0.41
	Logistic Regression	0.36
	Decision Tree	0.33
Interpretable	Boosted Rules	0.44
	DL8.5	0.47
	EBM	0.39
Deep Learning Transformers	Kim CNN	0.57
	LSTM	0.51
	PsychBERT	**0.63**

Local explanations for PsychBERT on a true positive (top) and true negative (bottom) via feature attribution (integrated gradients in Captum library)

[CLS] Getting worried I ' ve had su ##icidal thoughts since I was 17 I ' m 23 now . Every year the thoughts get worse . 2018 has been the most challenging year of my life with family my girl my friends I have lost everyone had people who I gave my trust to betray me and now I can ' t trust anyone . I have no one who cares about me or checks up on me . Maybe I just need to get used to being alone , but the past few days I ' ve found myself planning to kill myself without trying . It just happens I see it in my head . I ' m not sure if this is a phase or if I need help . [SEP]

[CLS] What ' s the best strength training I can do at home without equipment ? I ' ve been go ##og ##ling and this comes up as the top result : 20 body weight sq ##ua ##ts . 10 push ups . 20 walking lung ##es . 10 dumb ##bell rows (using a gal ##lon milk j ##ug) 15 second plan ##k . 30 jumping Jack ##s . Rep ##eat for 3 rounds . Is this a decent enough home routine ? I ' m trying to switch from card ##io to strength training (or trying to incorporate strength training at least) after reading the article from the begin ##ner guide on the side ##bar , and wanted to look for a good routine to follow . [SEP]

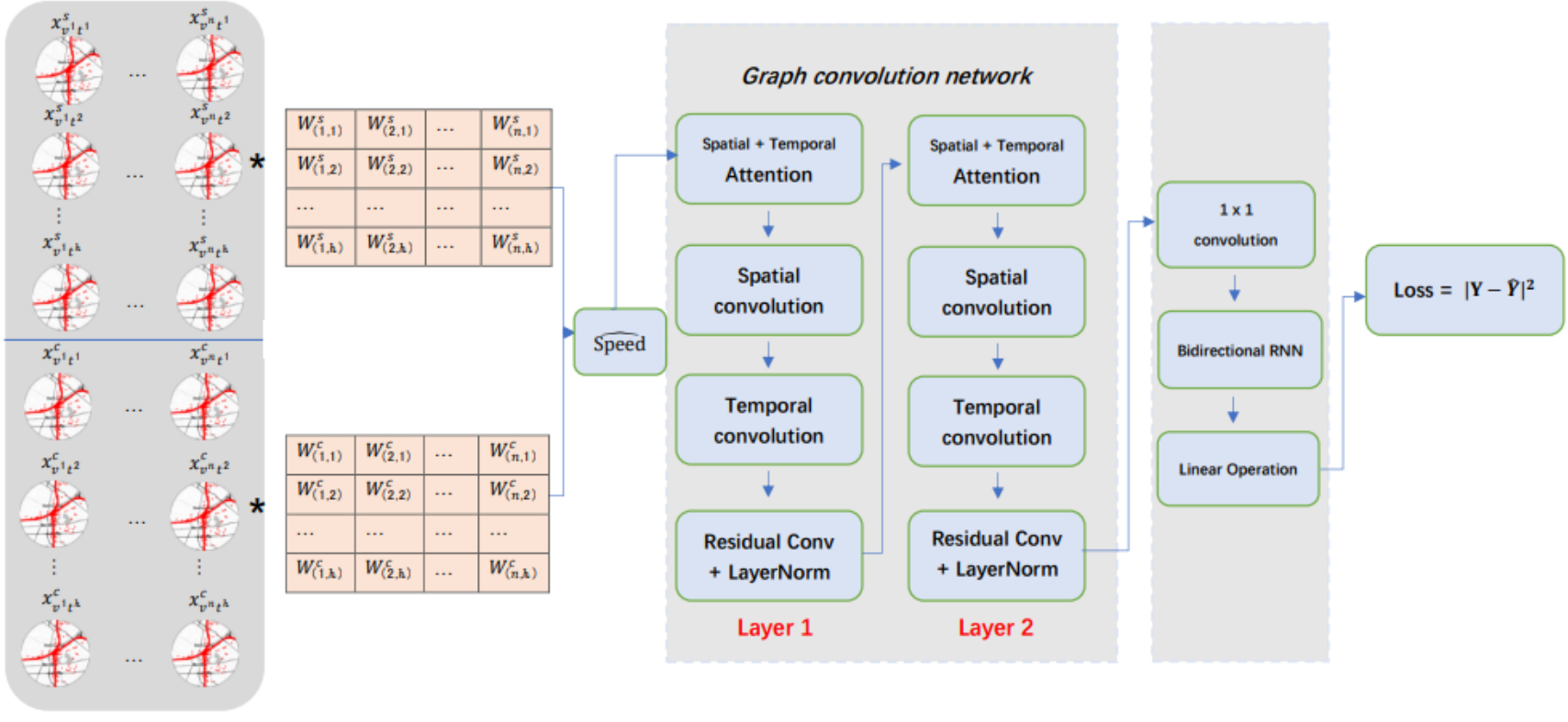
Traffic Speed Forecasting with Work Zones

