Parallel Processing & Applied Mathematics

Ostrava, Czech Republic, September 8-11, 2024

POLITECHNIKA Czestochowa University of Technology

VSB TECHNICAL UNIVERSITY OF OSTRAVA

# Strengthening AI to Enable Scientific Discovery


Amarda Shehu, PhD

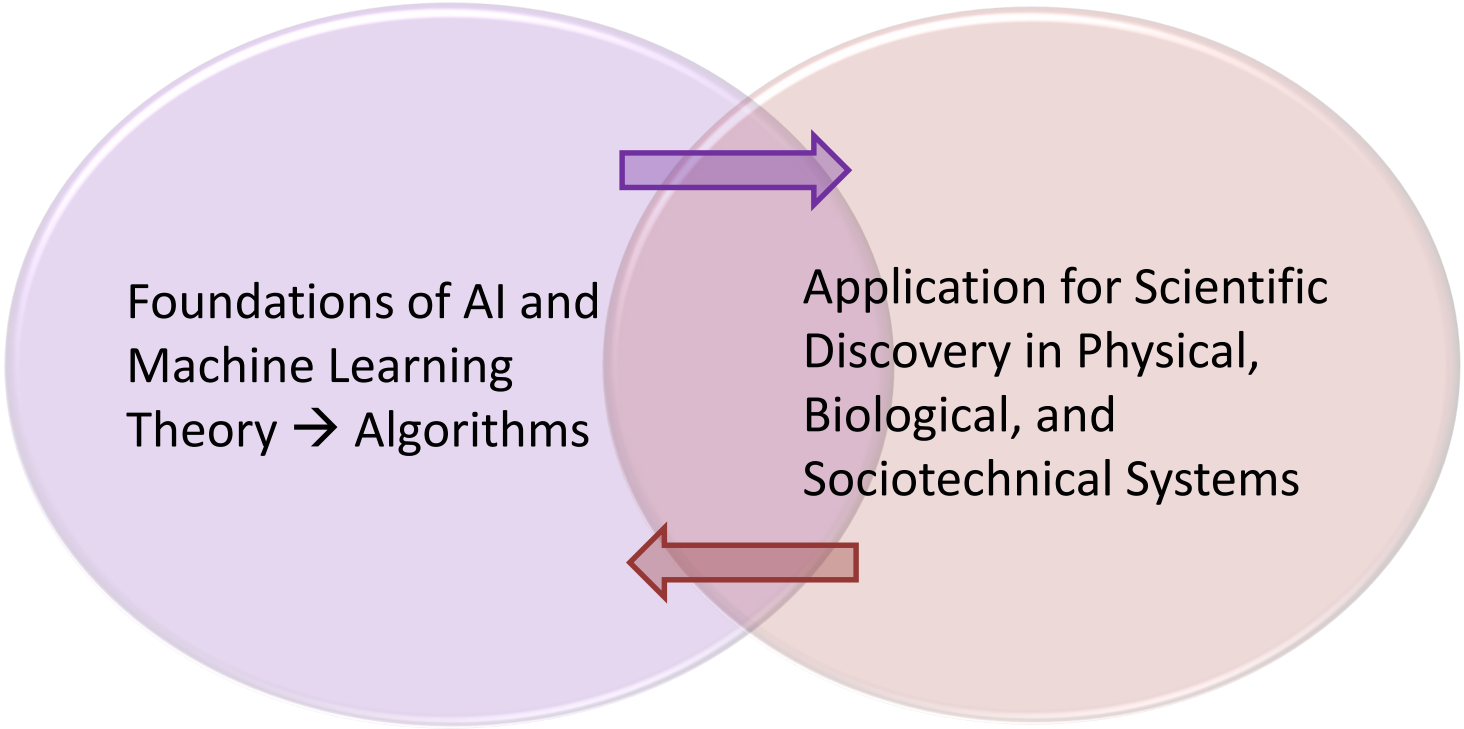
*Inaugural Vice President and Chief AI Officer,*

*Associate Dean for AI Innovation, College of Engineering and Computing*

*Professor of Computer Science*

GEORGE MASON UNIVERSITY

# 10K Mile-up View: *4AI and AI4*

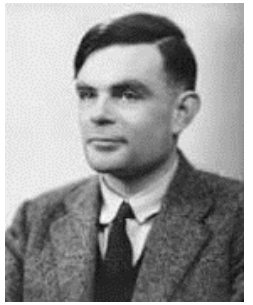



NSF
Jeffress Trust Awards Program
Health Resources in Action
Advancing Public Health and Medical Research
Commonwealth of Virginia Alzheimer's & Related Diseases Research Award Fund
Virginia Foundation for Healthy Youth
Department of Defense Minerva Program
Department of Education United States of America
SBA U.S. Small Business Administration
Air Force Research Laboratory

# Function → Form → Function-encoding Representation of Form

*"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.

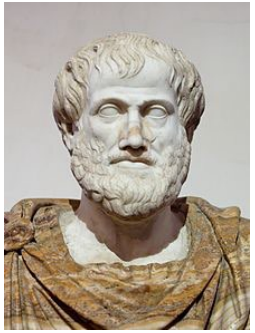
**How does the form of matter determine function?**



Sir Alan Turing
1912-1954

Guiding Principle

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



Roman copy in marble of Greek bronze bust of Aristotle by Lysippos, c. 330 BC

*"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 (ith approximative) Function-encoding Representation of Form***

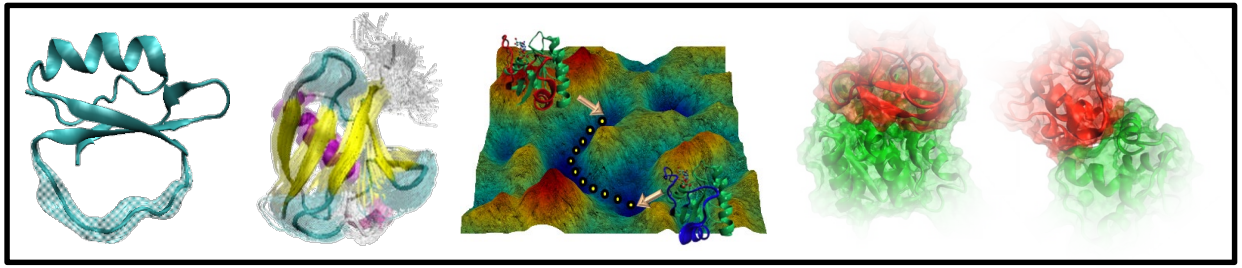
# Representation Learning in my Lab over the Years

**No to very little data – explicit knowledge**

**2002-2016**

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

-- **molecular structural biology**

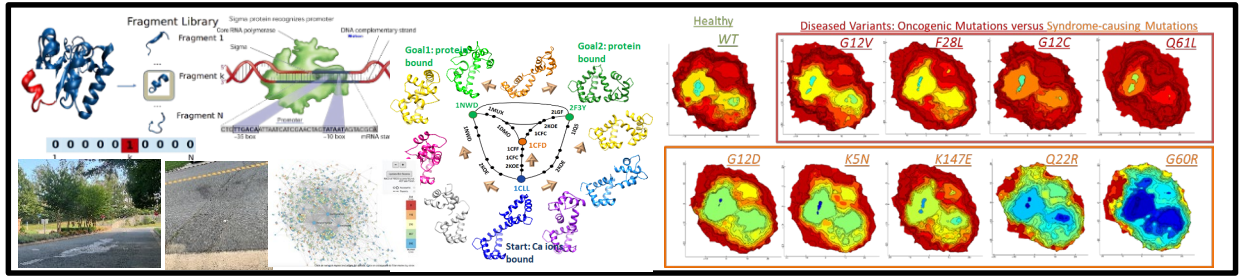


**Some data –explicit and tacit knowledge**

**2010-**

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

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

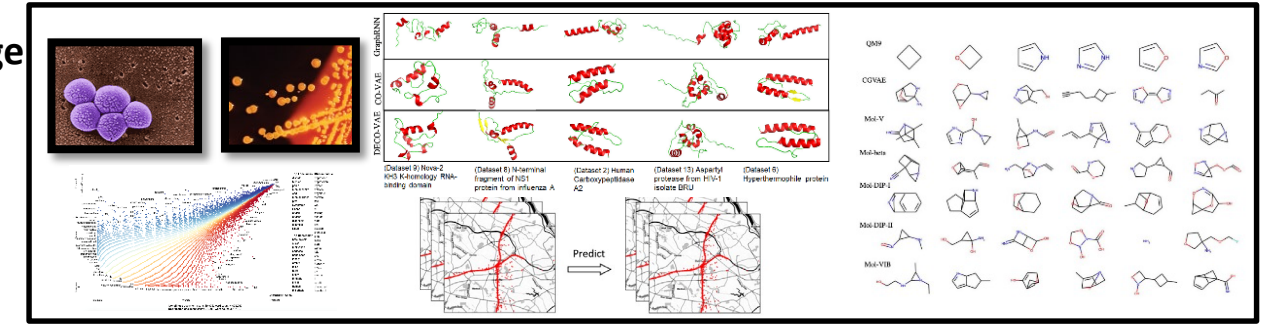


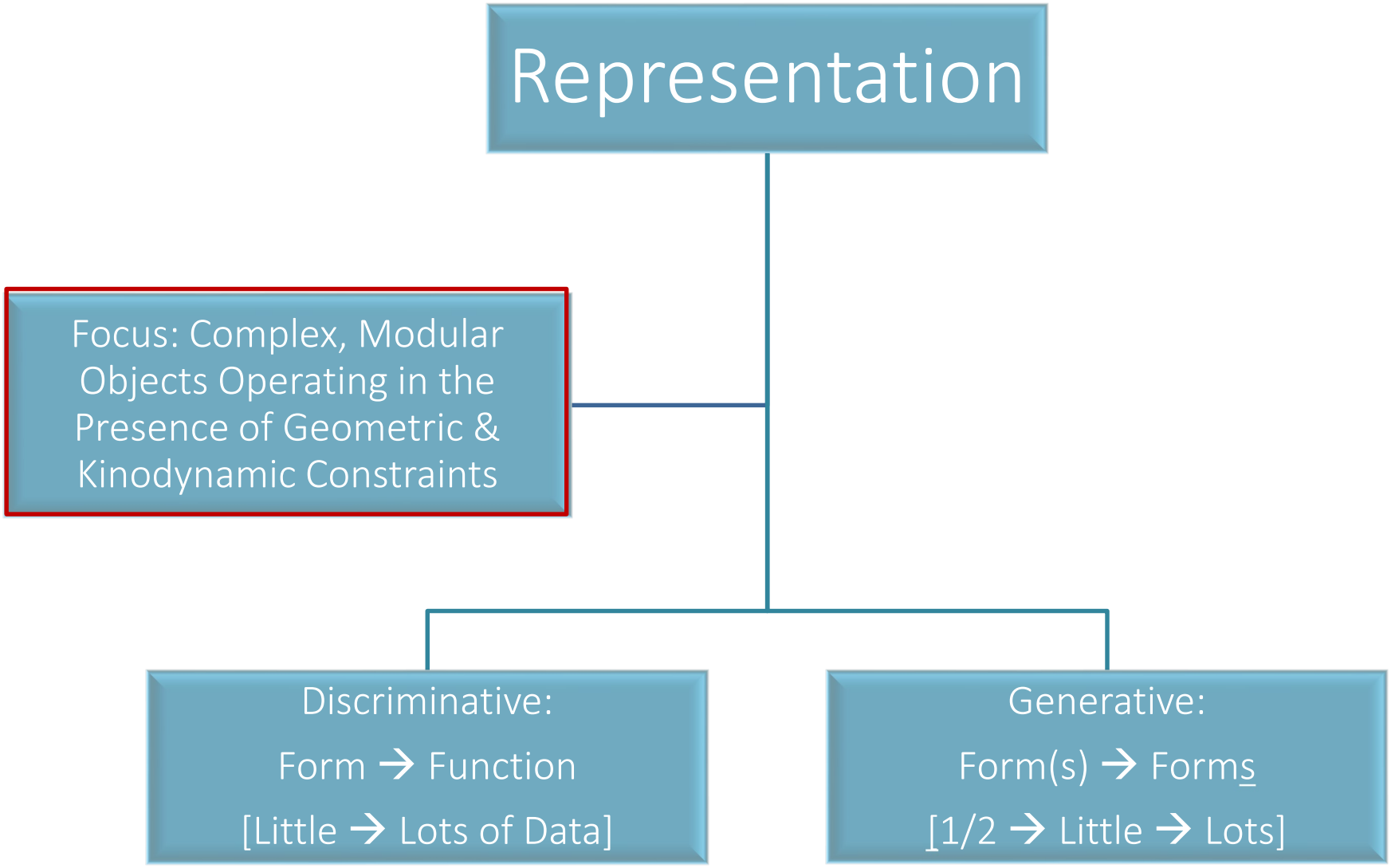

**Lots of data – AI romance w/ tacit knowledge**

**2018-**

Deep Learning, NLP, Deep generative models

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

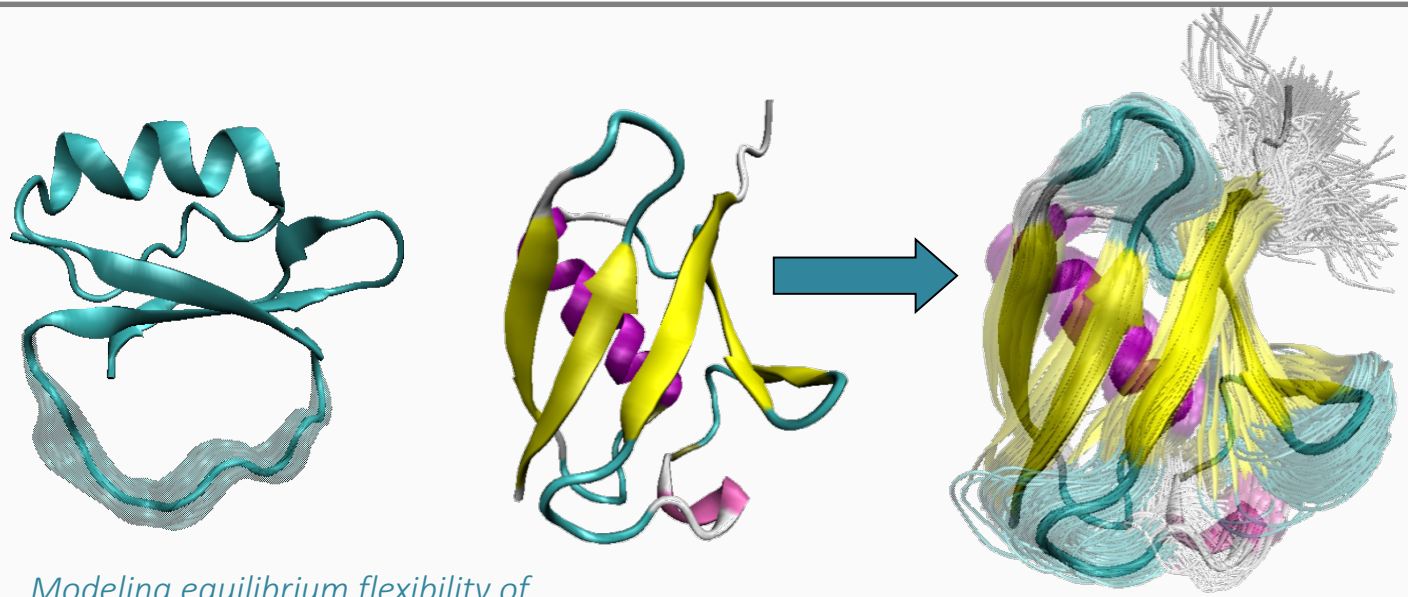



Representation
Focus: Complex, Modular Objects Operating in the Presence of Geometric & Kinodynamic Constraints
Discriminative:
Form → Function
[Little → Lots of Data]
Generative:
Form(s) → Forms
[1/2 → Little → Lots]

Generative AI before NNs, DL
Good old-fashioned configuration sampling, searching, planning

Focus: Complex, Modular Objects Operating in the Presence of Geometric & Kinodynamic Constraints

Discriminativ
Form → Funct
[Little → Lots of



*Modeling equilibrium flexibility of highly-mobile segments*

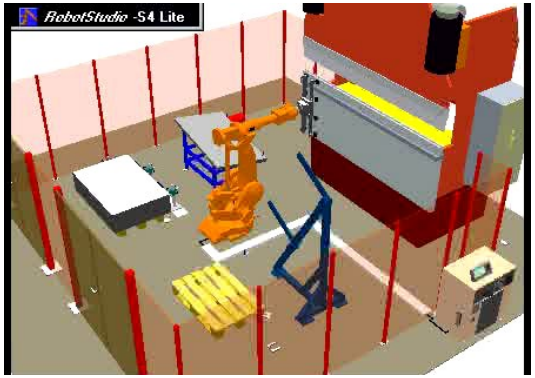
*Modeling equilibrium flexibility of entire protein chain*

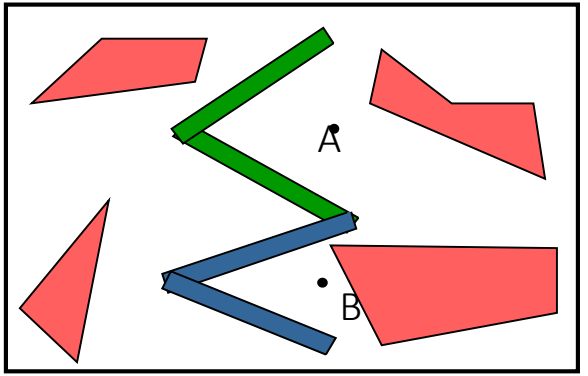
Goal: Partial or full characterization of protein flexibility by combining fast molecular kinematics (inspired from robotics/geometry of articulated objects) with physics-based treatments (molecular mechanics).

