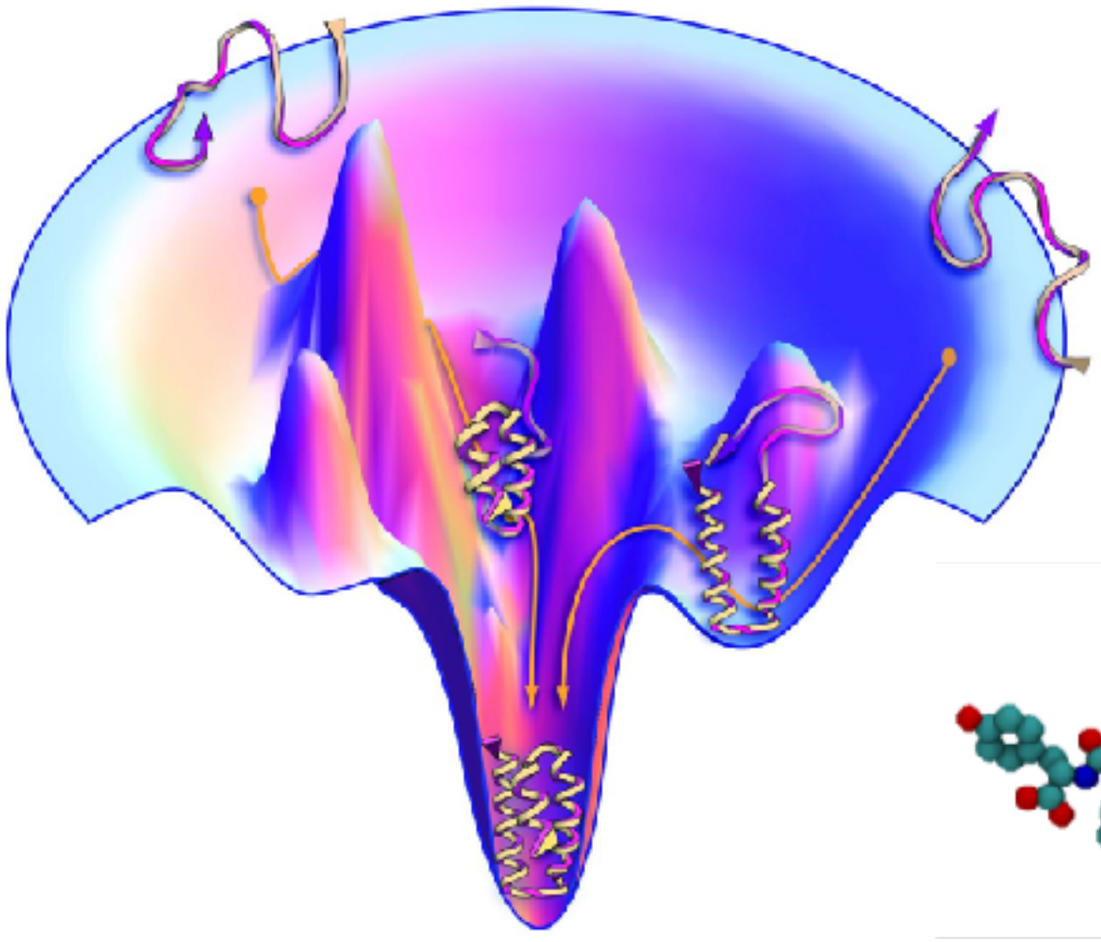
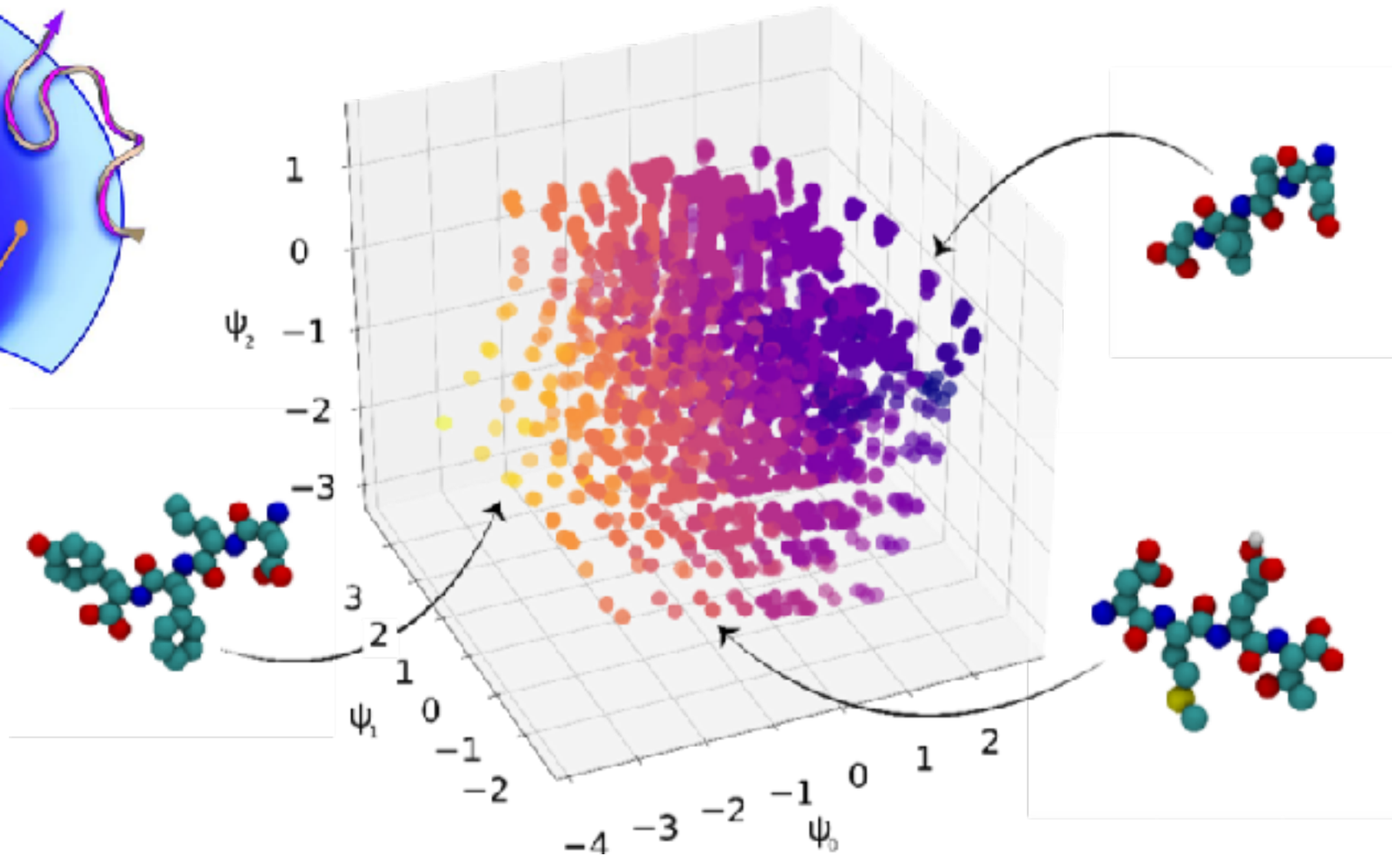
Machine learning and data science in soft materials design and engineering





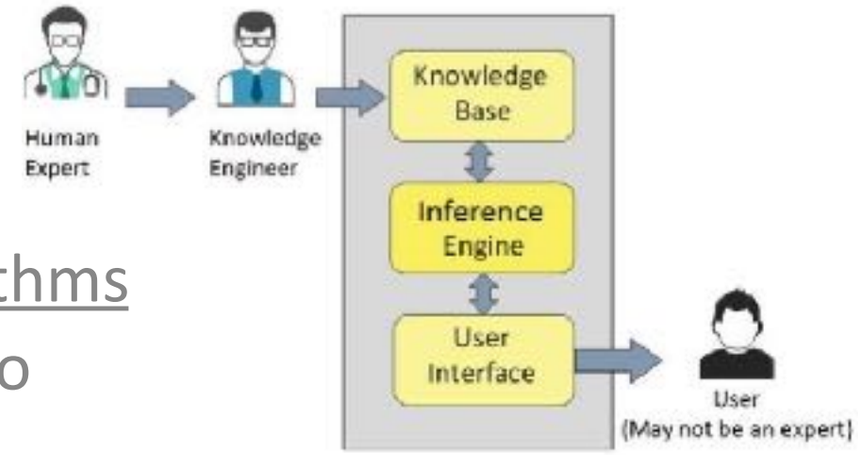
IPAM MLP Tutorial Talk
9 September 2019
Andrew Ferguson, U. Chicago