Road segments: Using a sensor in each area to record the location of the current road conditions and vehicle speed

Tyson's Corner in Fairfax, Virginia

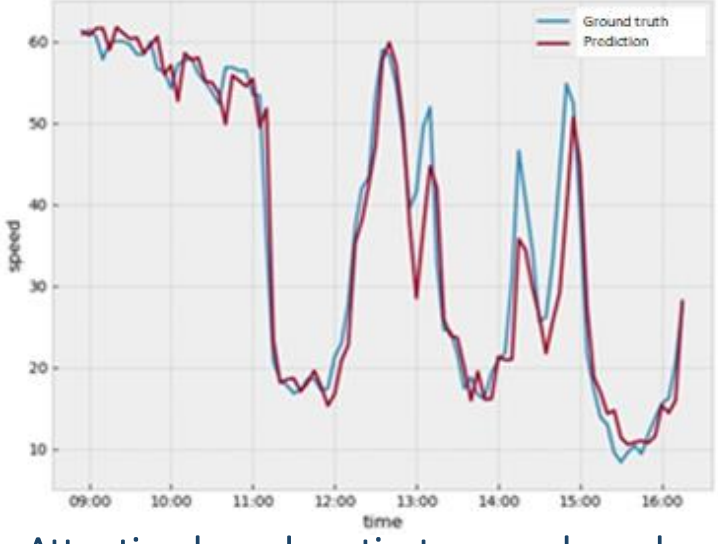
- 131 road segments, include interstate highway, Virginia state route, etc.
- 12 months timeframe (2019)
- 10 traffic attributes
- 478 construction work events
- 10677 rows of traffic speed and construction work information