Think deeply about coupled and uncoupled degrees of freedom to control dimensionality of the search space

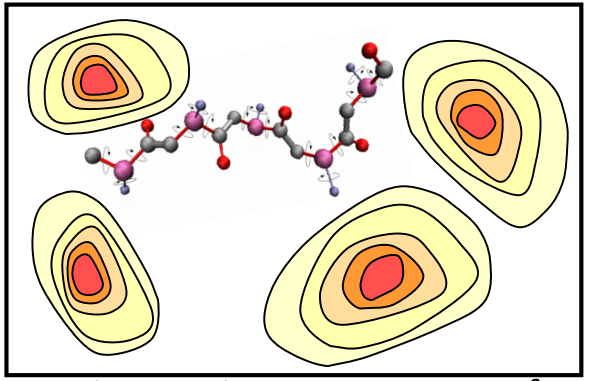
Algorithmica; Proteins: Structure & Function; Protein Science; Biophysics J; …

# ROBOT → MACROMOLECULAR MOTION PLANNING

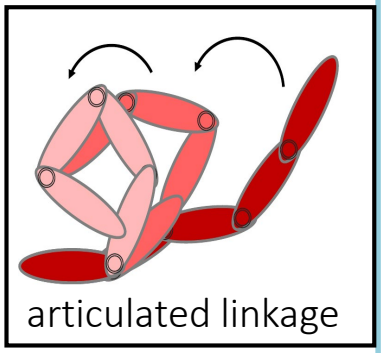



articulated robot: 0/1 obstacles



protein: continuous energy surface

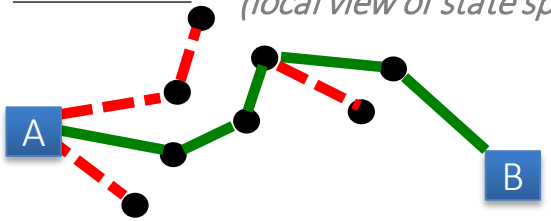


articulated linkage

---

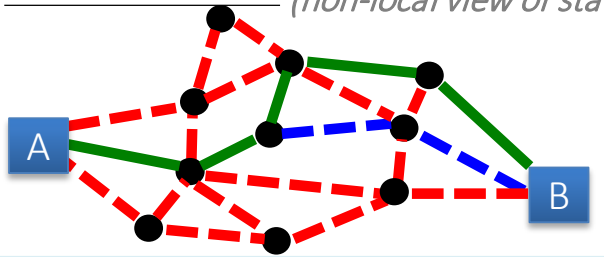
Tree-based

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

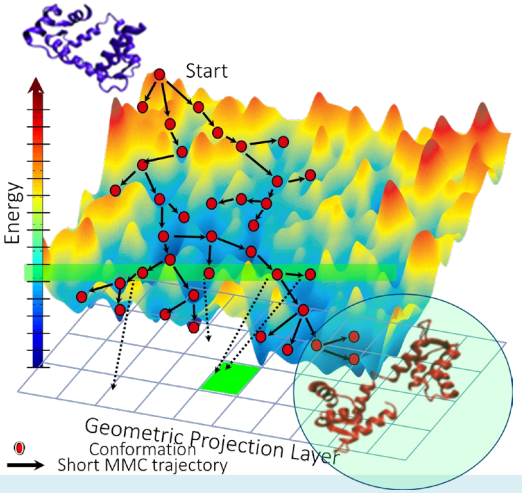



Roadmap-based

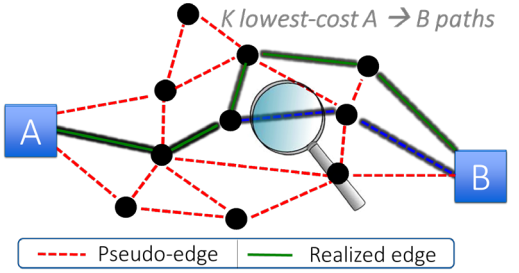
*Build roadmap in state space (non-local view of state space)*



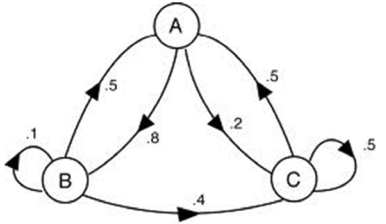

*Adaptive Search Guided by Projection-based Discretization Layers to Constraint-Satisfying Regions*




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




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

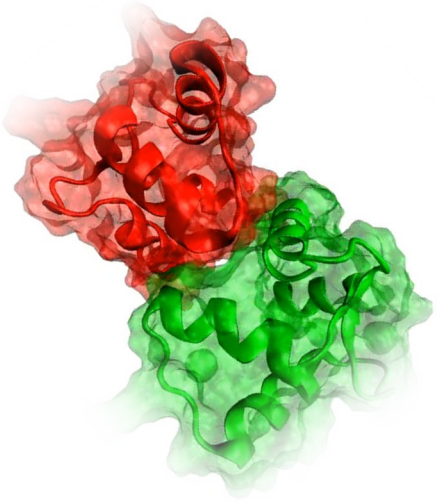



How not to get lost in hundred-dof space and apportion compute

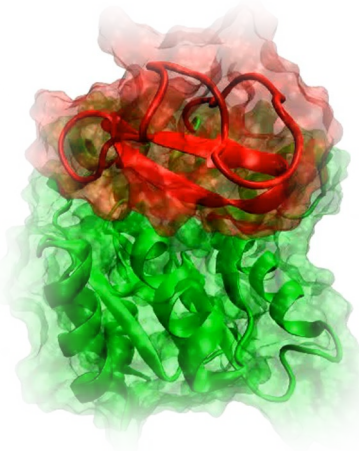
Connect summary statistics with wet-lab kinetic measurements

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

Calmodulin

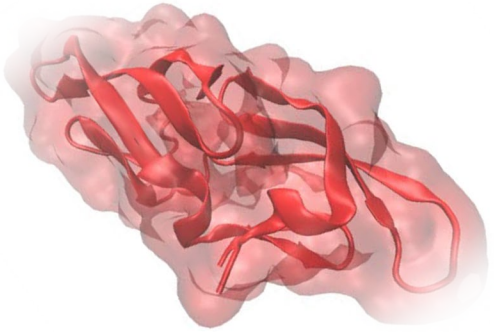


Adenylate Kinase



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

Cyanovirin-N



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


Robotica
IEEE/ACM Trans Comput Biol
BMC Struct Biol
Proteome Sci
J Artif Intel Res
Int J Robot Res
Robotics Sci & Sys
PLoS Comput Biol
J Comput Biol

# Representation

Focus: Complex, Modular Objects Operating in the Presence of Geometric & Kinodynamic Constraints

Discriminative:

Form → Function

[Little → Lots of Data]

Generative:

Form(s) → Forms

[1/2 → Little → Lots]

# Strengthening AI: Discriminative, Little Data Setting

Key Question: How does our understanding of the domain, of the problem, of the underlying process inform representation learning?

Function-encoding representations of form

In Discriminative Setting: Guide Feature Engineering with Prior Knowledge

# Strengthening AI: Discriminative, Little Data Setting

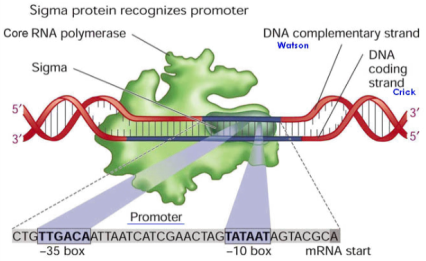
Key Question: How does our understanding of the domain, of the problem, of the underlying process inform representation learning?

<u>Function-encoding representations of form</u>

In Discriminative Setting: Guide Feature Engineering with Prior Knowledge

**Prediction Task: Sequence → Function**

**Representation Q: What Constraints does Function Impose on Sequence?**




Key insight: encode implicit constraints in linear representation

- <u>Non-local/distal</u> constraints
- Capture them as features (compositional, positional, correlational, …)
- Explicit, *interpretable* answer to how sequence encodes function

# Strengthening AI: Discriminative, Little Data Setting

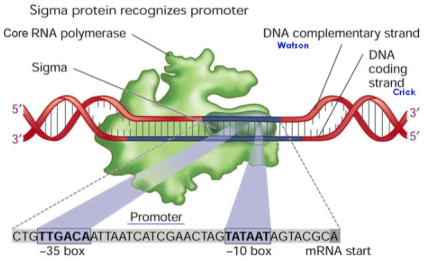
Key Question: How does our understanding of the domain, of the problem, of the underlying process inform representation learning?

<u>Function-encoding representations of form</u>

In Discriminative Setting: Guide Feature Engineering with Prior Knowledge

**Prediction Task: Sequence → Function**

**Representation Q: What Constraints does Function Impose on Sequence?**




Key insight: encode implicit constraints in linear representation

- <u>Non-local/distal</u> constraints
- Capture them as features (compositional, positional, correlational, …)
- Explicit, *interpretable* answer to how sequence encodes function

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

# Strengthening AI: Discriminative, Little Data Setting

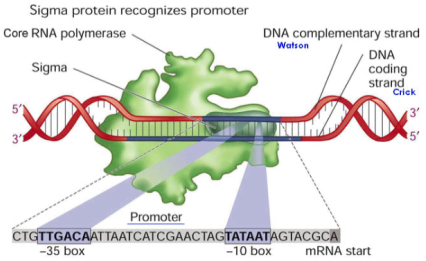
Key Question: How does our understanding of the domain, of the problem, of the underlying process inform representation learning?

Function-encoding representations of form

In Discriminative Setting: Guide Feature Engineering with Prior Knowledge

**Prediction Task: Sequence → Function**

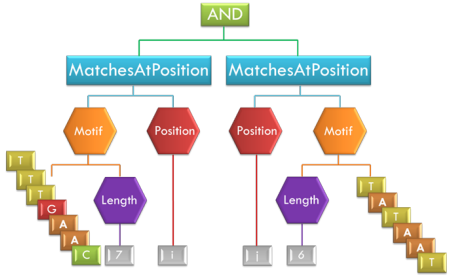
**Representation Q: What Constraints does Function Impose on Sequence?**




Key insight: encode implicit constraints in linear representation

- Non-local/distal constraints
- Capture them as features (compositional, positional, correlational, …)
- Explicit, *interpretable* answer to how sequence encodes function

- 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 | | Operators |
| 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} | Terminals |
| ERC-int | | Integer | | |

Productions: AND, OR, NOT

Problem: Exponential explosion of feature space!

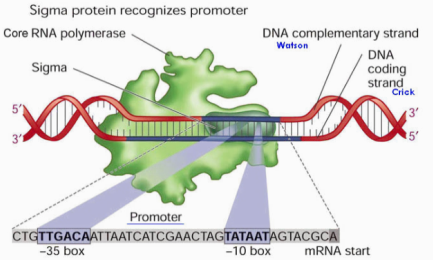
- Contribution #2
  - Genetic Programming for feature space exploration
  - Surrogate fitness function instead of wrapper classification model
  - Complete treatment in classification context

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

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


## Prediction Task: Sequence → Function

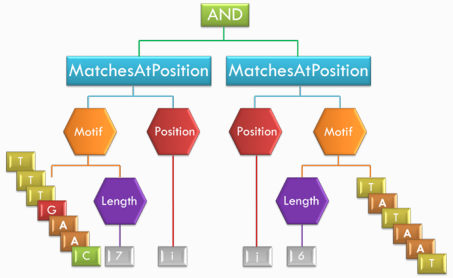
## Representation Q: What Constraints does Function Impose on Sequence?




Key insight: encode implicit constraints in linear representation

- Non-local/distal constraints
- Capture them as features (compositional, positional, correlational, …)
- Explicit, *interpretable* answer to how sequence encodes function

- 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} |
| ERC-int | | Integer | |

Operators / Terminals

Productions: AND, OR, NOT

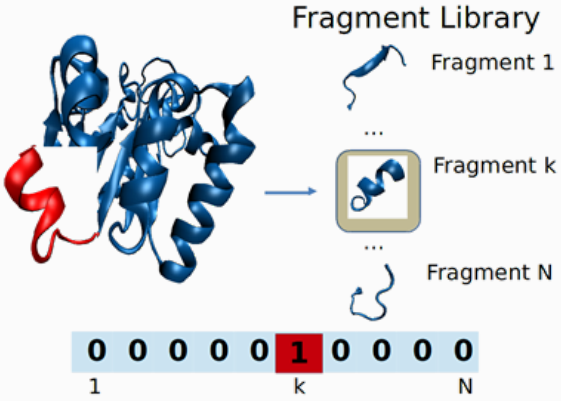
Problem: Exponential explosion of feature space!

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

---

## Prediction Task: Structure → Function

## Representation Q: What Constraints does Function Impose on (Protein) Structure?




Key insight: protein structure == bag of fragments/words

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

IEEE/ACM Trans Comput Biol
PLoS One
BMC Boinformatics
ACM Genet and Evol Comput Conf
IEEE Congress on Evol Comput

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 |

Inspired foundational ML research
GP-based exploration over (structured representations of) kernels

Feature and Kernel Evolution for Improved Classification via SVM

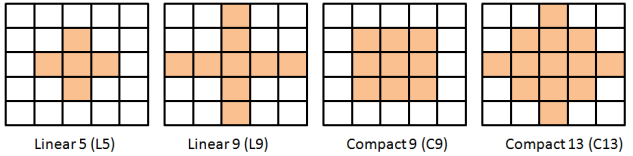
Other contributions: detection of remote homologs; organization of protein structure preserving function co-localization, …

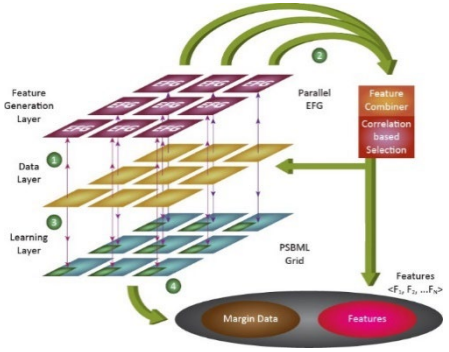
# Cross-fertilization of sub-domains: ML + Evolutionary Computation

# Scalable Learning on Big Data via A Meta-Learning Paradigm

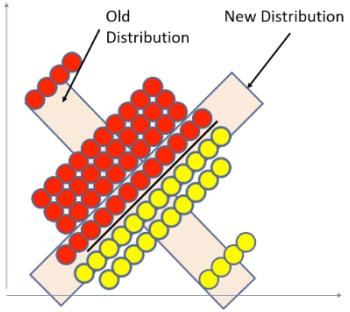
- Parallel Spatial Boosting Machine Learner (PSBML)
  - Distribute training data and evolve toward instances that determine decision boundary

Intl Conf on Parallel Problem Solving from nature (PPSN)




- At each topological grid node, a local algorithm is run that has only local interactions with immediate neighbors
- Each node/classifier trains on own data but tests on own data and training data of neighbors
- Local set of training instances updated via stochastic sampling over confidence-weighted instances



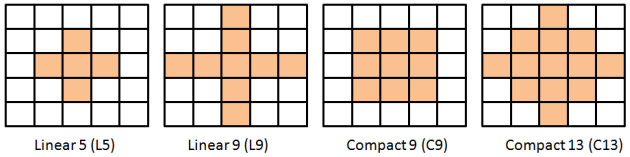

Boosting-like behavior: Hard instances migrate over nodes

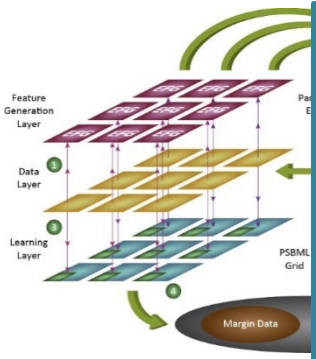
- The more difficult an instance is, the more likely it will be selected for sampling → local training sets evolve $cs_i = \min_{n \in N_i} c_{ni}$ $cs_i^{\text{norm}} = \frac{cs_i - cs_{\min}}{cs_{\max} - cs_{\min}}$ $w_i = 1 - cs_i^{\text{norm}}$
- Theoretical results: PSBML converges to a data distribution whose modes are centered around the margin (hardest points) – behavior modeled via Gaussian mixtures and mean shift