THE UNIVERSITY OF CHICAGO

PRITZKER SCHOOL OF MOLECULAR ENGINEERING
THE UNIVERSITY OF CHICAGO

Three waves of AI

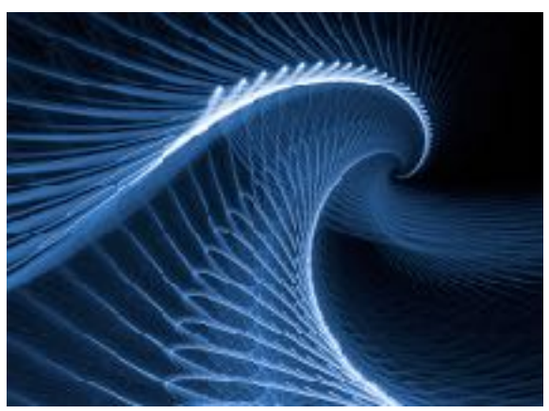
1 Expert systems



Conventional rule based if-then algorithms programmed with expert knowledge to mimic human decision making

2 Machine learning

Algorithms that train themselves to learn the rules from the data



3 Explainable AI (XAI) + Physics-aware AI (PAI)

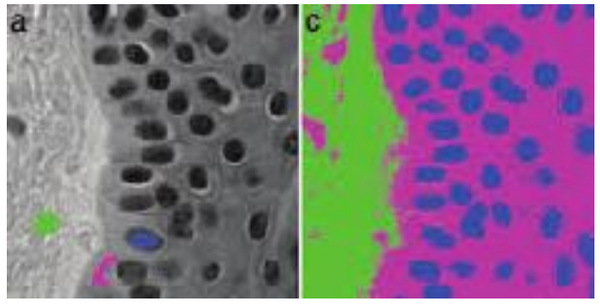
Algorithms that explain their actions and/or respect and exploit physical laws

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf https://www.fbo.gov/spg/ ODA/DARPA/CMO/DARPA-BAA-16-53/listing.html

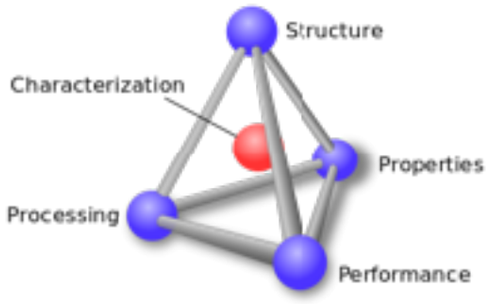
AI in molecular and materials science

- AI particularly valuable for:
  (i) high-dimensional and/or complex systems that foil human intuition
  (ii) large conformational or combinatorial search spaces
  (iii) inverse problems — data but not models, goals but not mechanisms

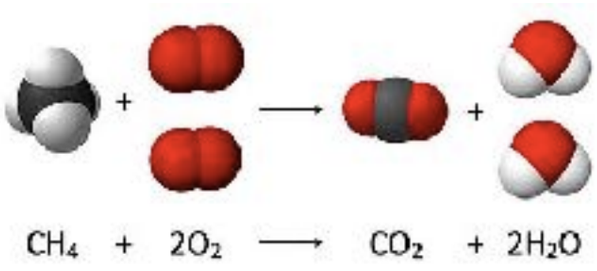
chemical discovery (i, ii)
— teaching machines chemical intuition and search



medical diagnosis (i, iii)
— pathology identification, intervention design



materials engineering (i, ii, iii)
— inverse materials design via the what (and the why)

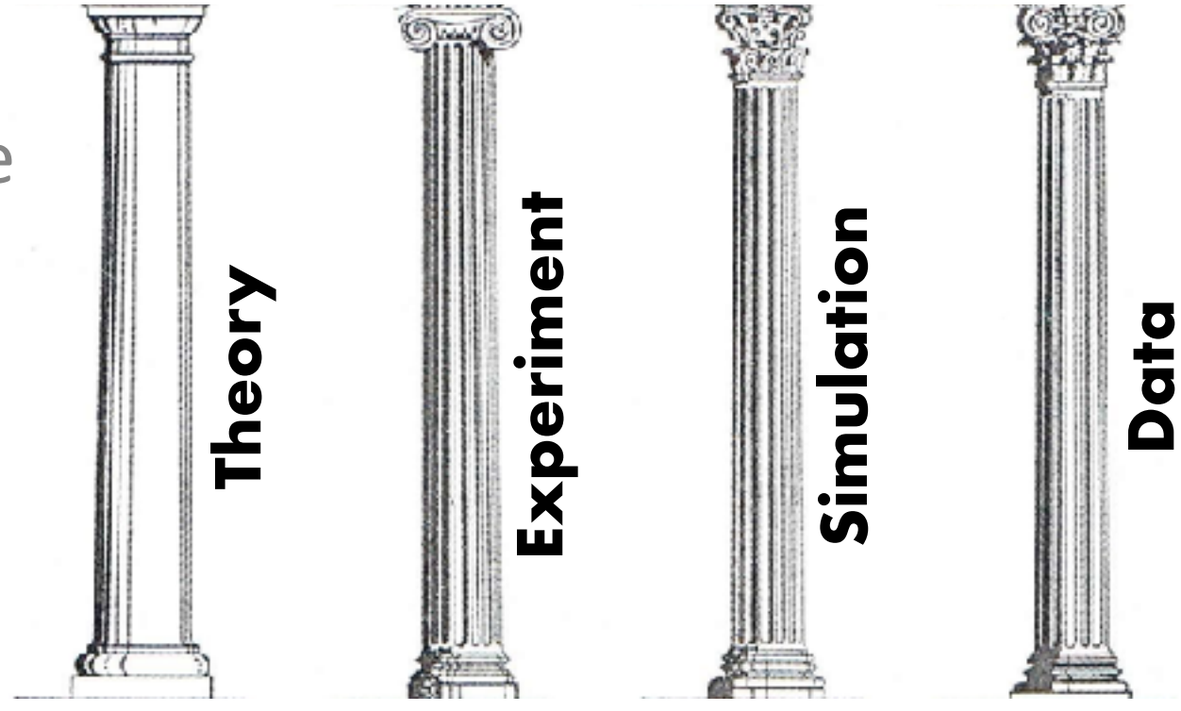


reaction engineering (ii, iii)
— optimizing conditions, predictive (retro)synthesis

ALF *ACS Central Science* 4 8 938-941 (2018)

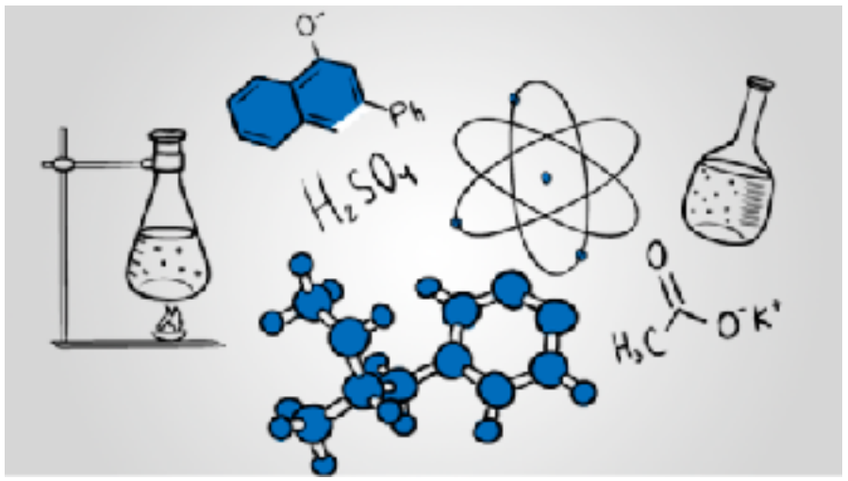
Data science & domain science

- Data-driven inquiry emerging as a **"fourth pillar"** of science — knowledge discovery from data (KDD)