Traffic Speed Forecasting with Work Zones via Spatial and Temporal Attention

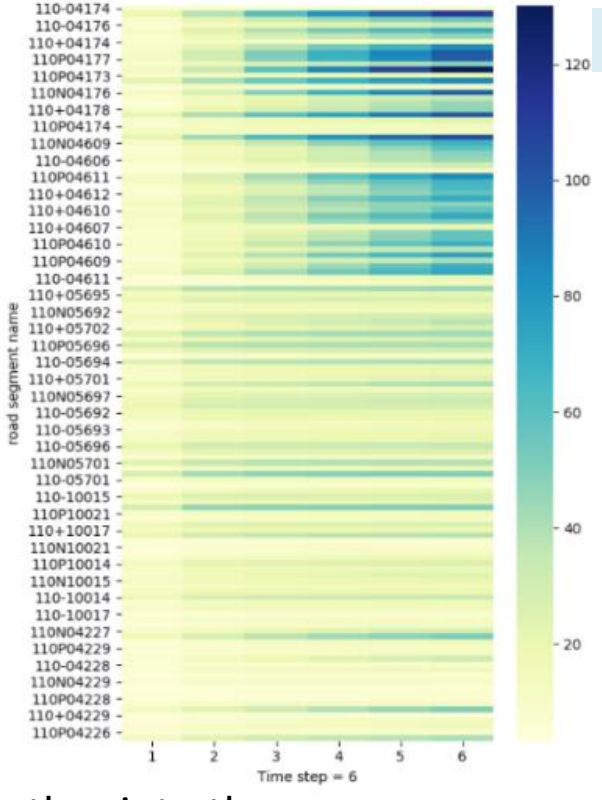


Traffic Speed Forecasting with Work Zones via Spatial and Temporal Attention

Predicting Traffic Speed in the Presence of Construction (project between Mason Center of Transportation and VDOT)



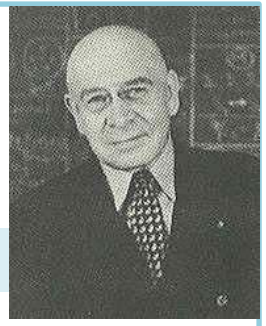
15min+ : Good
30min+: Fair
60min+: Bad



Attention based spatio-temporal graph convolutional network performs well

Challenge: predictions further into the future become increasingly unreliable

The Map is not the Territory



Alfred Korzibsky
1879-1950

Science and Sanity *(1933)*

Greek, pre-scientific (idealism)
Observer-centric; observed reality does not matter

Classical scientific (materialistic)
Observed reality-centric; observer does not matter

Modern scientific (pre-digital) [no to small data]
Human knowledge depends on *both* the observed reality and the observer

... - 2010

Modern digital (data-centric) [big data → Largely Deep Learning]
Human knowledge entirely derived from the data
Observed reality and the observer do not matter

2010 - …

?

Model-based Research
"Classical AI"
Explicit Knowledge

Data Analytic Modeling
"Machine Learning"
Tacit Knowledge

Search, Optimization, Planning

Machine Learning, Deep Learning

- We know how to do optimization
- We have done it for decades
- We understand *loss/objective* functions
- We understand multiple objectives
- We understand exploration versus exploitation deeply
- We know how to design optimization algorithms to high-dimensional, non-linear variable spaces

Problems ←

Solutions →

- DL literature current version of gold rush
- Ad-hoc approaches
- Confounding of terms
  - Multi-objective for aggregated functions
- Heuristics abound
- Papers on +.5 improvements! (SOTA chasing)
- Need better connection between engineering exercise and foundational computing research

Wishes for Scientific Discovery, Innovation, and Education

- Increasingly, all our students want to do is deep learning to respond to the market
  - Promotes group think and narrows scientific ingenuity and discovery
  - Important to train students in interdisciplinary setting [AI/ML + X]

- Vast uncharted territory for AI/ML-based discoveries
  - Algorithmic-mediated society
  - *Challenging problems → foundations of AI/ML*

- Scientific AI & ML frameworks over brute-force engineering facilitated by big data
  - Polanyi's Paradox → Polanyi's Revenge [Kambhampati. CACM (61):2, 2021]

Wishes for Scientific Discovery, Innovation, and Education

- Increasingly, all our students want to do is deep learning to resp
  - Promotes group think and narrows scientific ingenuity and
  - Important to train students in interdisciplinary setting [AI/
- Vast uncharted territory for AI/ML-based discoveries
  - Algorithmic-mediated society
  - *Challenging problems → foundations of AI/ML*
- Scientific AI & ML frameworks over brute-force engineering fac
  - Polanyi's Paradox → Polanyi's Revenge [Kambhampati. CA

"Human, grant me the serenity to accept the things I cannot learn, data to learn the things I can, and wisdom to know the difference."