# Scalable Learning on Big Data via A Meta-Learning Paradigm

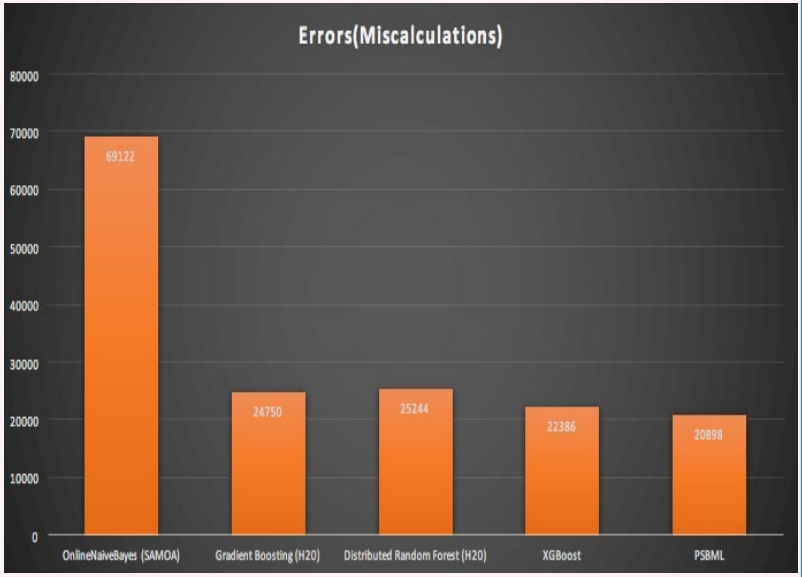
- Parallel Spatial Boosting Machine Learner (PSBML)
  - Distribute training data and evolve toward instances that determine decision boundary

Intl Conf on Parallel Problem Solving from nature (PPSN)

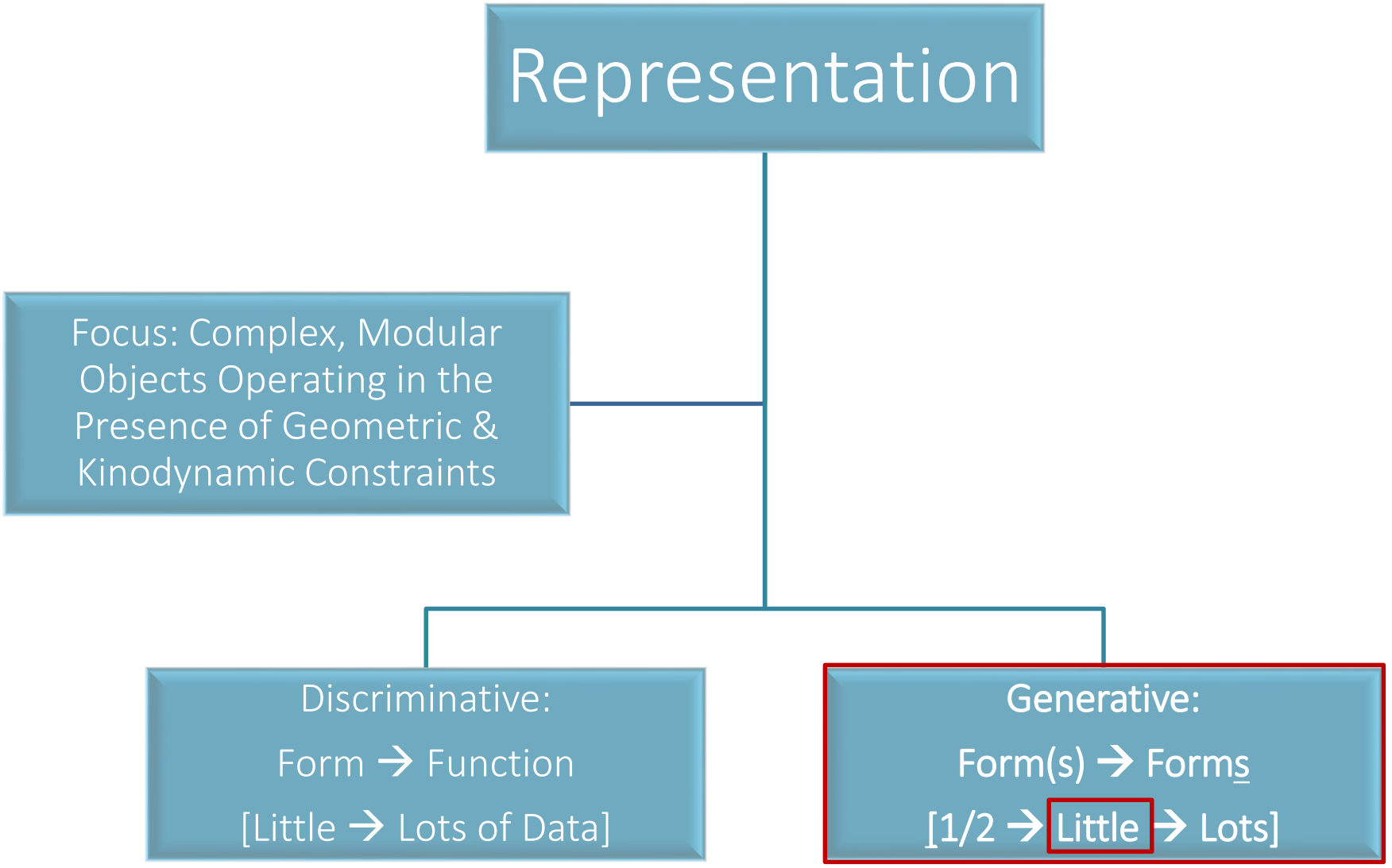






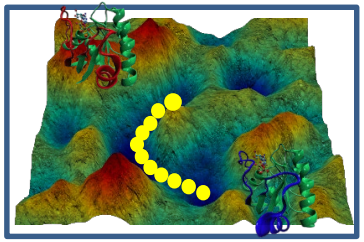

  - The more di
    evolve $cs_i$ =
  - Theoretical
    margin (hard


Representation
Focus: Complex, Modular Objects Operating in the Presence of Geometric & Kinodynamic Constraints
Discriminative:
Form → Function
[Little → Lots of Data]
Generative:
Form(s) → Forms
[1/2 → Little → Lots]

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

*The energy landscape underscores the inherent nature of biomolecules as dynamic systems interconverting between structures with varying energies*

H-Ras switching between its GTP-bound (red) and GDP-bound (blue) structures

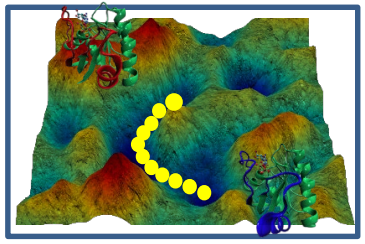


Schematic illustration of H-Ras structures on energy landscape

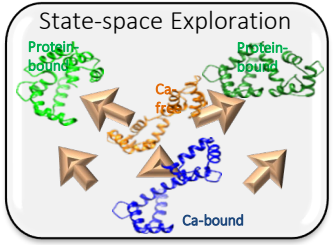
# Knowledge- and Data-driven AI: EA Sampling of Protein Energy Landscape

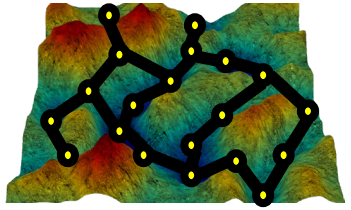
*The energy landscape underscores the inherent nature of biomolecules as dynamic systems interconverting between structures with varying energies*

H-Ras switching between its GTP-bound (red) and GDP-bound (blue) structures



Schematic illustration of H-Ras structures on energy landscape



Thousands of dimensions/dofs – can leverage little data (deposited structures)


Theoretical basis? Conformational selection/population shift:

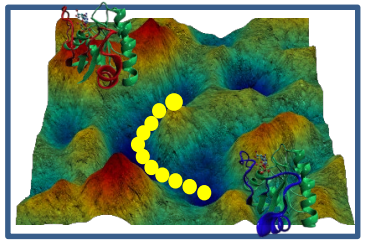
Experimentally-determined structures of bound, unbound, WT and diseased variants are ***known points*** in the state space!

→ **Leverage them to define and initialize variable space**

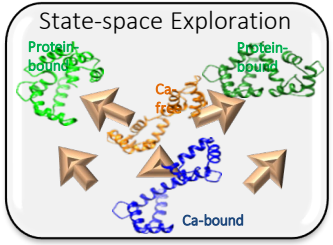
# Knowledge- and Data-driven AI: EA Sampling of Protein Energy Landscape

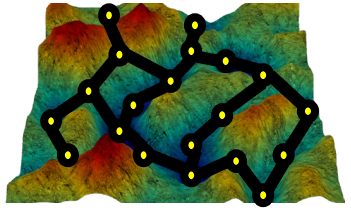
*The energy landscape underscores the inherent nature of biomolecules as dynamic systems interconverting between structures with varying energies*

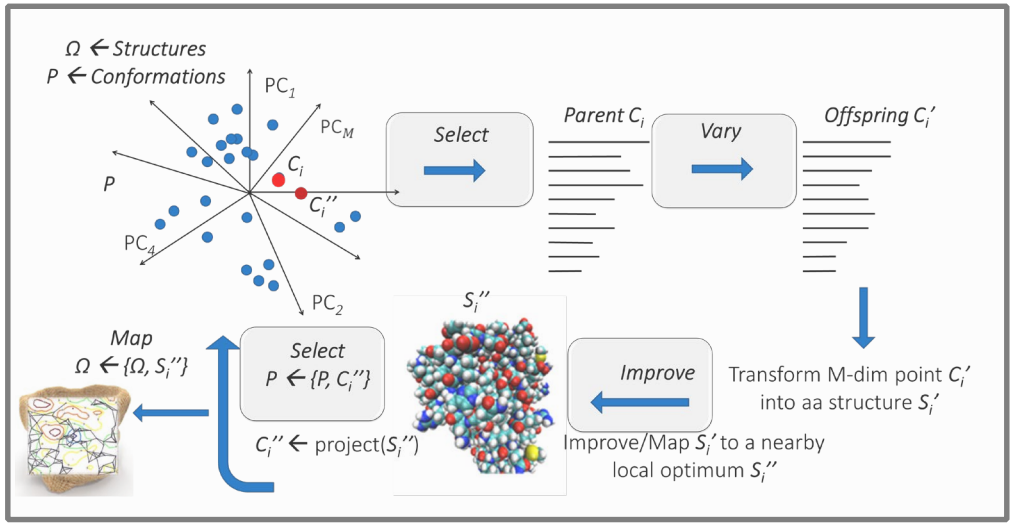
H-Ras switching between its GTP-bound (red) and GDP-bound (blue) structures



Schematic illustration of H-Ras structures on energy landscape






Thousands of dimensions/dofs – can leverage little data (deposited structures)

Theoretical basis? Conformational selection/population shift:

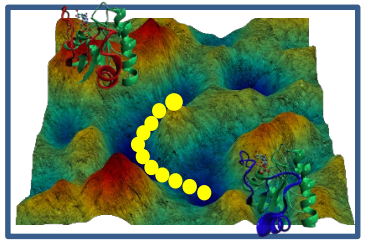
Experimentally-determined structures of bound, unbound, WT and diseased variants are ***known points*** in the state space!

→ **Leverage them to define and initialize variable space**

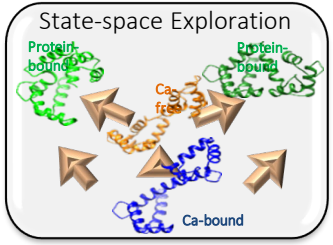
# Knowledge- and Data-driven AI: EA Sampling of Protein Energy Landscape

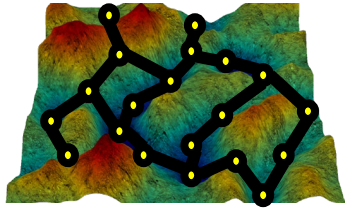
*The energy landscape underscores the inherent nature of biomolecules as dynamic systems interconverting between structures with varying energies*

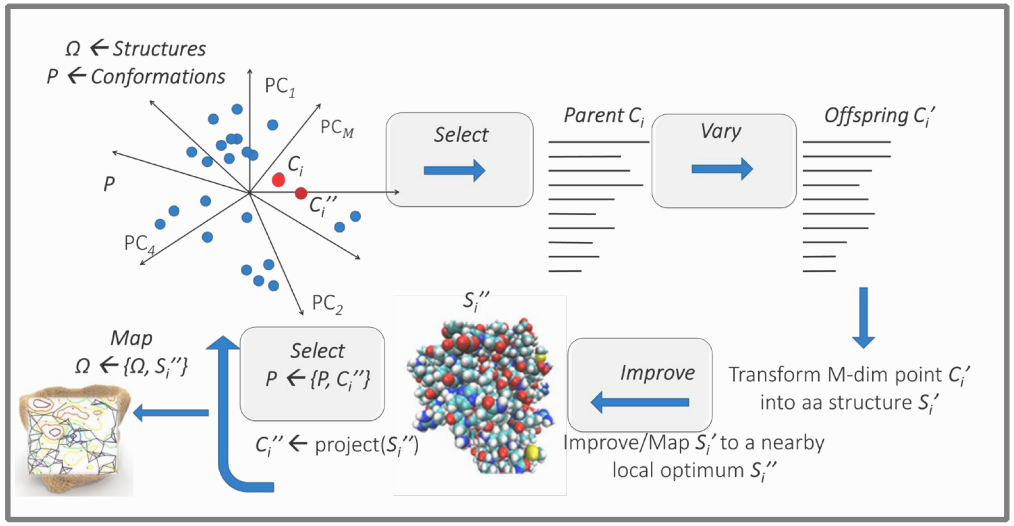
H-Ras switching between its GTP-bound (red) and GDP-bound (blue) structures



Schematic illustration of H-Ras structures on energy landscape





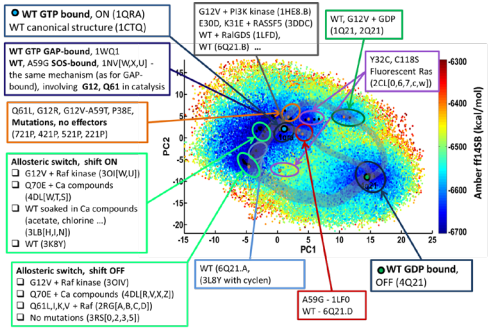

Thousands of dimensions/dofs – can leverage little data (deposited structures)

Theoretical basis? Conformational selection/population shift:

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

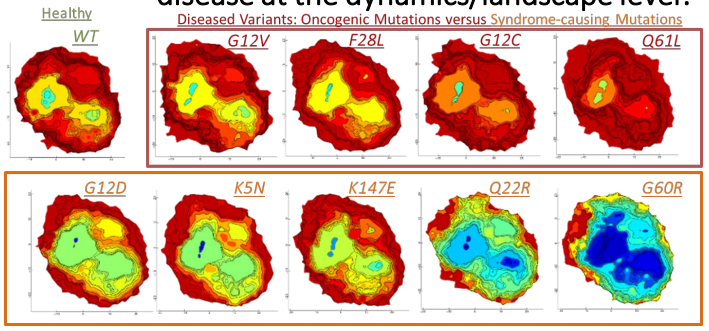
→ Leverage them to <u>define</u> and <u>initialize</u> variable space


## First-ever view of H-Ras WT energy landscape




## We can now see distinct mechanisms of disease at the dynamics/landscape level!

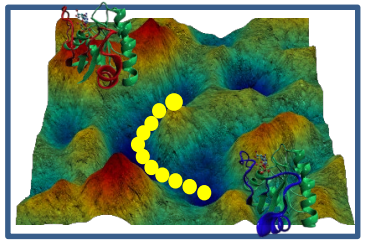
<u>Diseased Variants: Oncogenic Mutations versus Syndrome-causing Mutations</u>



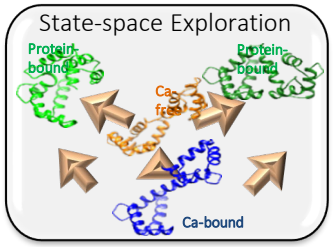
# Knowledge- and Data-driven AI: EA Sampling of Protein Energy Landscape

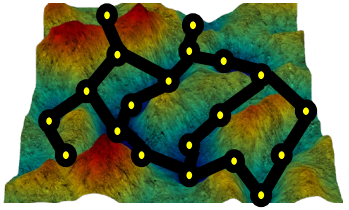
*The energy landscape underscores the inherent nature of biomolecules as dynamic systems interconverting between structures with varying energies*