- Success is contingent on integration of **data science paradigms and tools** with **domain specific knowledge and expertise** (thermo, QM, rxn eng, …)

+



=

A. Agrawal and A. Choudhary APL Mater. 4 053208 (2016) https://d3f1iyfxxz8i1e.cloudfront.net/courses/course_image/4a0821d298ee.png

Review articles

MOLECULAR SIMULATION
2018, VOL. 44, NOS. 13–14, 1090–1107
https://doi.org/10.1080/08927022.2017.1400164

Taylor & Francis
Taylor & Francis Group

Check for updates

Nonlinear machine learning in simulations of soft and biological materials

J. Wang[a] and A. L. Ferguson[a,b,c]

[a]Department of Physics, University of Illinois Urbana-Champaign, Urbana, IL, USA; [b]Department of Materials Science and Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA; [c]Department of Chemical and Biomolecular Engineering, University of Illinois Urbana-Champaign, Urbana, IL, USA

ABSTRACT

Interpretable parameterisations of free energy landscapes for soft and biological materials calculated from molecular simulation require the availability of 'good' collective variables (CVs) capable of discriminating the metastable states of the system and the barriers between them. If these CVs are coincident with the slow collective modes governing the long-time dynamical evolution, then they also furnish good coordinates in which to perform enhanced sampling to surmount high free energy barriers and efficiently explore and recover the landscape. Non-linear manifold learning techniques provide a means to systematically extract such CVs from molecular simulation trajectories by identifying and extracting low-dimensional manifolds lying latent within the high-dimensional coordinate space. We survey recent advances in data-driven CV discovery and enhanced sampling using non-linear manifold learning, describe the mathematical and theoretical underpinnings of these techniques, and present illustrative examples to molecular folding and colloidal self-assembly. We close with our outlook and perspective on future advances in this rapidly evolving field.

ARTICLE HISTORY
Received 11 August 2017
Accepted 29 October 2017

KEYWORDS
Enhanced sampling; free energy landscapes; non-linear manifold learning; protein folding; self-assembly

IOP Publishing
J. Phys.: Condens. Matter **30** (2018) 043002 (27pp)
Journal of Physics: Condensed Matter
https://doi.org/10.1088/1361-648X/aa98bd

Topical Review

Machine learning and data science in soft materials engineering

Andrew L Ferguson

Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, 1304 West Green Street, Urbana, IL 61801, United States of America
Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, 600 South Mathews Avenue, Urbana, IL 61801, United States of America
Department of Physics, University of Illinois at Urbana-Champaign, 1110 West Green Street, Urbana, IL 61801, United States of America
Frederick Seitz Materials Research Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States of America
Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States of America

E-mail: alf@illinois.edu

Molecular Systems Design & Engineering

ROYAL SOCIETY OF CHEMISTRY

EDITORIAL

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Mol. Syst. Des. Eng.*, 2018, 3, 429

DOI: 10.1039/c8me90007h

rsc.li/molecular-engineering

Machine learning and data science in materials design: a themed collection

Andrew Ferguson [abc] and Johannes Hachmann [def]

Guest Editors Andrew Ferguson and Johannes Hachmann introduce this themed collection of papers showcasing the latest research leveraging data science and machine learning approaches to guide the understanding and design of hard, soft, and biological materials with tailored properties, function and behaviour.

ACS central science

EDITORIAL

Andrew L. Ferguson
Institute for Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States

ACS Central Science Virtual Issue on Machine Learning

EDITORIAL

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Mol. Syst. Des. Eng.*, 2019, 4, 462

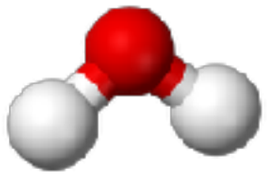
Conference report: 2018 materials and data science hackathon (MATDAT18)

Andrew L. Ferguson, *[a] Tim Mueller, [b] Sanguthevar Rajasekaran[c] and Brian J. Reich[d]

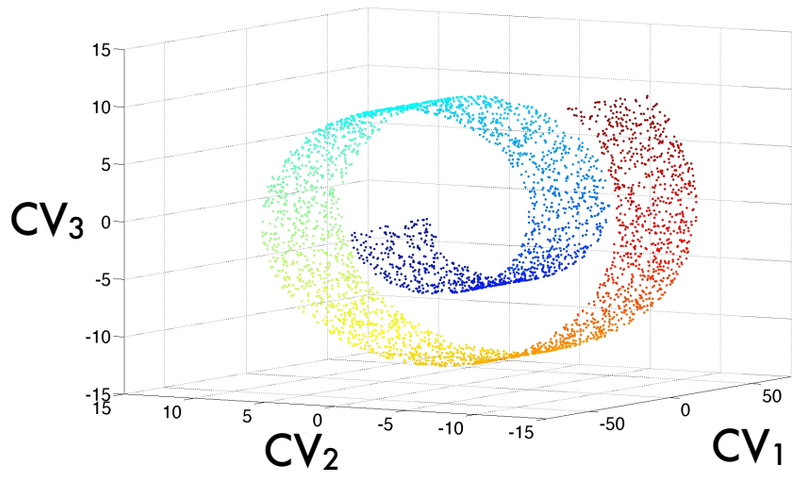
The National Science Foundation (NSF) 2018 Materials and Data Science Hackathon (MATDAT18) took place at the Residence Inn Alexandria Old Town/Duke Street, Alexandria, VA over the period May 30–June 1, 2018. This three-day collaborative "hackathon" or "datathon" brought together teams of materials scientists and data scientists to collaboratively engage materials science problems using data science tools. The materials scientists brought a diversity of problems ranging from inorganic material bandgap prediction to

ML and DS in materials design and engineering

1 Classical molecular dynamics in 15 minutes



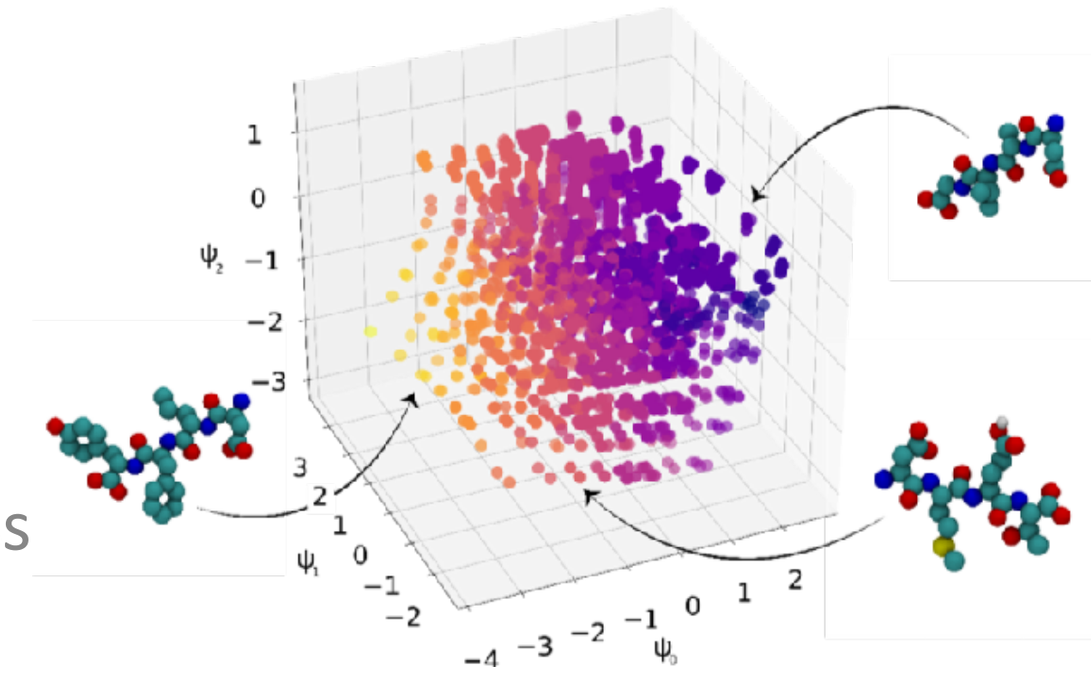
2 Enhanced sampling in molecular simulation
[ML-driven search of <u>conformational</u> space]