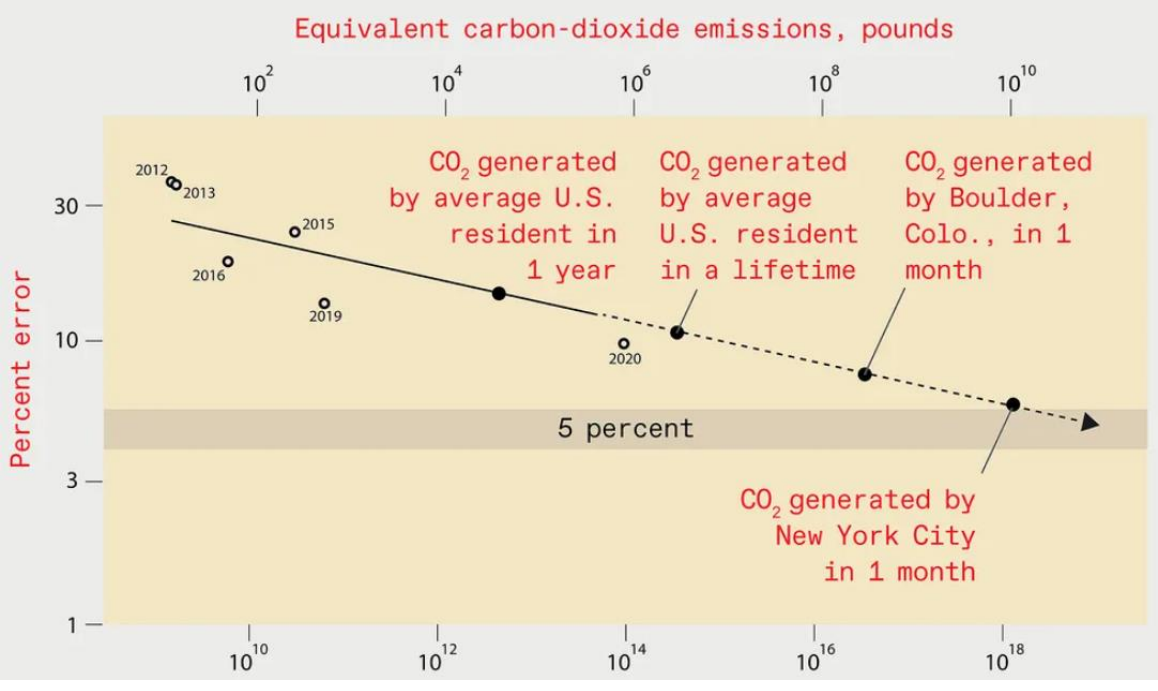
Wishes for Scientific Discovery, Innovation, and Education

- Increasingly, all our students want to do is deep learning to respond to the market
  - Promotes group think and narrows scientific ingenuity and discovery
  - Important to train students in interdisciplinary setting [AI/ML + X]

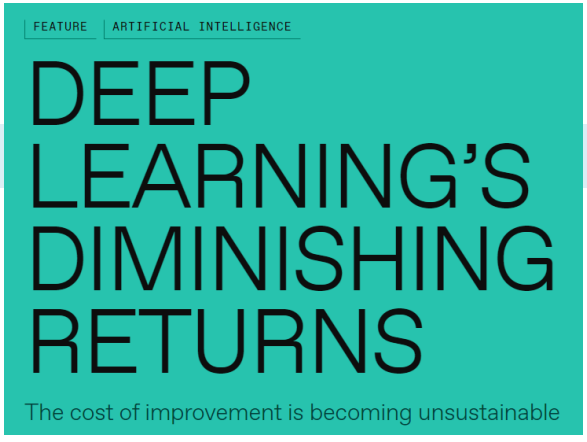
- Vast uncharted territory for AI/ML-based discoveries
  - Algorithmic-mediated society
  - *Challenging problems → foundations of AI/ML*