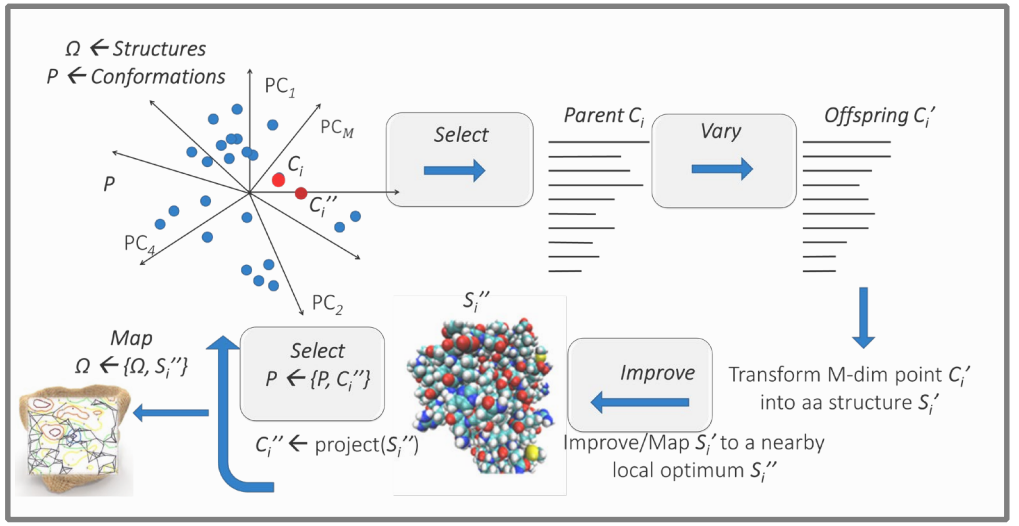
H-Ras switching between its GTP-bound (red) and GDP-bound (blue) structures

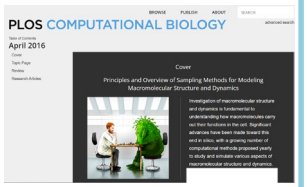


Schematic illustration of H-Ras structures on energy landscape





(top 50 most downloaded in 2016 and *featured on April issue front cover. Also featured in the PLoS Comp Biol blog.*)

PLoS Comput Biol, Robotica J Comput Biol, IEEE Trans NanoBioScience, BMC Genomics, IEEE/ACM TCBB, BMC Bioinformatics, Molecules, Biomolecules, ACM Bioinf and Comput Biol, ACM Genet and Evol Comput Conf, IEEE Intl Conf on Bioinformatics and Biomedicine

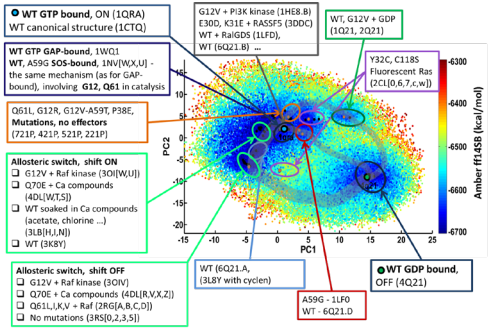
Thousands of dimensions/dofs – can leverage little data (deposited structures)

Theoretical basis? Conformational selection/population shift:

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

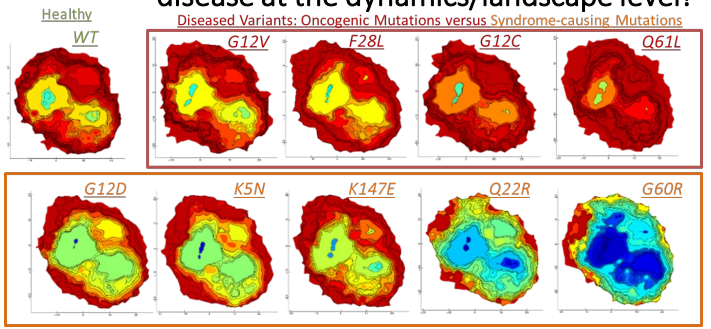
→ Leverage them to define and initialize variable space


## First-ever view of H-Ras WT energy landscape



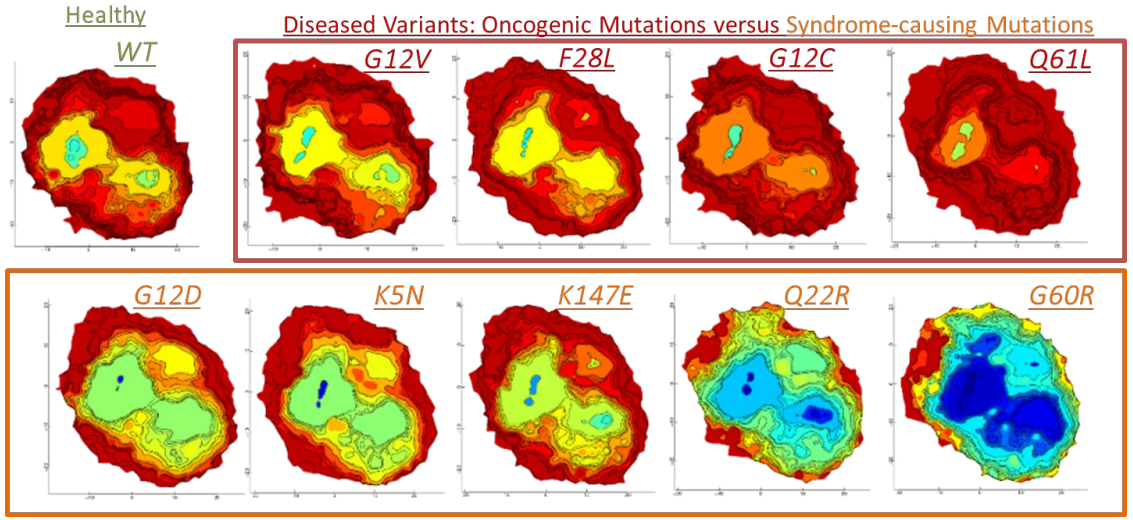

## We can now see distinct mechanisms of disease at the dynamics/landscape level!

Diseased Variants: Oncogenic Mutations versus Syndrome-causing Mutations

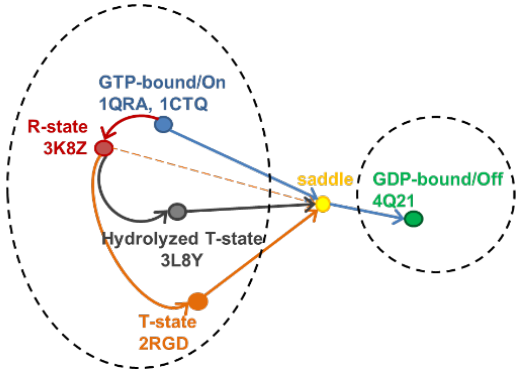


# Predicting Phenotypical Impact of Mutations

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




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.




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

# Representation

Focus: Complex, Modular Objects Operating in the Presence of Geometric & Kinodynamic Constraints

Foundational Optimization Research

- Brought structurization of search spaces (from motion planning-inspired algorithms) to evolutionary algorithms
- Recast motion planning algorithms under umbrella of evolutionary computation (unified treatment → better understanding → new ideas unlocked)
- Foundational advances back to AI (distributed ML)

Discriminative:

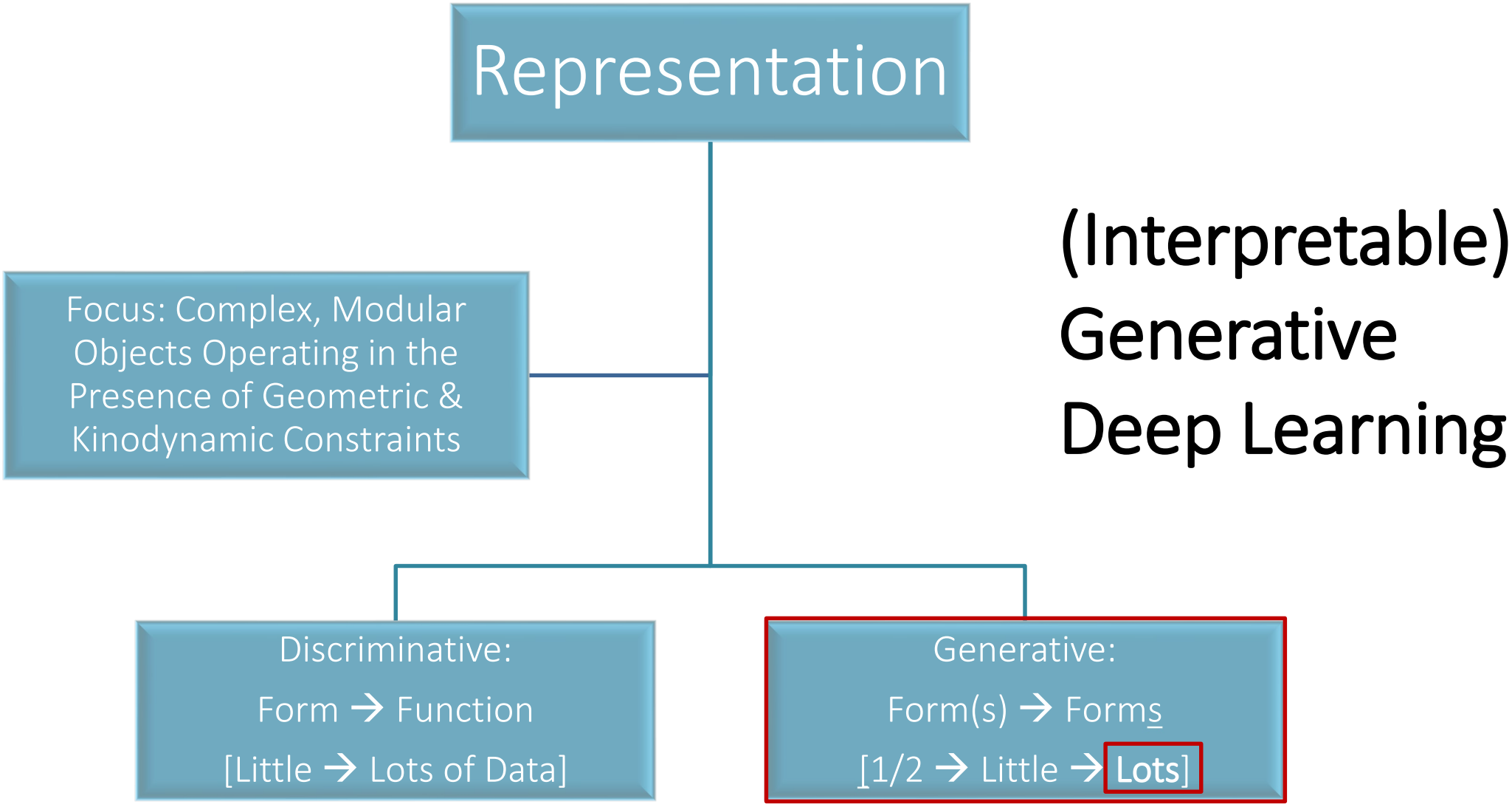
Form → Function

[Little → Lots of Data]

Generative:

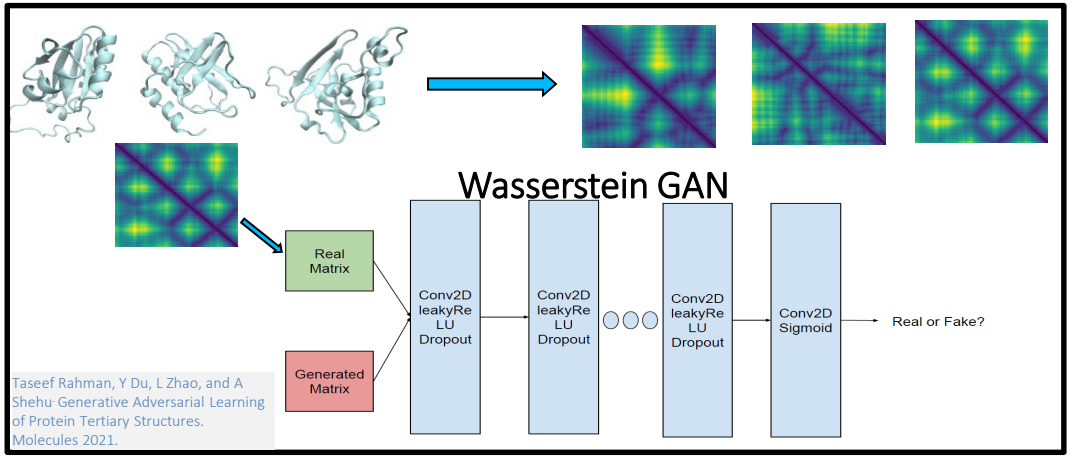
Form(s) → Forms

[1/2 → **Little** → Lots]

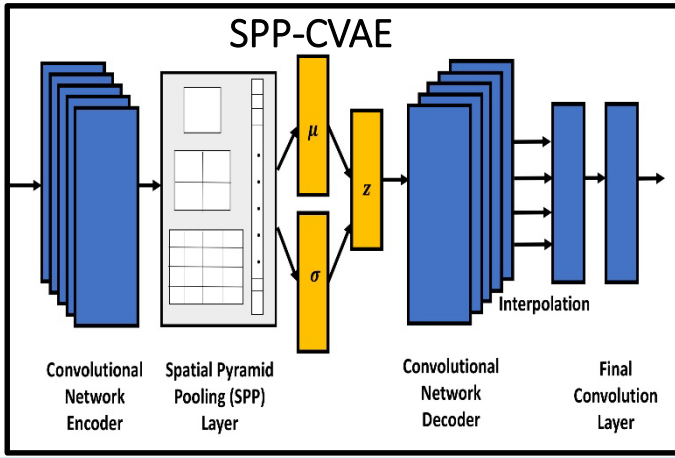



# (Interpretable) Generative Deep Learning

# Paradigm Shift: Protein Conformation Sampling via Deep Latent Variable Models




Taseef Rahman, Y Du, L Zhao, and A Shehu. Generative Adversarial Learning of Protein Tertiary Structures. Molecules 2021.



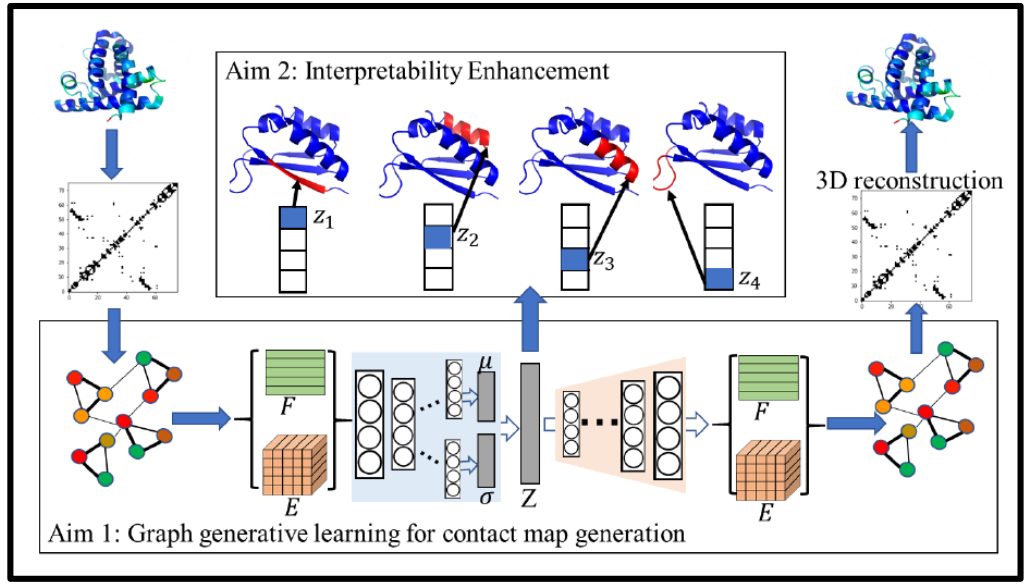

- Treats contact maps as 2D images
- Leverages 2D convolution
- SPP allows training over structures of proteins of variable lengths

Fardina Alam and A Shehu. Data Size and Quality Matter: Generating Physically-realistic Distance Maps of Protein Tertiary Structures. Biomolecules 2022 (**featured as title story for 60 days & Editor's Choice**).