- collective variable (CV) discovery
- accelerated sampling in molecular simulations
- **APPLICATION**: enhanced sampling of protein folding w/ auto-encoding ANNs

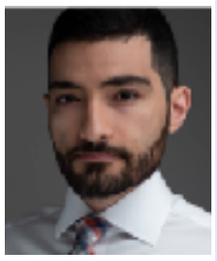
3 Data-driven design of self-assembling π-conjugated oligopeptides
[DS-driven search of <u>chemical</u> space]

- surrogate model construction
- high-throughput virtual screening
- **APPLICATION**: oligopeptide discovery w/ coarse-grained molecular simulation, variational autoencoders, Gaussian process regression, and active learning



Acknowledgements



Collab: **Prof. J.D. Tovar (JHU)**
Prof. Howard Katz (JHU)
Prof. André Schleife (UIUC)

GS: **Wei Chen**
Yutao Ma
Kirill Shmilovich
Beryl Zhang
Xinran Lian
Nick Rego
Shiqi Chen

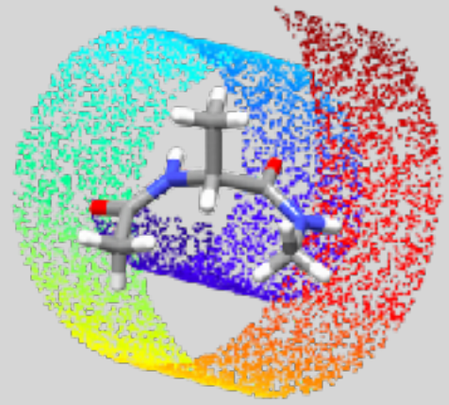
UG: Olivia Dunne
Gillian Shen
Joseph Aulicino

Post-Docs:
Dr. Hythem Sidky
Dr. Brandon Peters
Dr. Mingfei Zhao

THE UNIVERSITY OF CHICAGO

PRITZKER SCHOOL OF MOLECULAR ENGINEERING
THE UNIVERSITY OF CHICAGO

Ferguson Laboratory



THE INSTITUTE FOR MOLECULAR ENGINEERING

http://andrewferguson.uchicago.edu

MolSSI

NSF

ACI-1547580

MICCoM

5J-30161-0021A

DMR-1841800
CHE-1841805
DMS-1841810
DMR-1841807
DMR-1844505

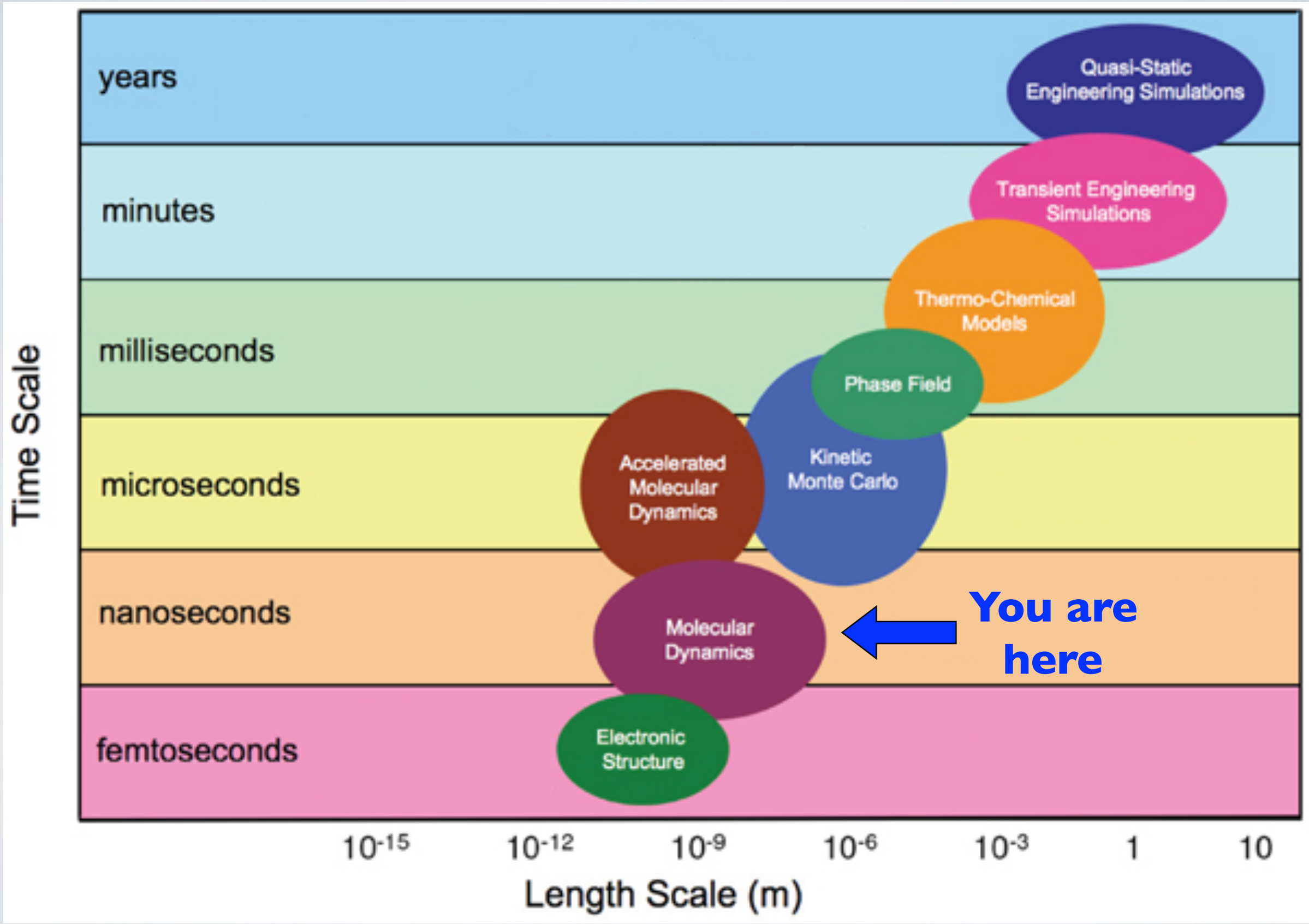
U.S. DEPARTMENT OF ENERGY

DE-SC0019288

1. Classical molecular dynamics in 15 minutes

2. ANN accelerated sampling of molecular free energy landscapes **[ML-driven search of conformational space]**

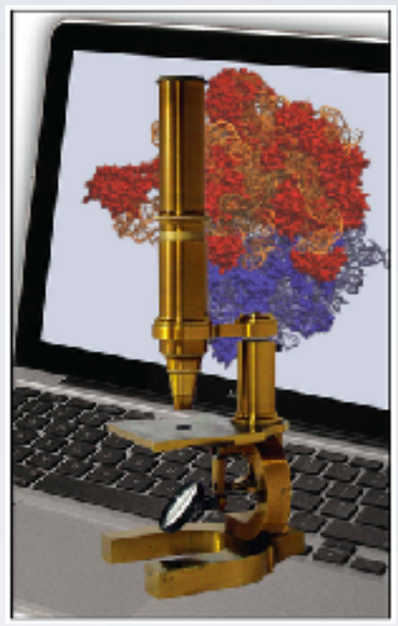
3. Data-driven design of π-conjugated oligopeptides **[DS-driven search of chemical space]**



http://www.lbl.gov/CS/html/exascale4energy/nuclear.html

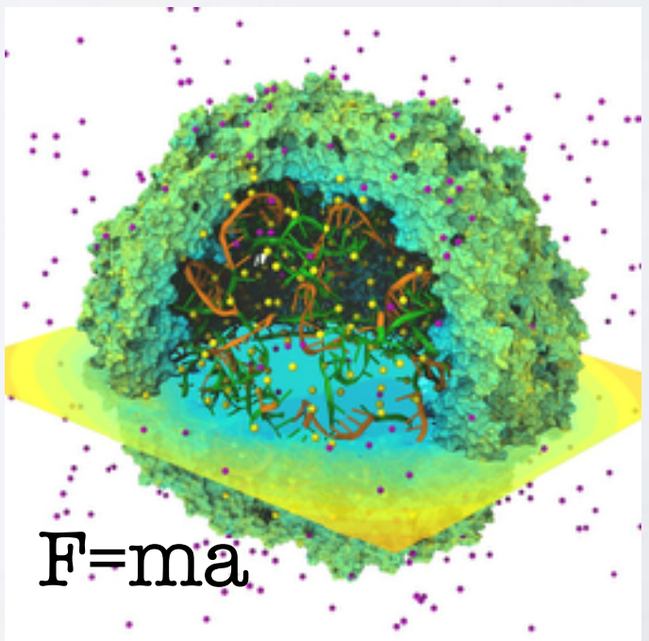
What is molecular dynamics?