- Scientific AI & ML frameworks over brute-force engineering facilitated by big data
  - Polanyi's Paradox → Polanyi's Revenge [Kambhampati. CACM (61):2, 2021]
  - DL models do not generalize well and are unsustainable

Wishes for Scientific Discovery, Innovation, and Education



NEIL C. THOMPSON KRISTJAN GREENEWALD KEEHEON LEE GABRIEL F. MANSO. IEEE Spectrum. 24 SEP 2021



By 2025, the error level in the best deep-learning systems recognizing objects in the ImageNet data set should be reduced to just 5 percent.

But the computing resources and energy required to train such a future system would be enormous, leading to the emission of as much carbon dioxide as New York City generates in one month SOURCE: N.C. THOMPSON, K. GREENEWALD, K. LEE, G.F. MANSO

Students Behind Highlighted Work

Uday Kamath
Now chief analytical officer at Digital Reasoning

Daniel Veltri
Now Bioinformatics Lead at NIH-NIAID

Kevin Molloy
Now Assistant Professor at James Madison University

Jenniffer Van (Radel)
Now Machine Learning Engineer at Rebellion Defense
420 Council

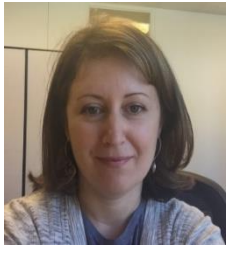
Nasrin Akhter
Now Assistant Professor of Teaching at University of Buffalo

Brian Olson
Now Engineering Manager, Machine Learning at LinkedIn

Irina Hashmi
Now Assistant Professor of Teaching at George Mason University

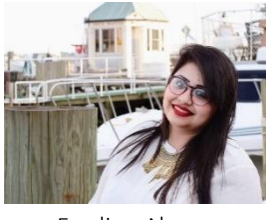
Ahmed Bin Zaman
Now Assistant Professor of Teaching at George Mason University

Emmanuel Sapin
Now Research Associate at University of Colorado Boulder



Tatiana Maximova
Now Research Associate at Ben Gurion University

Zahra Rajabi
Now Machine Learning Engineer @ Adobe

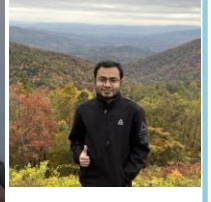


Fardina Alam

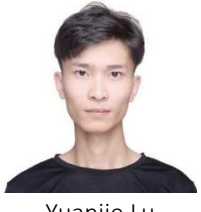
Manpriya Dua

Lutful Kazi Kabir

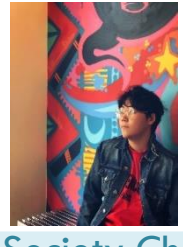
Parastoo Kamranfar



Taseef Rahman



Yuanjie Lu



Yuanqi Du
Undergraduate student heading to Cornell

Vedant Vajre
Senior at Stone Bridge High School

Acknowledgements

Highlighted work supported in part by:

- ❑ DoD Minerva Program
- ❑ NSF IIS:III, CCF:AF, CCF:FET, and OAC:SSE Programs
- ❑ Jeffress Trust Award Program
- ❑ Virginia Youth Tobacco Program
- ❑ NIH National Cancer Institute
- ❑ INOVA Translational Medicine Institute

Highlighted work in collaboration with:

- ❑ Liang Zhao (Emory University)
- ❑ Ruth Nussinov, Buyong Ma (NIH NCI)
- ❑ Daniel Barbarà, Kenneth De Jong, Daniel Lattanzi, Erion Plaku, Wanli Qiao, Ozlem Uzuner (George Mason University)

- ❑ Papers and accompanying software at cs.gmu.edu/~ashehu
- ❑ Contact for any inquiries: amarda@gmu.edu