**Goal:** Learn Underlying Distribution of Experimentally-determined Protein Tertiary Structures

**Novel AI&ML Methodologies:** Generative Models – Generative Adversarial Learning, Convolutional Variational Autoencoders, Graph Variational Autoencoders with Disentanglement, and more

## Our State of the Art NN Framework

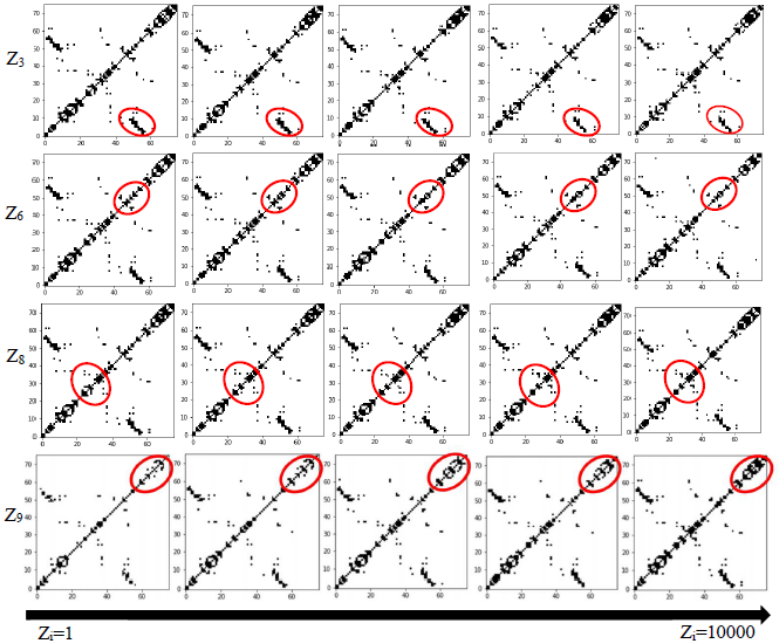


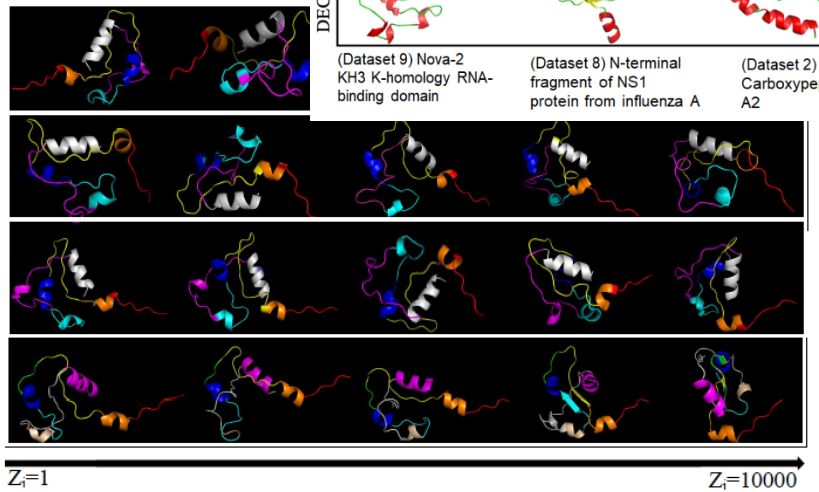

Xiaojie Guo, Y Du, S Tadepalli, L Zhao, and A Shehu. Generating Tertiary Protein Structures via Interpretable Graph Variational Autoencoders. Bioinformatics Advances 2021

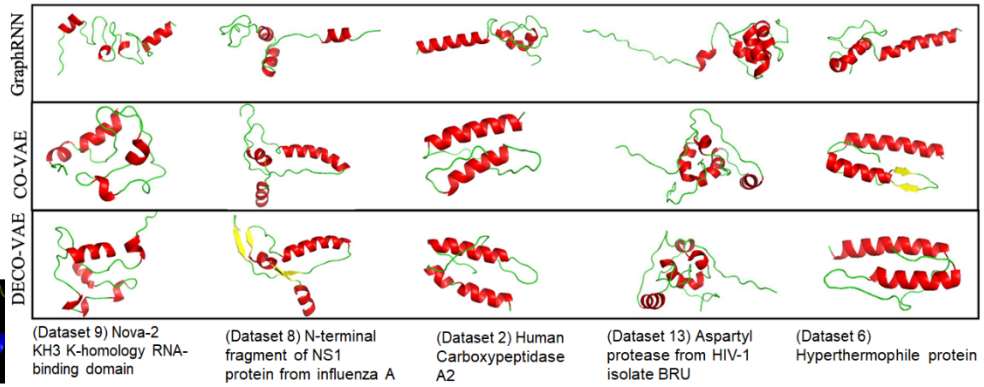
Taseef Rahman, Y Du, and A Shehu. Graph Representation Learning for Protein Conformation Sampling. IEEE Intl Conf on Comput Adv in Bio and Medical Sciences 2021.

# Graph Learning and Generation (with Disentanglement)

- *Better, physically-realistic structures*
- *Learned latent factors allow structural control*



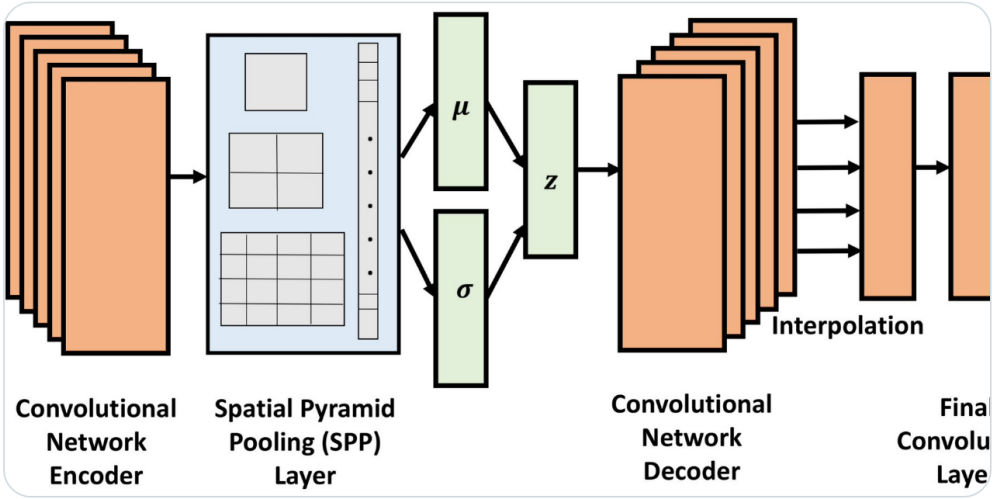

*One of few labs advancing generative deep learning for modeling structural plasticity*

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.

# Brick by Brick

#EditorsChoice
Read now ➡️ "Data Size and Quality Matter: Generating Physically-Realistic Distance Maps of Protein Tertiary Structures" by Fardina Fathmiul Alam and Amarda Shehu, et al.

The article 👉 bit.ly/3FWm3Tj
The list 👉 bit.ly/3VzbpHC

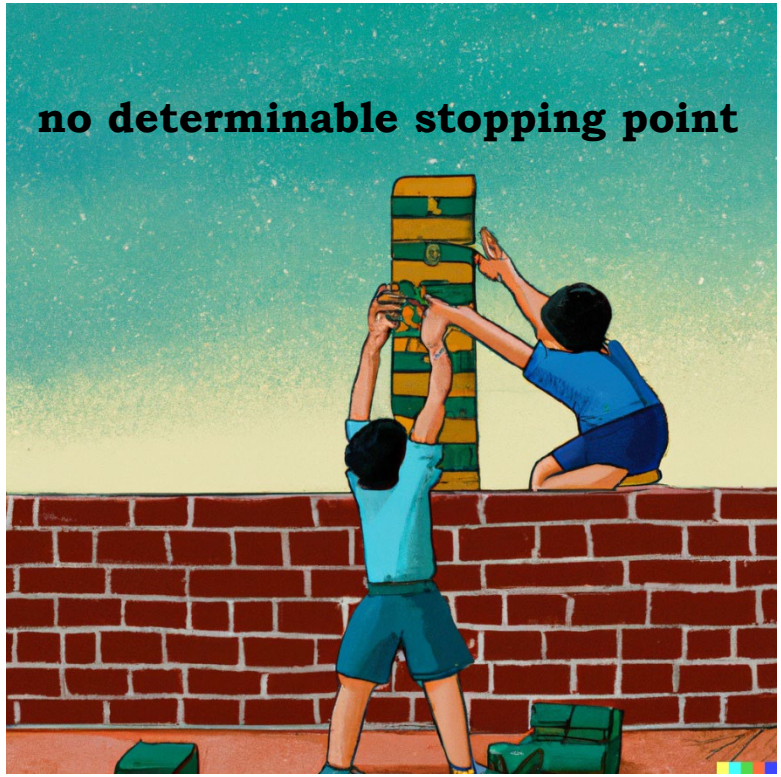



3:49 AM · Dec 16, 2022

Generated with DALL-E

# Latent Factors == Concept Space?

- **Are there other scientific domains where:**
  - Latent factors provide us with an advantage
- Latent factors as concepts to explain generation process and relate to our theoretical and empirical understanding of underlying physics
- Latent factors as concepts to control generation process and link the chemical with the biological space

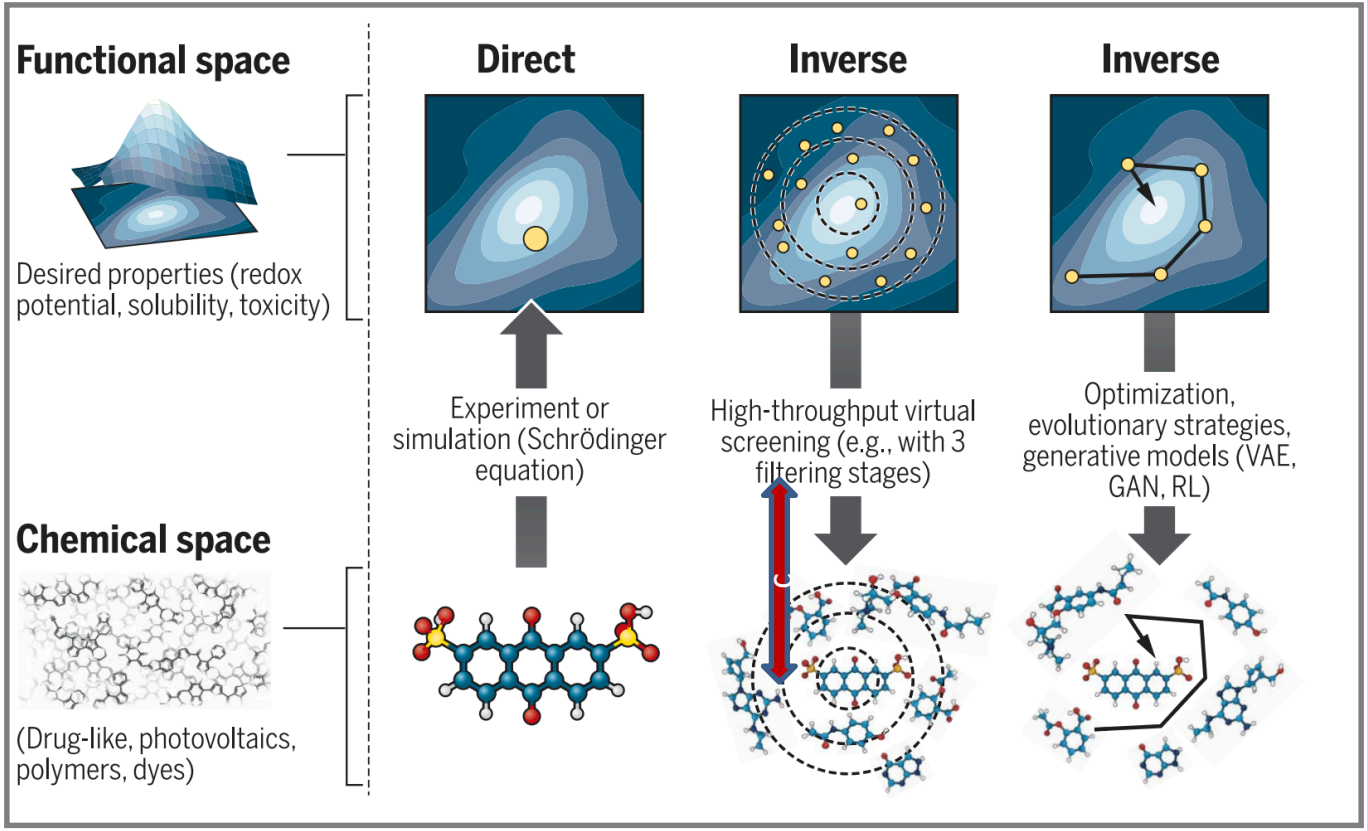



Generated with DALL-E

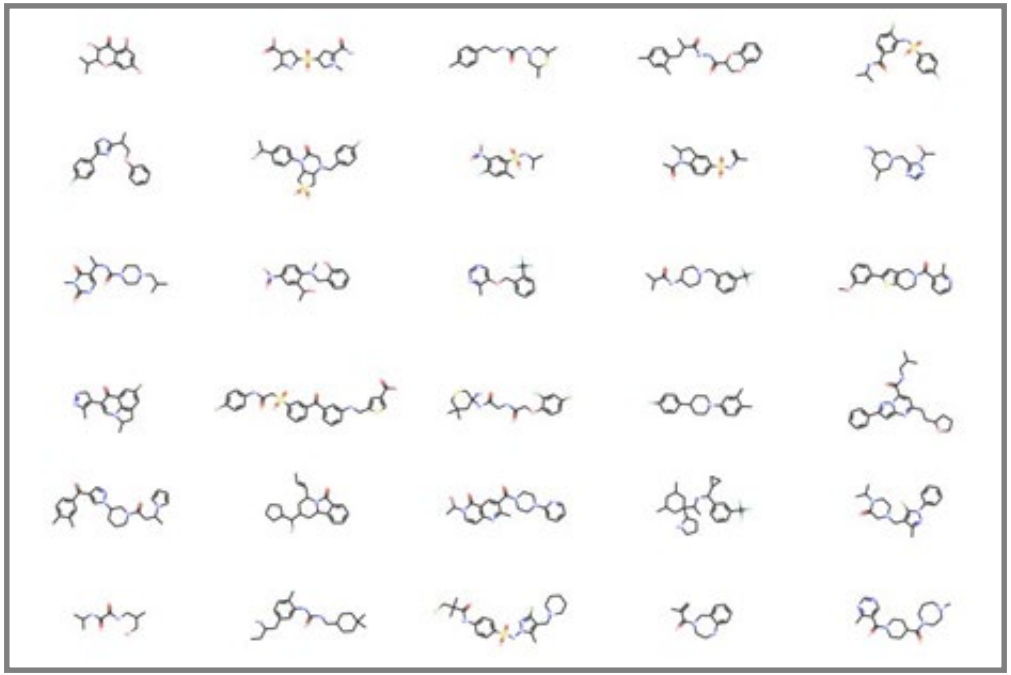
# Small Molecule Generation: Linking Chemical and Biological Space

Material Science

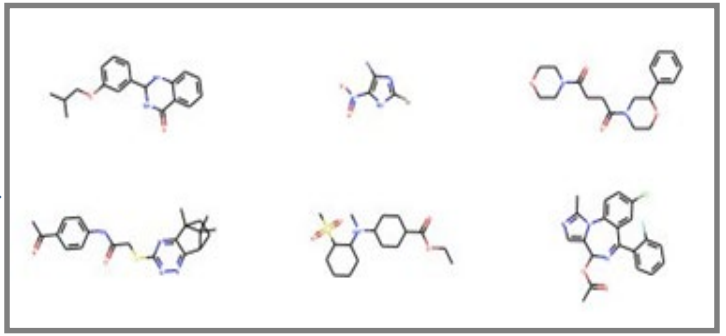
Drug Discovery



# Small Molecule Generation: Typical Setup

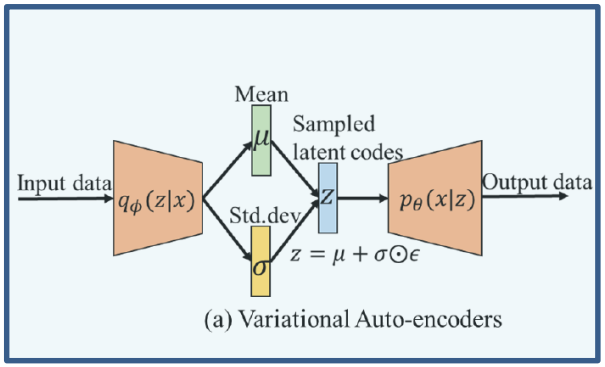


Input: Small molecules deposited in publicly-available databases

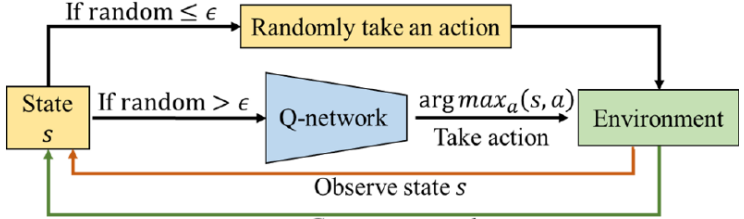


Output: Small molecules generated by a deep model

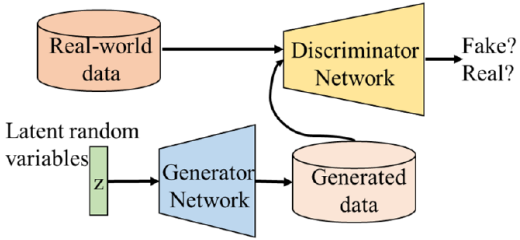
# Small Molecule Generation: Typical Approaches