- A computational microscope



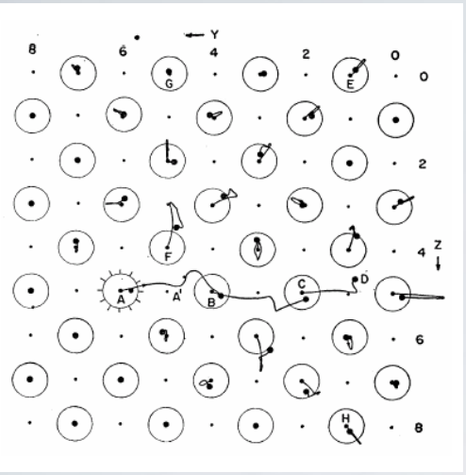
- An experiment on a computer

- A simulation of the classical mechanics of atoms

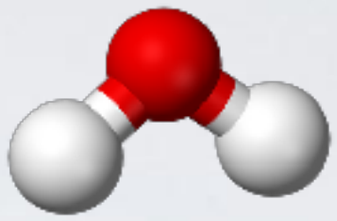


http://www.ks.uiuc.edu http://www.123rf.com

Milestones in MD



1960
Gibson *et al.*
Simulation of Cu radiation damage
Gibson, J.B., Goland, A.N., Milgram, M., and Vineyard, G.H. Phys. Rev. 120 1229 (1960)

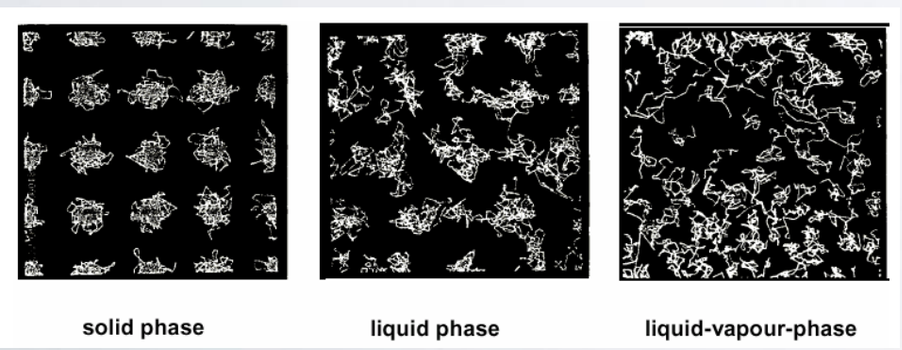


1974
Rahman & Stillinger
First simulation of liquid water
Stillinger, F.H. and Rahman, A.J. Chem. Phys. 60 1545 (1974)

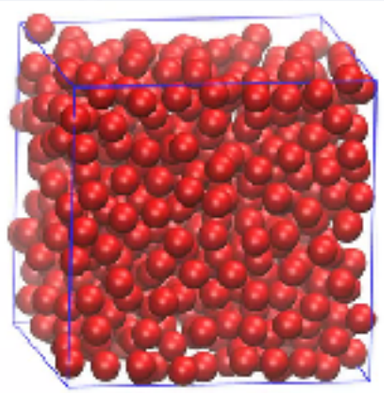
1994
York *et al.*
BPTI hydrated xtal [1ns]
York, D.M., Wlodawer, A., Pedersen, L.G. and Darden, T.A. PNAS 91 18 8715 (1994)

2010
Shaw *et al.*
BPTI in water [1ms]
Shaw, D.E. et al. Science 330 341 (2010)

1957
Alder & Wainwright
First MD simulation of hard sphere fluid
Alder, B.J. and Wainwright, T..E. J. Chem. Phys. 27 1208 (1957)



1964
Rahman
First simulation of liquid Ar using realistic potential
Rahman, A. Phys. Rev. A136 405 (1964)



1977
McCammon *et al.*
First protein simulation (BPTI) [8.8ps]
McCammon, J.A., Gelin, B.R., and Karplus, M. Nature 267 585 (1977)

1998
Duan & Kollman
Villin headpiece in water [1μs]
Duan, Y., and Kollman, P.A. Science 282 5389 740 (1998)

The fundamental idea

- MD simulates atomic motions using classical mechanics
- Running a simulation is like cooking - just follow the recipe
- Three ingredients:

1. An initial system configuration

$$[\vec{r}(t=0), \vec{v}(t=0)]$$

2. A (classical) interaction potential for the system

$$V(\vec{r})$$

3. A way to integrate F=ma

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2$$

The fundamental idea

- Laplace's Demon / "The Clockwork Universe"

"Given for one instant an intelligence which could **comprehend all the forces by which nature is animated** and the **respective positions of the beings which compose it**, if moreover this intelligence were vast enough to **submit these data to analysis**, it would embrace in the same formula both the movements of the largest bodies in the universe and those of the lightest atom; to it **nothing would be uncertain, and the future as the past would be present to its eyes**."

- Pierre Simon de Laplace (1749-1827)

This is essentially molecular dynamics

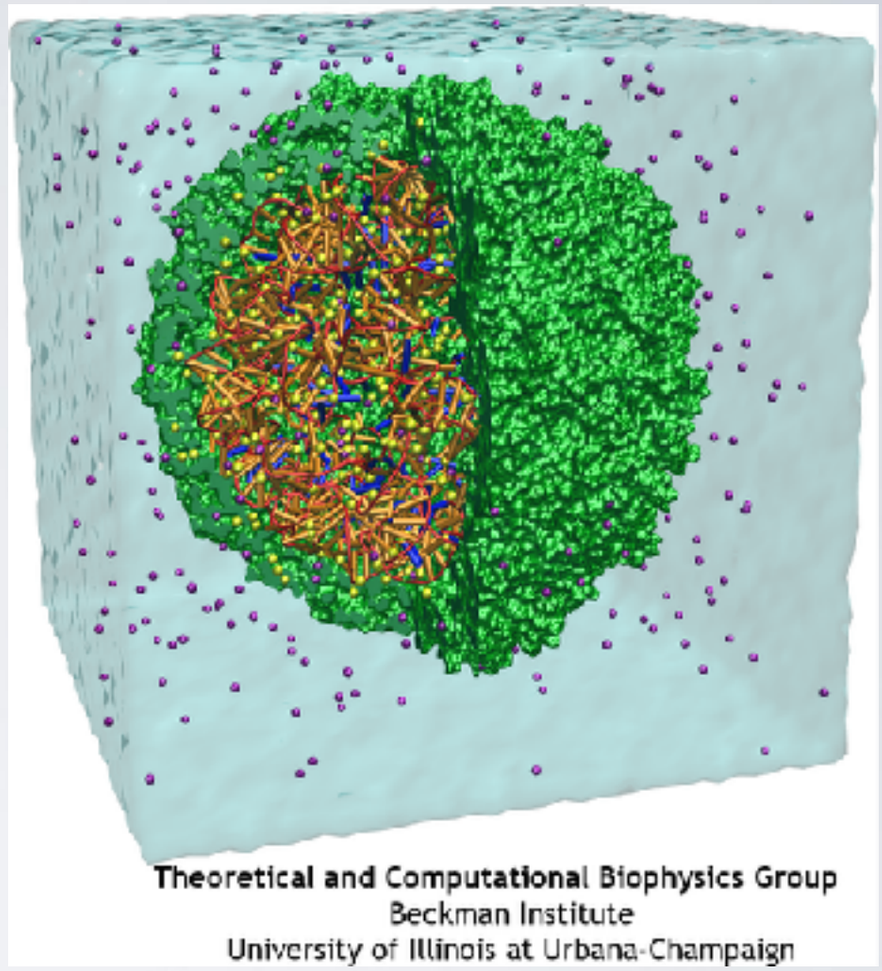
Ingredient 1: Initial configuration

- Specification of initial atomic coordinates and velocities

- Classical mechanics is deterministic: **initial state and interaction rules fully specify the system's future***