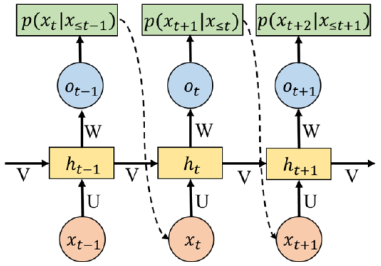
(a) Variational Auto-encoders



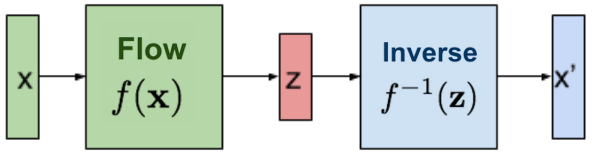
(b) Deep-Q Network



(c) Generative Adversarial Nets



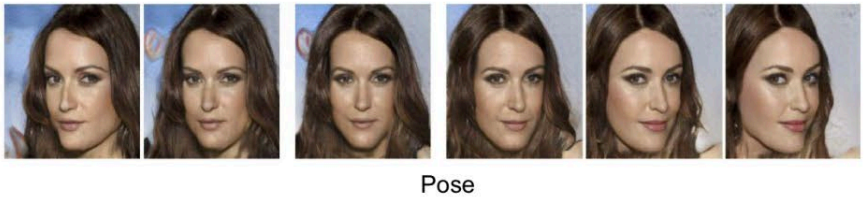
(d) Generative RNN



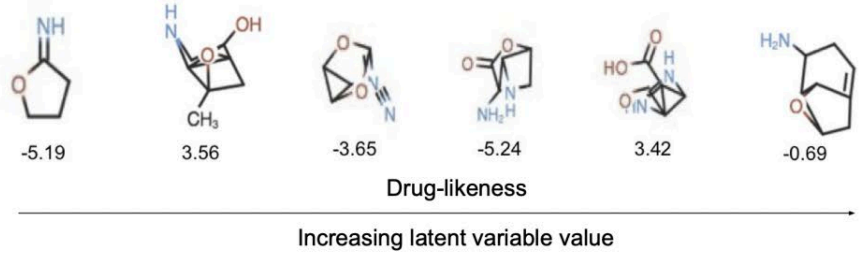
(e) Flow-based generative models

# Our Objective:
# Generate with (Property) Control

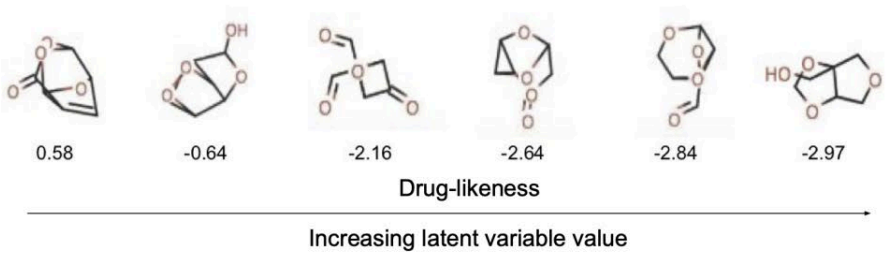

(a) Image generation without control




(b) Image generation with control



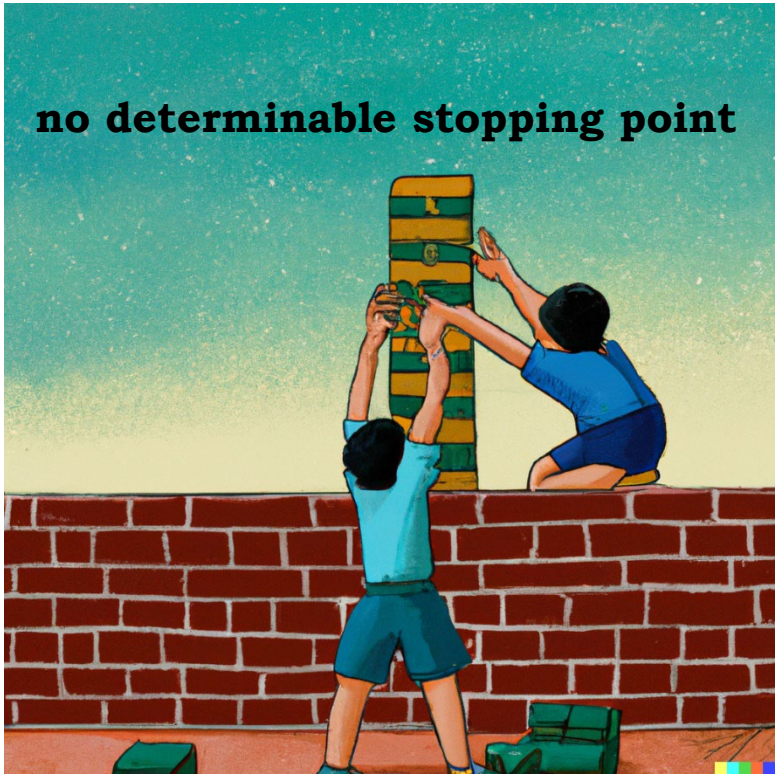

(c) Molecule generation without control




(d) Molecule generation with control

# Strengthening Encoder-Decoder Frameworks: Instructibility

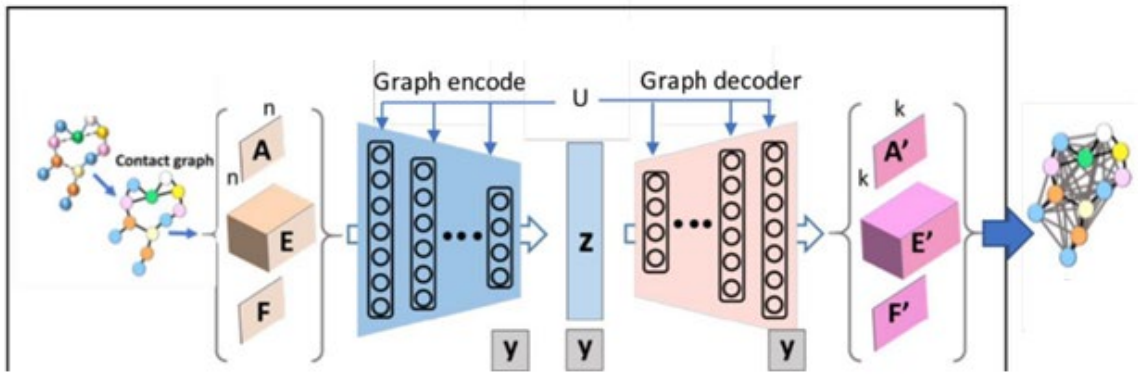
- Projecting model behavior onto a concept space provides opportunities for adapting behavior based on explicit feedback from chemists in the wet laboratory
- VAE-based frameworks provide an interesting platform amenable to achieve model instructibility
- Leveraging latent factors as axis of a concept space, disentangled graph variational autoencoder (D-GraphVAE) frameworks allow control of small molecule generation in the biological/phenotypical/concept space




Generated with DALL-E

# Our Approach: Graph Variational Autoencoder-based Learning

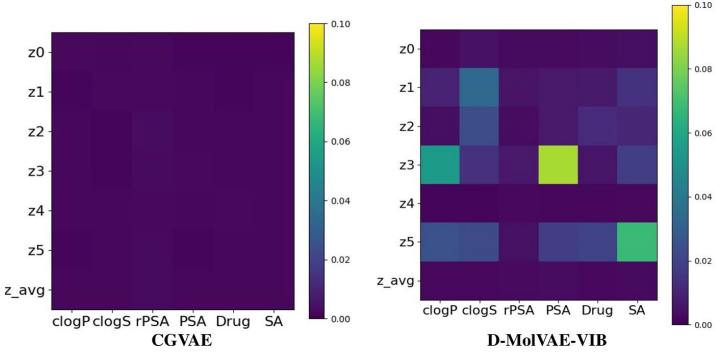
- Input: Molecular Graph $G = (V, E, A, F)$
- Output: Novel Molecular Graph $G'$
- Nodes – Atoms $V$, Edges – Bonds $E \subseteq V \times V$
- Node Attribute – Node Features $F \in \mathbb{R}^{N \times K'}$, where k' = number of atom types
- Edge Attribute – Adjacency Matrix $A \in \mathbb{R}^{N \times N \times K}$, where k = number of bond types




- The deep latent-variable framework parameterizes VAEs to learn a joint distribution over a molecular graph G and desired properties **Y**, given a group of learned disentangled latent variables Z.
- The generative process is formulated as p(G|Y,Z). The objective is to learn to (i) encode a molecular graph into a continuous latent space with p(Z,Y|G) and (ii) decode a molecule from the learned latent space with p(G|Z,Y).

# From Small Molecule Generation to Generation with Property Control

- Disentanglement does not degrade performance: in fact, evidence that some disentangled factors (additionally) control for molecular properties

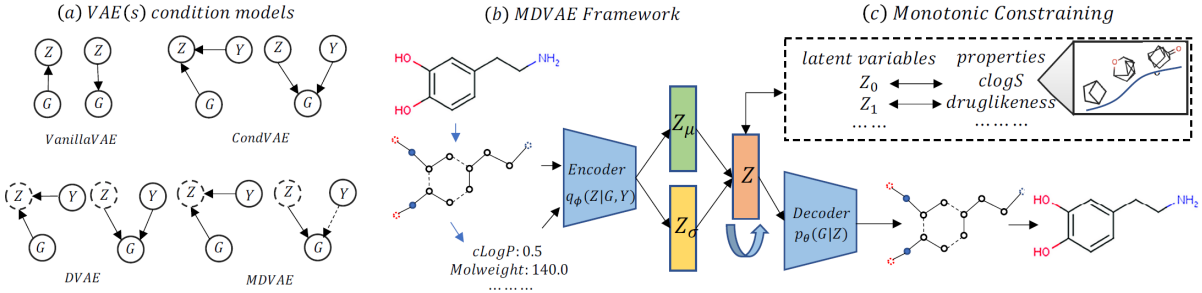



Yuanqi Du, X Guo, Y Wang, A Shehu, and L Zhao. Small Molecule Generation via Disentangled Representation Learning. *Bioinformatics*, btac296, 2022.

- Let us then reserve some latent factors for given molecular properties and show control

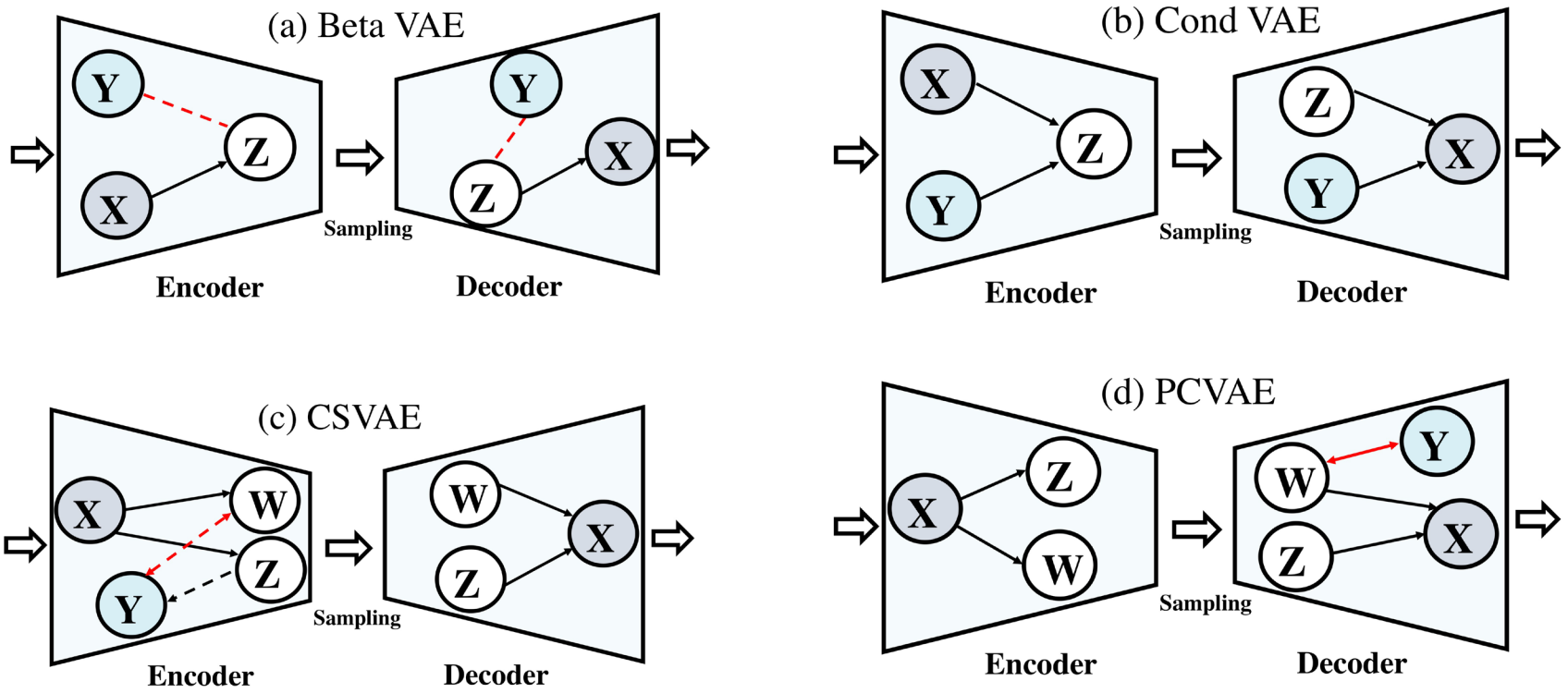
Yuanqi Du, Y Wang, F Alam, Y Lu, X Guo, L Zhao, and A Shehu. Deep Latent-Variable Models for Controllable Molecule Generation. *IEEE Intl Conf on Bioinformatics and Biomedicine*, 2021

- Let us enforce a monotonic relationship between latent factors and given properties and obtain *better* control




Yuanqi Du, X Guo, A Shehu, and L Zhao. Interpretable Molecular Graph Generation via Monotonic Constraints. *SIAM Data Mining*, 2022.

# Mechanisms to Connect Latent Factors with Properties

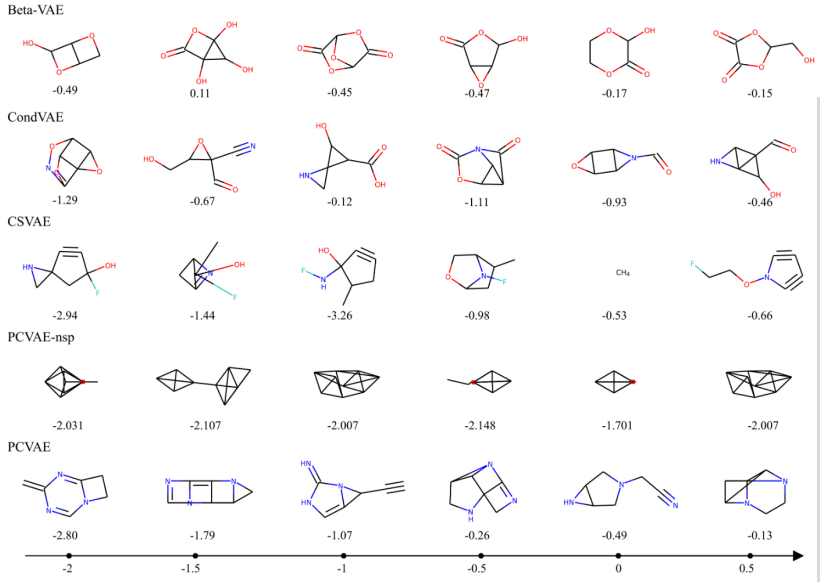



Enforcement of independence is shown by dotted red arrows; invertible dependence between two variables is represented by double arrows. Data is denoted by X and Z. W are subsets of latent variables. Y denotes targeted molecular properties.

# From Small Molecule Generation to Generation with Property Control

- *More effective models at generating valid, unique, and novel molecules*

| Model | QM9 Validity | QM9 Novelty | QM9 Uniqueness | ZINC Validity | ZINC Novelty | ZINC Uniqueness | MOSES Validity | MOSES Novelty | MOSES Uniqueness |
|---|---|---|---|---|---|---|---|---|---|
| $\beta$-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-nsp | **100.00%** | **98.57%** | 86.94% | **100.00%** | **100.00%** | 99.74% | **100.00%** | 99.90% | 99.80% |
| PCVAE | **100.00%** | 97.43% | 88.24% | **100.00%** | **100.00%** | 99.48% | **100.00%** | 99.96% | 98.62% |




Caption: We set out to control the cLogS property of generated molecules to be [-2, -1.5, -1, -0.5, 0, 0.5], respectively. The models will show different levels of controllability. PCVAE achieves superior performance. The model learns a monotonic relationship; an increasing value of z relates very closely here with an increase in the value of cLogS.

- *More effective models at providing us with control over properties*

| Method | cLogP $[-2, 2]$ | cLogS $[-2, 2]$ | Drug $[-5, 5]$ | Weight $[120, 130]$ | PSA $[20, 60]$ | SA $[2, 5]$ |
|---|---|---|---|---|---|---|
| $\beta$-VAE | 2.45 | 1.01 | 43.83 | 264.59 | 249.72 | 7.03 |
| CondVAE | 2.20 | 0.99 | 22.27 | 42.03 | **183.43** | 4.87 |
| CSVAE | **0.67** | 0.96 | 9.24 | 39.73 | 810.45 | 1.86 |
| PCVAE-nsp | 2.15 | 3.18 | 8.99 | **38.45** | 765.44 | **1.84** |
| PCVAE | 1.13 | **0.62** | **5.41** | 38.59 | 1554.00 | 1.87 |

Caption: For each property, we generate 100 molecules with properties specified in the range with the highest density in the molecular property distribution (as in training dataset). We report the discrepancy via MSE between properties of molecules generated in this manner and expected properties predefined within the highest property density region.


Y Du, X Guo, Y Wang, A Shehu, and L Zhao. Small Molecule Generation via Disentangled Representation Learning. *Bioinformatics*, btac296, 2022.

Y Du, X Guo, A Shehu, and L Zhao. Interpretable Molecular Graph Generation via Monotonic Constraints. *SIAM Data Mining*, 2022.

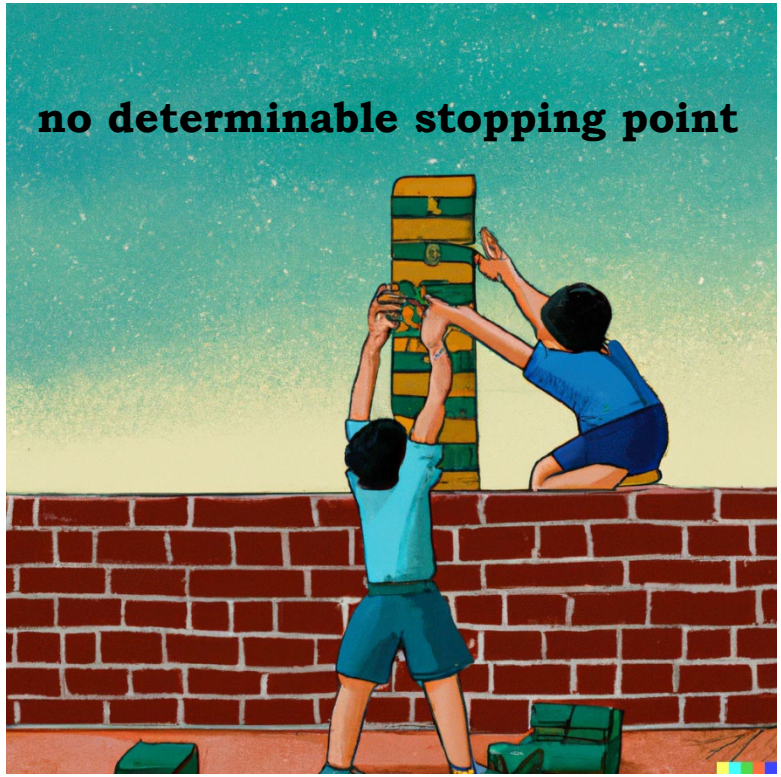
Y Du, Y Wang, F Alam, Y Lu, X Guo, L Zhao, and A Shehu. Deep Latent-Variable Models for Controllable Molecule Generation. *IEEE Intl Conf on Bioinformatics and Biomedicine*, 2021

# Brick by Brick

- What about correlated properties?

  Yuanqi Du, [..], A Shehu, and L Zhao. Multi-objective Deep Data Generation with Correlated Property Control. *NeurIPS*, 2022.

- What about incomplete and noisy data? → our KDD 2022 workshop paper
- What about very fine-granular control with a specific target in mind?
- What about specific binding affinities?
- Can you actually design these things in the lab and learn from the wet laboratory? → our 2023 NSF IIBR Grant




Generated with DALL-E

# Representation

How do you **incorporate** millions of sequences, hundreds of thousands of structures, thousands of known interactions, millions of short reads, thousands of wet-lab characterization data, etc.

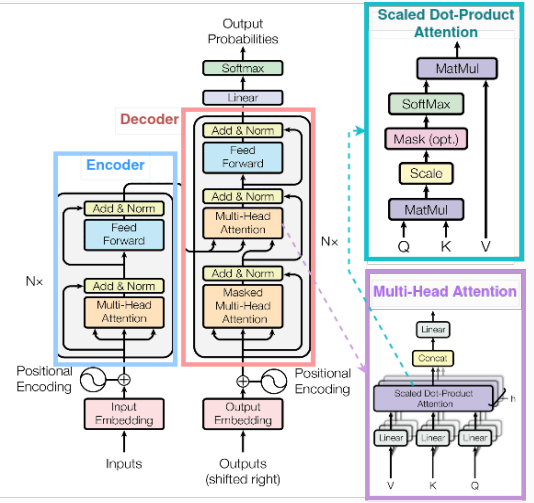
**Multi-modal Foundation Models**

How do you leverage knowledge for lighter models?

**Grounded Foundation Models**

How do you catch up to the transformer with alternative architectures?

**Small State Space Models**




Discriminative:

Form → Function

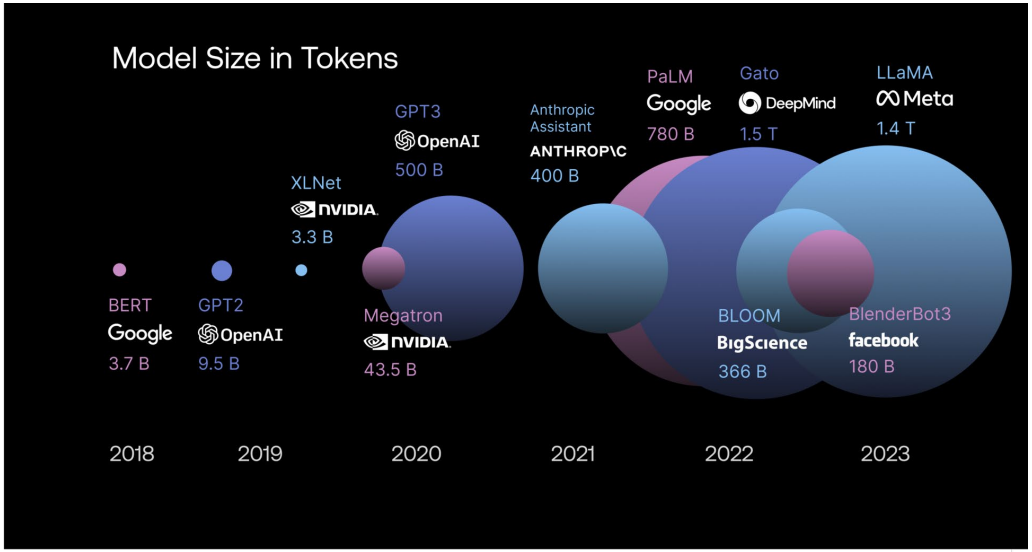
[Little → Lots of Data]

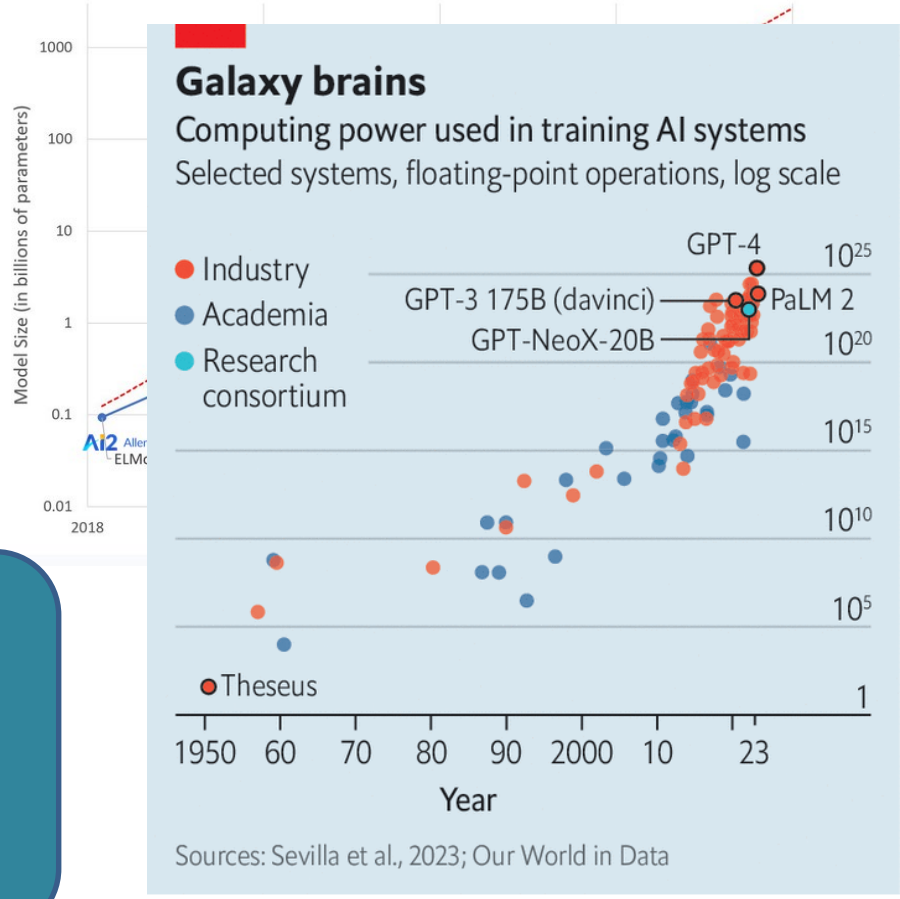
Generative:

Form(s) → Forms

[1/2 → Little → Lots]

# Progress NOT by Scale Alone!




**Objectives:**

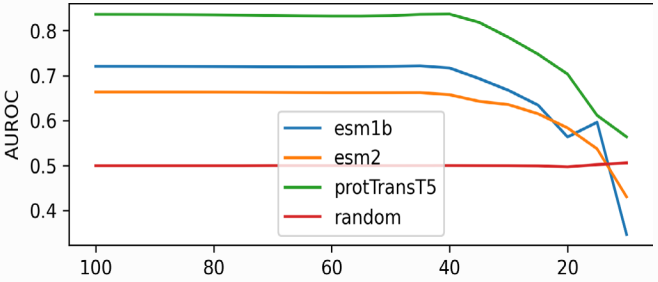
Lighter models by situating them in and exploiting knowledge and structure in biological data

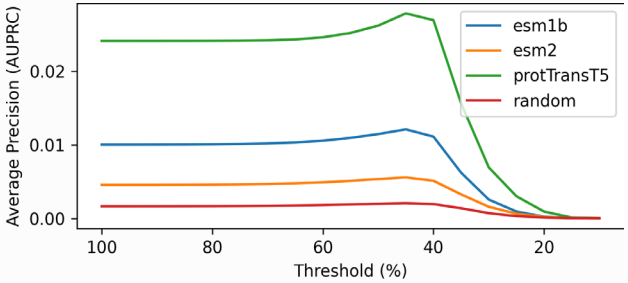
Beyond the transformer architecture (foundation model research) [2023 NSF III Medium Grant]

# Progress NOT by Scale!

## Remote Homology

**Finding**: Employed a rigorous setting, by lowering sequence identity, we report that current state-of-the-art, large protein language models are still underperforming in the twilight zone of very low sequence identity



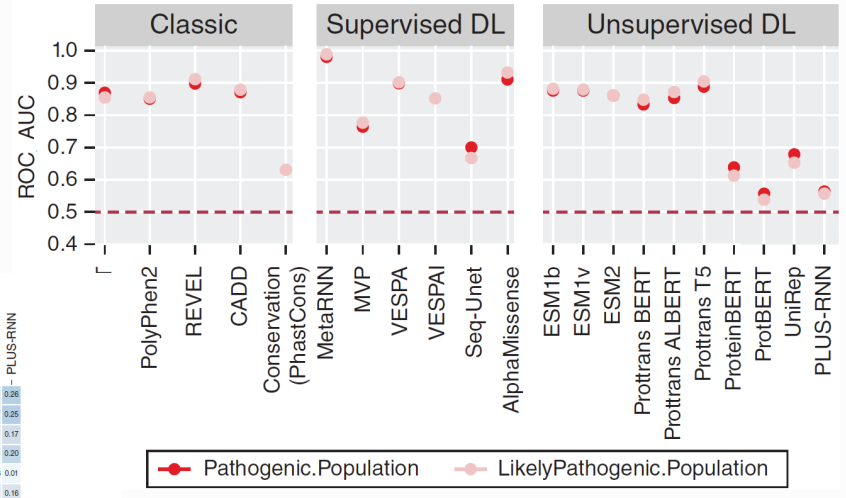



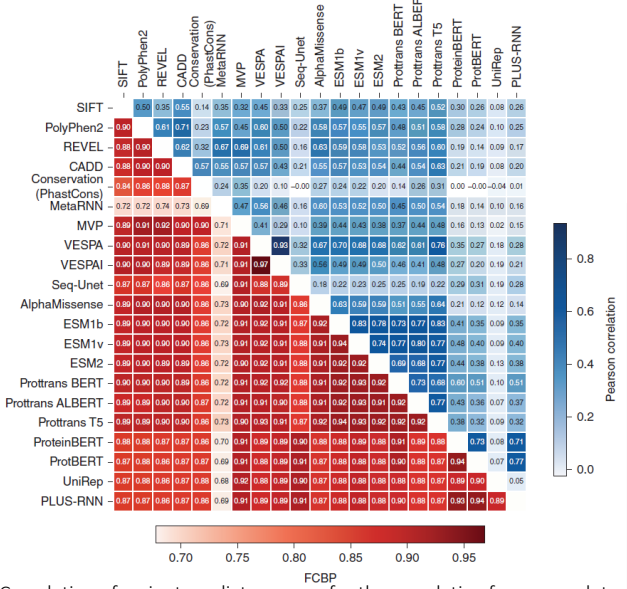

Kabir, Moldwin, and Shehu. A Comparative Analysis of Transformer-based Protein Language Models for Remote Homology Prediction. ACM-BCB 2023 and AAAI-W 2024. Best Paper Award.

## Variant Prediction

There are three broad variant effects: (evolutionary) *fitness effect, pathogenicity (disease causation), and (molecular) function change*.

Existing computational methods may have largely failed to recognize the difference between these three types of effect.




Correlation of variant predictor scores for the population frequency data set is reported as the Pearson correlation coefficient (blue), and the fraction of consensus binary predictions (FCBPs) (red). Higher scores (darker colors) indicate a better correlation.

**Major Findings:** Of the unsupervised DL methods, ESM (650M) and ProtTrans (11B) were able to differentiate clinically significant variants from the general population better than simply using conservation. These models were also as good as or better than many of the supervised methods. **However, as neither of the unsupervised methods captured variant population frequency well, the rarity of pathogenic variants is an unlikely cause of these models' pathogenicity classification abilities.**

Bromberg, Kabir, Ramakrishnan, and Shehu. Variant prediction in the Age of Machine Learning. Machine Learning and Protein Science, Cold Spring Harbor Laboratory Press, 2024.

# Strengthening Biological Foundation Models via Grounding

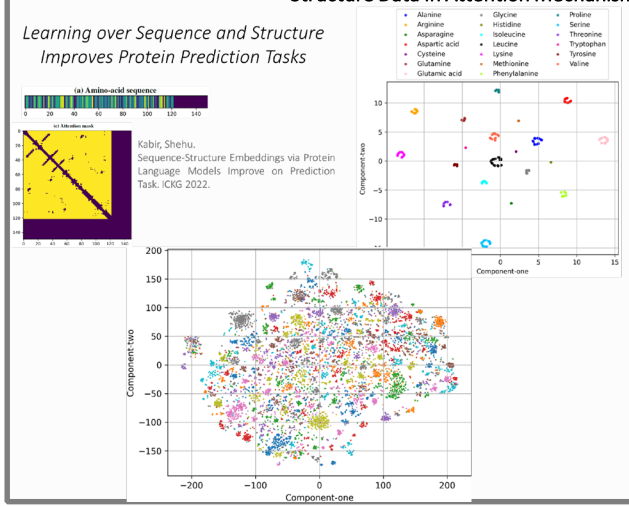
- The data at our disposal are in large part imperfect proxies of the underlying physics that governs the detailed behavior of molecular machineries in the living cell

- *Biological priors*:
- additional data modalities
- physics-based knowledge

# Prior Knowledge in Data or First Principles (Physics)

## Structure Data in Attention Mechanism for Lighter Models

*Learning over Sequence and Structure Improves Protein Prediction Tasks*




Kabir, Shehu. Sequence-Structure Embeddings via Protein Language Models Improve on Prediction Task. ICKG 2022.