- Wind up Laplace's clockwork universe and — in principle — a "vast intelligence" could compute the future of the system

- Our intelligence is insufficiently vast — the equations are hard! — and thus **we resort to numerical simulation**



* neglecting numerical integration errors and finite precision (i.e., uncertainty)

Initializing coordinates

- Initial configurations can be generated "by hand" or short scripts for simple systems (e.g., liquid Ar, bulk Al)

- Software tools for complex systems (e.g., proteins, complex defect structures)

PRODRG (http://davapc1.bioch.dundee.ac.uk/prodrg/)
ATP (http://compbio.biosci.uq.edu.au/atb/)
PyMOI (http://www.pymol.org/)
Chimera (http://www.cgl.ucsf.edu/chimera/)

- Common protein structures are in Protein Data Bank

PDB (www.rcsb.org/pdb)

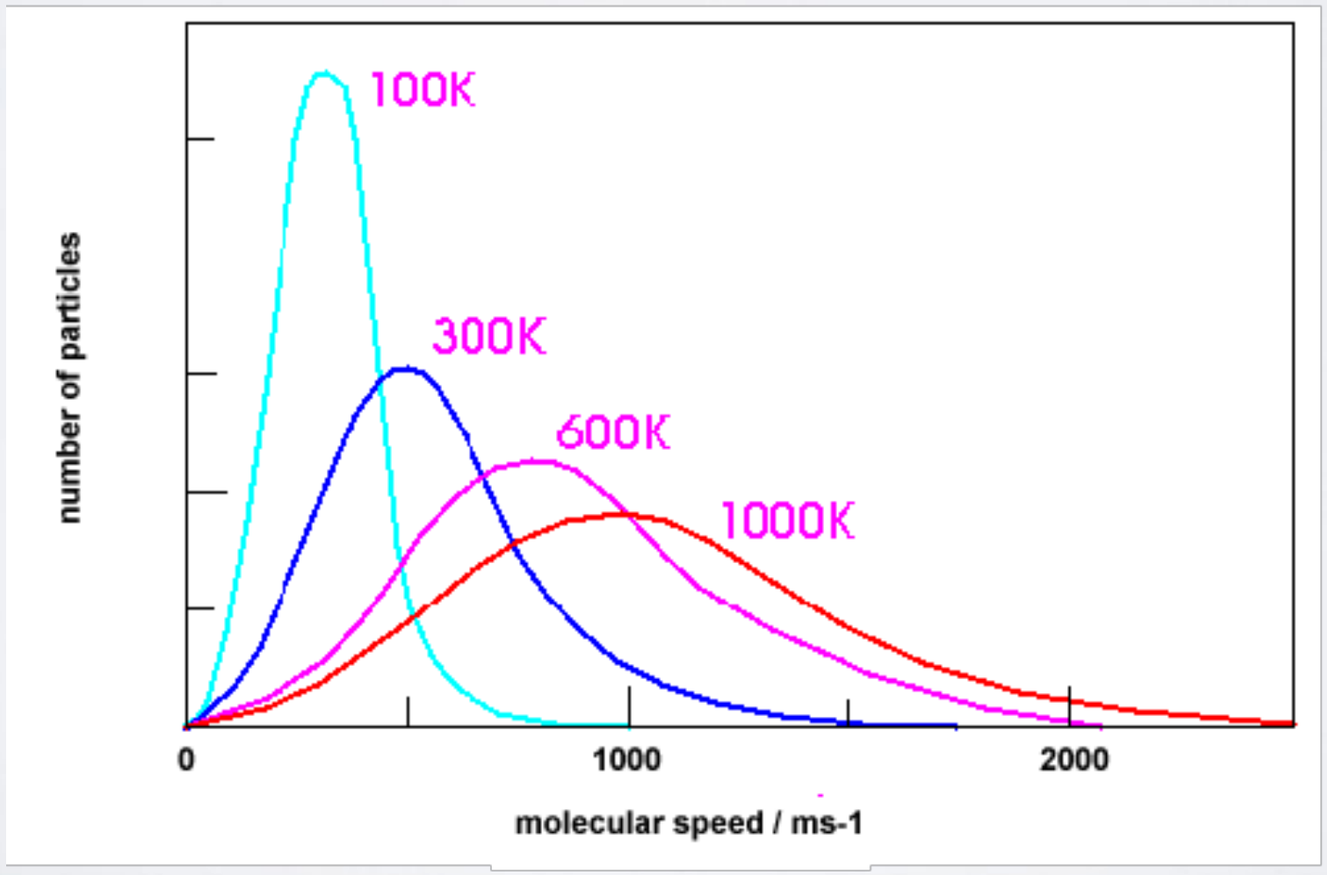
```
Protein in water
 2626
    1ACE     CH3    1   0.654   2.519   0.492  0.1151 -0.0284  0.0138
    1ACE    HH31    2   0.740   2.540   0.554  0.2235  0.0824 -0.1715
    1ACE    HH32    3   0.605   2.433   0.538  3.1239 -1.7508  0.2704
    1ACE    HH33    4   0.684   2.482   0.394  0.2995  1.4351 -0.5063
    1ACE       C    5   0.553   2.633   0.481 -0.0173 -0.1643 -0.2114
    1ACE       O    6   0.445   2.613   0.535 -0.0062 -0.0674 -0.1518
    2ALA       N    7   0.582   2.739   0.405  0.1733  0.1955  0.3558
    2ALA       H    8   0.510   2.806   0.379  2.0591  1.7509 -1.1449
    2ALA      CA    9   0.705   2.781   0.341 -0.1656 -0.5238 -0.7826
    2ALA      HA   10   0.741   2.700   0.278 -1.5076 -1.1917 -0.7488
    2ALA      CB   11   0.674   2.911   0.267  0.4673 -0.0071 -0.1476
    2ALA     HB1   12   0.611   2.896   0.179 -2.0184 -0.1132  1.5667
    2ALA     HB2   13   0.628   2.977   0.340  0.9533 -0.2065  0.3439
    2ALA     HB3   14   0.763   2.957   0.225  0.9167 -0.2257  0.5469
    2ALA       C   15   0.813   2.805   0.445 -0.7286 -0.5024 -0.1928
    2ALA       O   16   0.783   2.866   0.547  0.1974 -0.4451  0.0528
    3NAC       N   17   0.941   2.777   0.419 -0.5125  0.1136  0.1784
    3NAC       H   18   1.000   2.799   0.497  0.1647 -1.3605  0.1187
    3NAC     CH3   19   1.001   2.723   0.298 -0.7672 -0.2750  0.2229
    3NAC    HH31   20   1.092   2.669   0.324  0.3722  1.1812 -0.5828
    3NAC    HH32   21   0.945   2.648   0.243  1.0207 -0.0997 -1.9789
    3NAC    HH33   22   1.030   2.810   0.238 -2.1192 -0.7269 -1.1621
    4SOL      OW   23   0.784   1.392   0.792  0.1855 -0.2071  0.1377
    4SOL     HW1   24   0.735   1.315   0.761 -1.0746  1.1108 -1.3153
    4SOL     HW2   25   0.719   1.445   0.839  1.3389 -0.5885  2.3128
    5SOL      OW   26   0.428   0.234   2.288  1.2957 -0.4548 -0.0720
    5SOL     HW1   27   0.411   0.170   2.219 -0.2175  0.3118 -0.4516
    5SOL     HW2   28   0.488   0.297   2.247  3.0259 -1.7375  0.3978
    6SOL      OW   29   0.166   0.601   2.571 -0.1148  0.6829 -0.6515
    6SOL     HW1   30   0.212   0.681   2.595 -0.5922  0.6213  0.5401
    6SOL     HW2   31   0.228   0.552   2.517  1.4295  0.3667  1.2935
    7SOL      OW   32   2.575   0.438   1.811  0.4391  0.2071  0.3094
    7SOL     HW1   33   2.581   0.469   1.721 -1.3349  0.1731  0.1541
    7SOL     HW2   34   2.481   0.429   1.828  0.6643  1.2137  2.4877
    8SOL      OW   35   0.492   2.063   2.222 -0.4334 -0.0059 -0.1953
    8SOL     HW1   36   0.570   2.035   2.269 -0.2720 -1.2784 -1.1564
    8SOL     HW2   37   0.450   2.127   2.279  0.5359 -0.3976  0.9797
    9SOL      OW   38   2.657   0.259   0.784  0.3737 -0.2806  0.0046
    9SOL     HW1   39   2.659   0.233   0.692 -1.4133  0.9624 -0.4269
    9SOL     HW2   40   2.714   0.335   0.789  1.6804 -1.2503  0.2641
   10SOL      OW   41  -0.009   1.802   0.210  0.2163  0.8744 -0.2151
   10SOL     HW1   42  -0.046   1.724   0.251 -0.3127  1.2546  0.0424
   10SOL     HW2   43   0.080   1.807   0.244  0.7693 -0.4235 -1.3548
   11SOL      OW   44   0.693   2.604   2.223 -0.8870 -0.4375  0.1438
   11SOL     HW1   45   0.641   2.585   2.302 -0.5618 -3.2331 -0.1923
   11SOL     HW2   46   0.772   2.647   2.256 -0.6655 -1.7422  1.4208
   12SOL      OW   47   2.600   2.648   2.637  0.3128 -0.3491  0.5421
   12SOL     HW1   48   2.615   2.621   2.547 -0.1552 -1.3876  0.7622
```