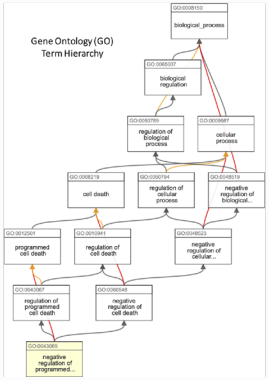
Table 1: Top: Architectural details of ProToFormer [170] and FT-PRoBERTa [5]. Bottom: Performance comparison along accuracy (ACC), precision (PRE), recall (REC), F1-Score (F1), and AUC-ROC score (AR). Consecutive rows per model show performance on validation and test set, respectively. The highest values on a metric on each set are highlighted in boldface font.

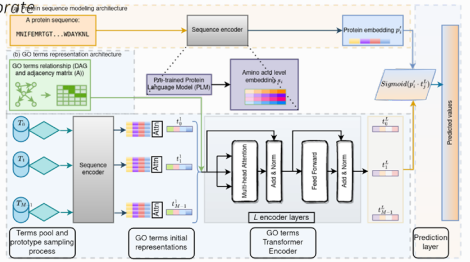
| HYPERPARAMS | PROTOFORMER | FT-PROBERTA |
|---|---|---|
| MAX-LENGTH | 512 | 512 |
| EMBED-DIM | 128, 256 | 768 |
| #-ATTN-HEADS | 8 | 12 |
| #-ENC-LAYERS | 5 | 5 |
| #-WEIGHTS | 13M | 44M |

| MODEL | ACC | PRE | REC | F1 | AR |
|---|---|---|---|---|---|
| FT-PROBERTA | 0.542 | 0.600 | 0.542 | 0.523 | 0.989 |
| | 0.472 | 0.614 | 0.472 | 0.432 | 0.966 |
| PROTOFORMER (128-SEQ) | 0.526 | 0.600 | 0.526 | 0.526 | 0.987 |
| | 0.486 | 0.578 | 0.486 | 0.461 | 0.984 |
| PROTOFORMER (128-SEQ+CM) | 0.720 | 0.762 | 0.720 | 0.713 | **0.992** |
| | 0.664 | 0.741 | 0.664 | **0.664** | **0.985** |
| PROTOFORMER (256-SEQ) | 0.565 | 0.636 | 0.565 | 0.562 | 0.985 |
| | 0.516 | 0.609 | 0.516 | 0.490 | 0.975 |
| PROTOFORMER (256-SEQ+CM) | **0.742** | **0.783** | **0.742** | **0.734** | **0.992** |
| | **0.678** | **0.752** | **0.678** | 0.643 | 0.979 |

## (Hierarchical) Function Relationships in Joint Representation Learning

*A powerful, versatile framework that can additionally incorporate hierarchical functional relationships*

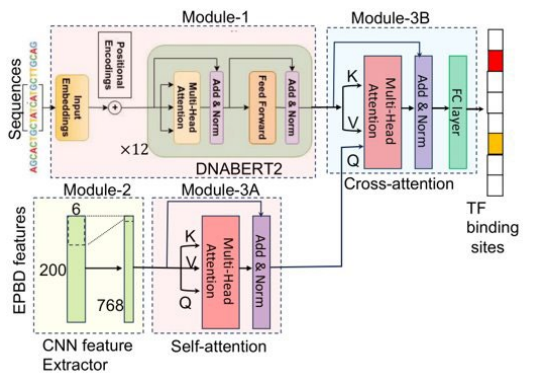



| Metrics | GO | DeepGoPlus Val | DeepGoPlus Test | DeepChoi-ESM-1 Val | DeepChoi-ESM-1 Test | GOProFormer Val | GOProFormer Test |
|---|---|---|---|---|---|---|---|
| Fmax | BP | 0.460 | 0.491 | 0.499 | 0.529 | **0.526** | **0.557** |
| | CC | **0.739** | 0.709 | 0.738 | 0.712 | **0.739** | **0.729** |
| | MF | 0.457 | 0.436 | 0.524 | 0.541 | **0.580** | **0.623** |
| Smin | BP | 10.499 | 9.721 | 10.074 | 9.080 | **9.600** | **8.810** |
| | CC | 4.598 | 5.487 | 4.608 | 5.359 | **4.541** | **5.073** |
| | MF | 7.719 | 8.267 | 7.565 | 7.748 | **6.338** | **6.308** |
| AUPRC | BP | 0.439 | 0.469 | 0.466 | 0.495 | **0.526** | **0.557** |
| | CC | 0.691 | 0.649 | **0.782** | **0.751** | 0.724 | 0.693 |
| | MF | 0.368 | 0.350 | 0.513 | 0.517 | **0.564** | **0.584** |

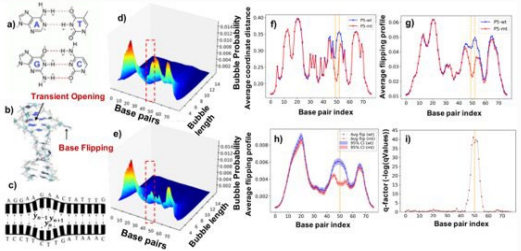
A Kabir and A Shehu. GOProFormer: A Multi-Modal Transformer Method for Gene Ontology Protein Function Prediction. Biomolecules 12(11):1709, 2022.

## Leveraging Cross-attention



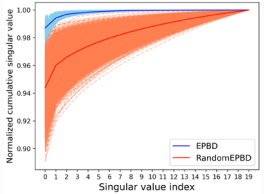

A Kabir, M Bhattarai, K Rasmussen, A Shehu, A Bishop, B Alexandrov, A Usheva. Advancing Transcription Factor Binding Site Prediction Using DNA Breathing Dynamics and Sequence Transformers via Cross Attention. BioRXiv 2024.

**Schematic** representation of the Metropolis-Hasting MCMC Algorithm for DNA State Transitions in pyDNA-EPBD. The EPBD model is a quasi-two-dimensional nonlinear model that describes the transverse opening motion of the complementary strands of double-stranded DNA, while distinguishing the two sides of the DNA double strand.

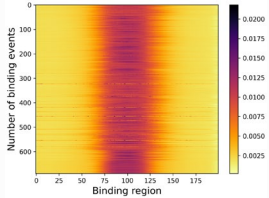



A Kabir, M Bhattarai, K Rasmussen, A Shehu, A Usheva, A Bishop, and B Alexandrov. Dynamically Active Zone of DNA Breathing. Bioinformatics 39(11):btad699, 2023.

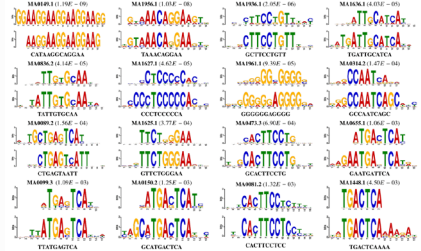
| Methods | AUROC | AUPR |
|---|---|---|
| DNABERT2-finetuned | 0.918 | 0.296 |
| RandEPBD-DNABERT2 | 0.917 | 0.296 |
| VanillaEPBD-DNABERT2-coord | 0.921 | 0.311 |
| VanillaEPBD-DNABERT2-coordflip | 0.920 | 0.308 |
| EPBD-DNABERT2-crossattn | 0.949 | 0.326 |




**Analysis**: Normalized cumulative singular values of 10,000 input DNA sequences corresponding to EPBD-DNABERT2-crossattn (blue) and randomly permuted EPBD features (red) shows that EPBD breathing dynamics guide DNABERT2 for extracting meaningful information.




**Analysis**: Average cross attention landscapes for all 690 TF-dna binding events on the test set shows that the model attends the proper binding regions as showing by the red color.



**Analysis**: Selected motifs from multi-modal framework analysis with cross attention weights for visual motif comparison using weblogos (**Fig. C**). The top panel shows the motif reported by the JASPAR 2022 database, and the bottom panel is the corresponding motifs found by EPBD-DNABERT2-crossattn in each panel. Our model can effectively pinpoint motifs with implications for understanding the complexities of gene regulation.

# Strengthening Biological Foundation Models via Grounding & Aligning

- The data at our disposal are in large part imperfect proxies of the underlying physics that governs the detailed behavior of molecular machineries in the living cell

- *Biological priors*:
- additional data modalities
- physics-based knowledge

- When contextualized to biological research, models need to be aligned with the underlying physics that governs all life
- However, our understanding of the underlying physics itself is distorted/narrowed through semi-empirical models
- Our approach: key metrics that capture physical realism *at various levels of sophistication* and *benchmark datasets* that truly expose model alignment in molecular biology

# Strengthening Biological Foundation Models via Grounding & Aligning

- The data at our disposal are in large part imperfect proxies of the underlying physics that governs the detailed behavior of molecular machineries in the living cell

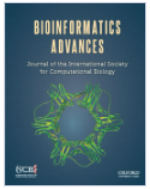
- *Biological priors*:
- additional data modalities
- physics-based knowledge

- When contextualized to biological research, models need to be aligned with the underlying physics that governs all life
- However, our understanding of the underlying physics itself is distorted/narrowed through semi-empirical models
- Our approach: key metrics that capture physical realism *at various levels of sophistication* and *benchmark datasets* that truly expose model alignment in molecular biology

OXFORD ACADEMIC | Journals | Books

BIOINFORMATICS ADVANCES

Issues | Advance articles | Submit ▾ | Alerts | About ▾ | Bioinformatics Advances



Volume 4, Issue 1
2024
(In Progress)

**Article Contents**



JOURNAL ARTICLE

## In the twilight zone of protein sequence homology: do protein language models learn protein structure?


Anowarul Kabir ✉, Asher Moldwin, Yana Bromberg, Amarda Shehu ✉



*Bioinformatics Advances*, Volume 4, Issue 1, 2024, vbae119,
https://doi.org/10.1093/bioadv/vbae119
**Published:** 17 August 2024 **Article history** ▾


PDF | Split View | Cite | Permissions | Share ▾


**Abstract**

**Motivation**

Protein language models based on the transformer architecture are increasingly improving performance on protein prediction tasks, including secondary structure, subcellular localization, and more. Despite being trained only on protein sequences, protein language models appear to implicitly learn protein structure. This paper investigates whether sequence representations learned by protein language models encode structural information and to what extent.

**Results**

We address this by evaluating protein language models on remote homology prediction, where identifying remote homologs from sequence information alone requires structural knowledge, especially in the "twilight zone" of very low sequence identity. Through rigorous testing at progressively lower sequence identities, we profile the performance of protein language models ranging from millions to billions of parameters in a zero-shot setting. Our


# oundation Models via Aligning

- When contextualized to biological research, models need to be aligned with the underlying physics that governs all life
- However, our understanding of the underlying physics itself is distorted/narrowed through semi-empirical models
- Our approach: key metrics that capture physical realism *at various levels of sophistication* and *benchmark datasets* that truly expose model alignment in molecular biology

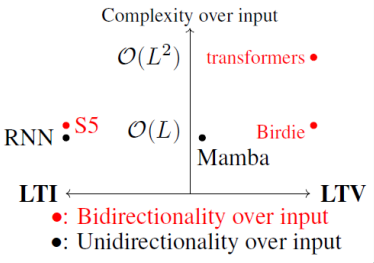
# The Quest for Small AI: Beyond the Transformer?

State Space Models (SSMs) are emerging as alternatives to Transformers but struggle with tasks needing long-range interactions, such as text copying and multi-query associative recall

Our contribution: a minimalist SSM architecture with novel pre-training objectives and a dynamic mixture of pre-training objectives via reinforcement learning

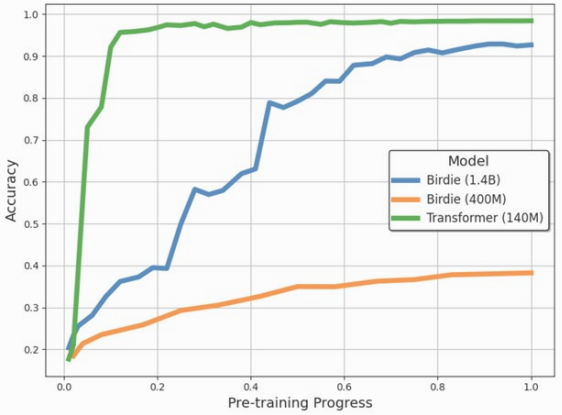
## Catching up to the Transformer with State Space Models: **Birdie**

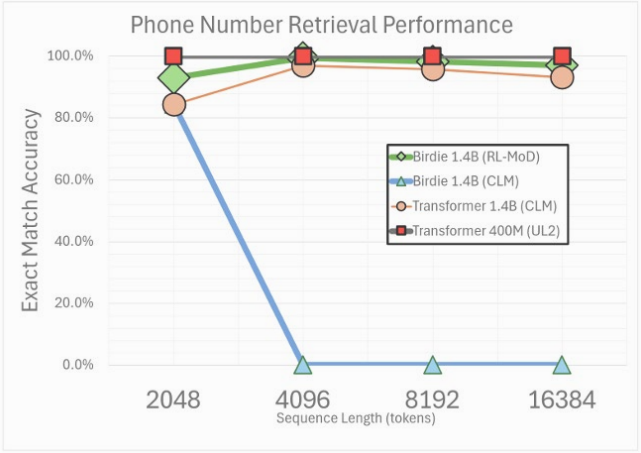
Linear time-varying (LTV) systems improve language modeling capabilities compared to linear time-invariant (LTI) counterparts. A comprehensive analysis of LTV vs. LTI systems for NLP, controlling for various training objectives and fine-tuning remains unexplored.




### Birdie

$$\mathbf{z}_t = \mathbf{x_t} \odot \tau(\mathbf{W}^K \mathbf{x}_t) \in \mathbb{R}^N$$
$$\mathbf{f}_t = \sigma(\mathbf{W}^f \mathbf{x}_t) \in \mathbb{R}^N$$
$$\mathbf{h}_t = \mathbf{f}_t \odot \mathbf{h}_{t-1} + \mathbf{z}_t$$
$$\mathbf{y}_t = \mathbf{W}^{out} \mathbf{h}_t \odot \gamma(\mathbf{W}^{side} \mathbf{x}_t)$$




A comprehensive comparison pitches Birdie against transformer-based models at various configurations (base versus instruction fine tuning, 400M versus 1.4Bparameters, and various pre-training objectives) over14max-likelihoodtasks from the Eleuther AI LMHarness

| Model | Objective | Avg Task Accuracy |
|---|---|---|
| **Instruction Tuned, 1.4B** | | |
| Birdie (RL-MoD) | RL-MoD | **45.5%** |
| Attention (CLM) | CLM | 43.0% |
| Birdie (PT5) | PT5 | 42.5% |
| Birdie (CLM) | CLM | 40.9% |
| **Base Models, 1.4B** | | |
| Birdie (PT5) | PT5 | **41.0%** |
| Birdie (CLM) | CLM | 40.9% |
| Birdie (RL-MoD) | RL-MoD | 40.6% |
| Attention (CLM) | CLM | 40.1% |
| **Instruction Tuned, 400M** | | |
| Birdie (UL2) | UL2 | **40.3%** |
| Attention (UL2) | UL2 | 40.2% |
| Hawk (PT5) | PT5 | 39.3% |
| Attention (CLM) | CLM | 39.2% |
| Hawk (CLM) | CLM | 38.4% |
| **Base Models, 400M** | | |
| Birdie (CLM) | CLM | **40.3%** |
| Birdie (RL-MoD) | RL-MoD | 40.1% |
| Attention (CLM) | CLM | 39.7% |
| Birdie (UL2) | UL2 | 39.5% |
| Birdie (PT5) | PT5 | 39.3% |
| Attention (UL2) | UL2 | 39.2% |
| Hawk (PT5) | PT5 | 38.8% |
| Hawk (CLM) | CLM | 38.4% |

# Students and Collaborators

Fardina Alam

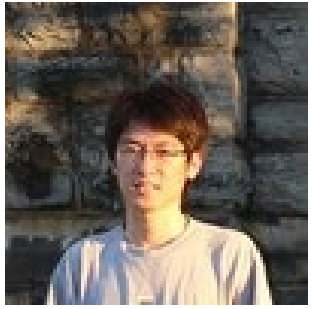

Toki Inan

Taseef Rahman

Samuel Blouir

Manpriya Dua

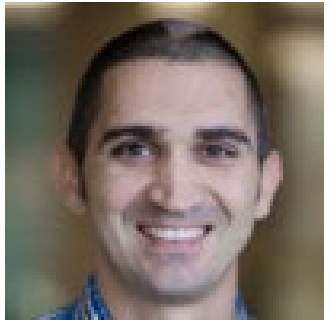
Anowarul Kabir



Liang Zhao
Emory University

Yana Bromberg
Emory University

Kevin Minbiole
Villanova University



William Wuest
Emory University

Saleh Alkhalifa
Amgen

J.P. Singh
George Mason

My heart-felt apologies to the 3 postdoctoral students, 23 PhD students, 14 MS students, 38 undergraduate students, and 19 high-school students that could not be placed in one slide but have been a privilege to work with and pursue scientific research together.

… and to my collaborators, Ruth Nussinov (NIH), Chen Keasar (Ben Gurion), Bojan Alexandrov (LANL), and many others that have enriched my scientific journey and have made me a better scientist and person.