Initializing velocities

- Bad idea to start atoms from rest (absolute zero = 0 K) due to thermal shock upon starting simulation

- Standard approach is to draw velocities randomly from a Maxwell-Boltzmann distribution at the temperature, T

$$f_{\mathbf{v}}(v_x, v_y, v_z) = \left(\frac{m}{2\pi kT}\right)^{3/2} \exp\left[-\frac{m(v_x^2 + v_y^2 + v_z^2)}{2kT}\right]$$



http://ibchem.com/IB/ibfiles/states/sta_img/MB2.gif

Ingredient 2: Interaction potentials

- The net force acting on each atom in the system is a result of its interactions with all other atoms

- These interaction amount to a set of rules known as a **force field** or **interaction potential**

- Accurate, robust, and transferable force fields are critical to perform physically realistic molecular simulations

- Force field development is an academic industry

metals:	EAM (Daw & Baskes), MEAM (Baskes)
biomolecules:	Amber (Kollman, UCSF), GROMOS (U. Groningen), CHARMM (Karplus, Harvard), OPLS (Jorgensen, Yale), MARTINI [coarse grained] (Marrink, U. Groningen)
polymers:	TraPPE (Siepmann, U. Minnesota), MM2 (Allinger, UGA)
water:	SPC (Berendsen), SPC/E (Berendsen), TIPnP(Jorgensen), ST2 (Stillinger & Rahman)
general:	DREIDING (Mayo et al.), DISCOVER(Rappe et al.), UFF (Hagler et al.)

Energy, force, and acceleration

- The potential energy of the system is a complicated function of atomic coordinates (this is why we have to *simulate numerically* rather than *calculate analytically*)

- The net force on atom i is the negative gradient of the potential energy wrt the atomic coordinates

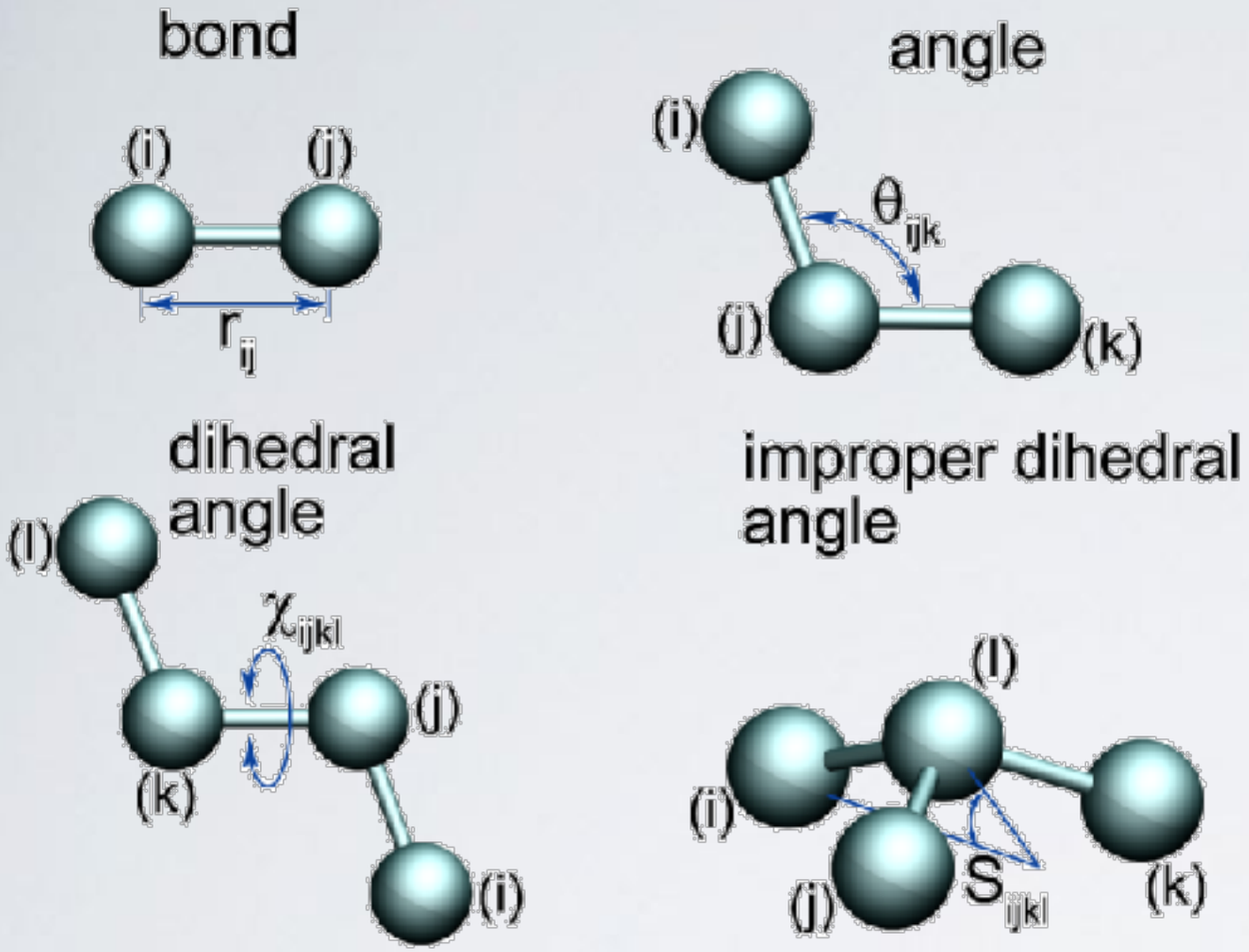
$$F_i = -\nabla_i [V(r_1, r_2, ..., r_N)]$$

$$a_i = \frac{F_i}{m_i}$$

- The potential energy is typically broken into four parts:

$$V(\vec{r}) = V_{bonded} + V_{non-bonded} + V_{restraints} + V_{field}$$

Bonded

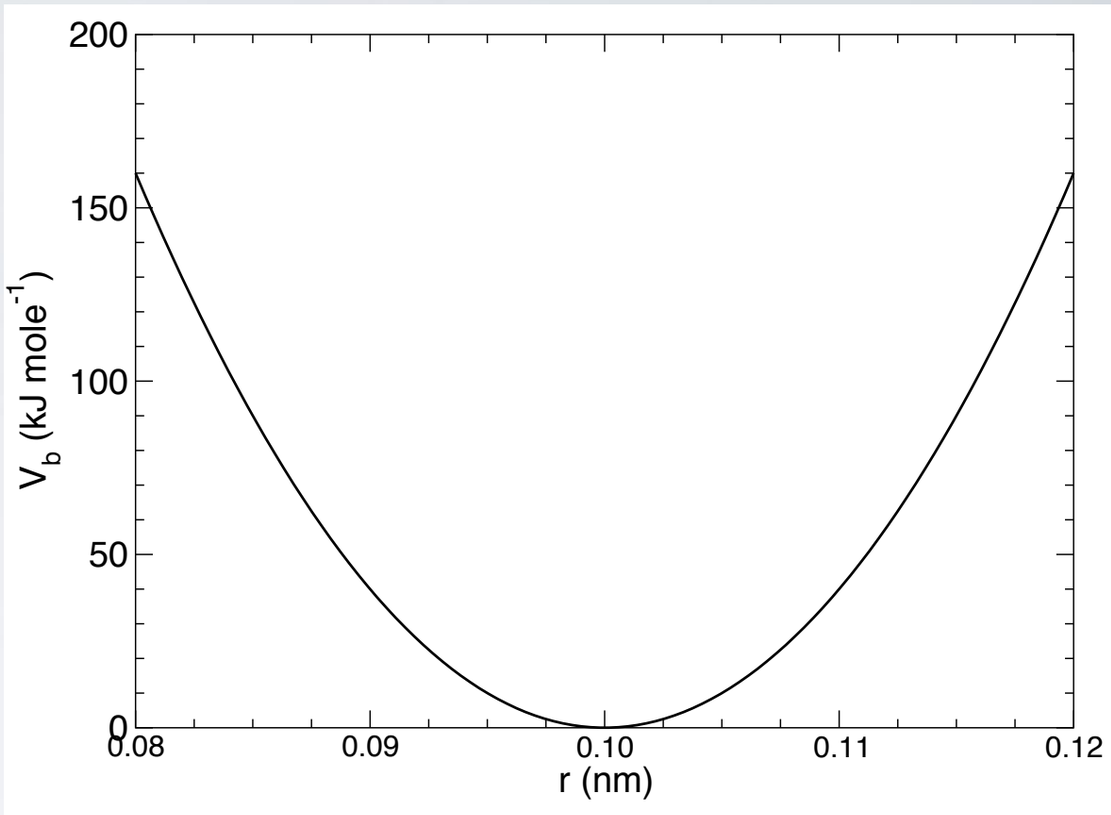


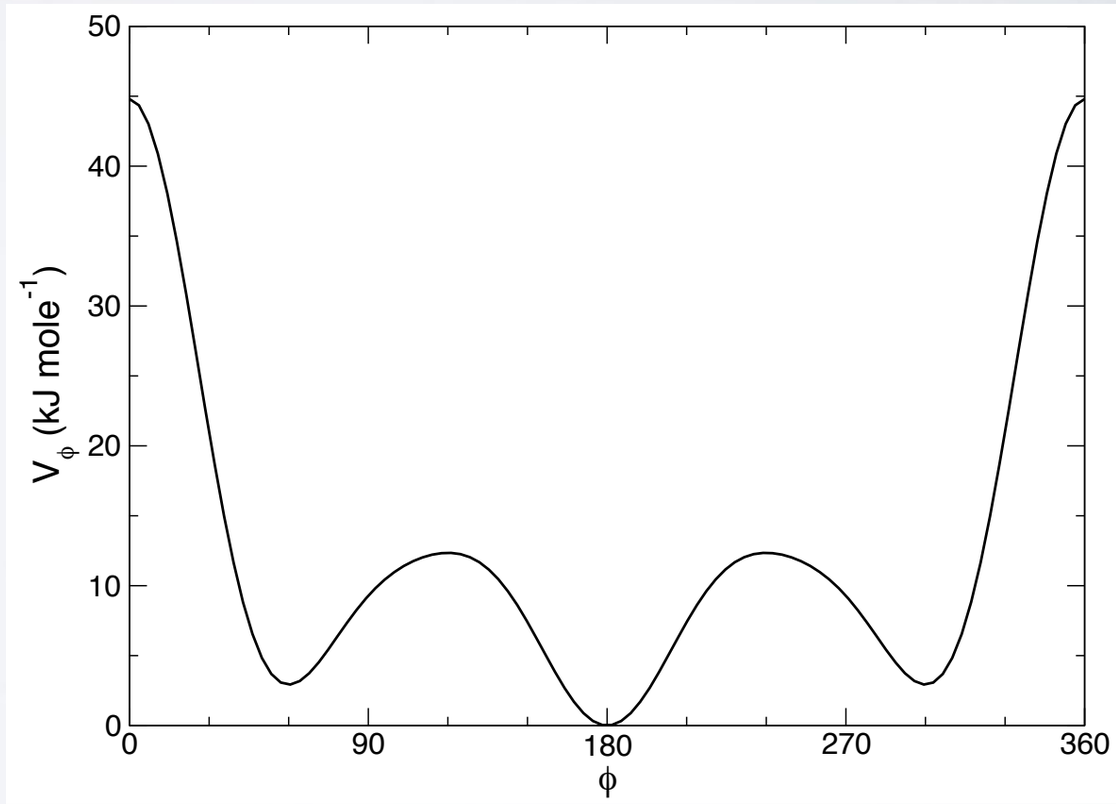
$$V_b\left(r_{ij}\right) = \frac{1}{2}k^b_{ij}(r_{ij} - b_{ij})^2$$

$$V_a(\theta_{ijk}) = \frac{1}{2}k^\theta_{ijk}(\theta_{ijk} - \theta^0_{ijk})^2$$

$$V_{rb}(\phi_{ijkl}) = \sum_{n=0}^{5} C_n(\cos(\psi))^n$$

$$V_{id}(\xi_{ijkl}) = \frac{1}{2}k_\xi(\xi_{ijkl} - \xi_0)^2$$





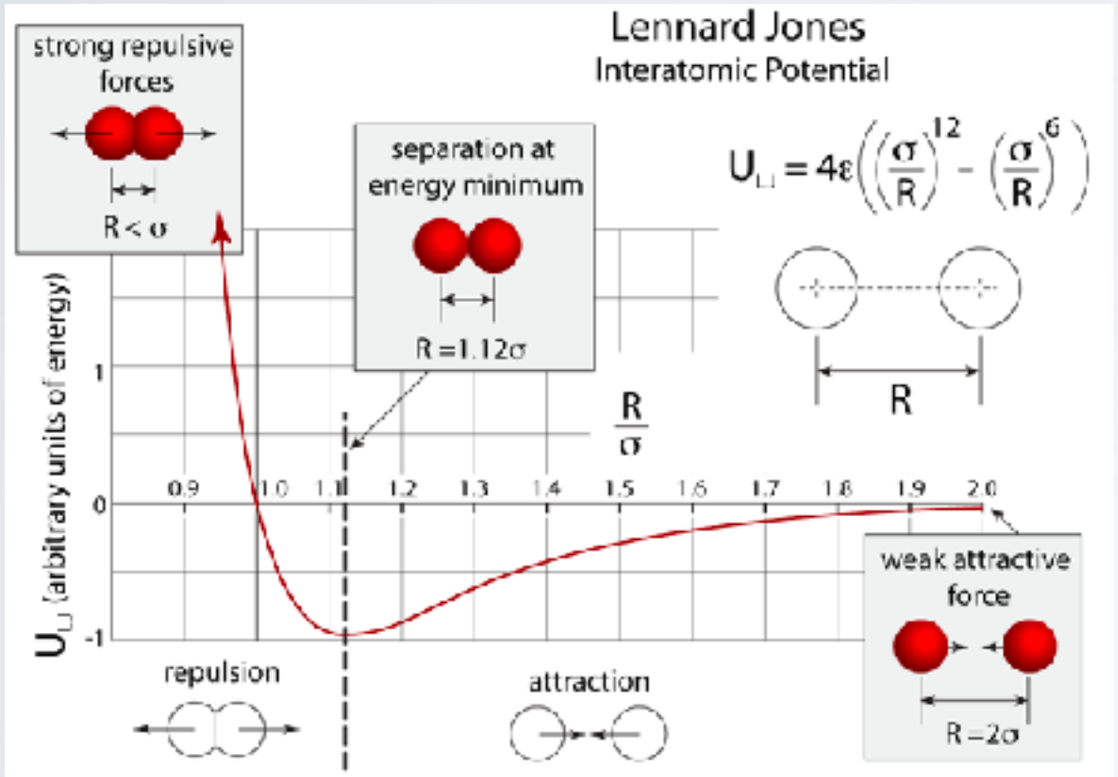
http://www.mbnexplorer.com/users_guide/users_guide743x.png

Hess et al. GROMACS Manual v. 4.5.4 (2010)

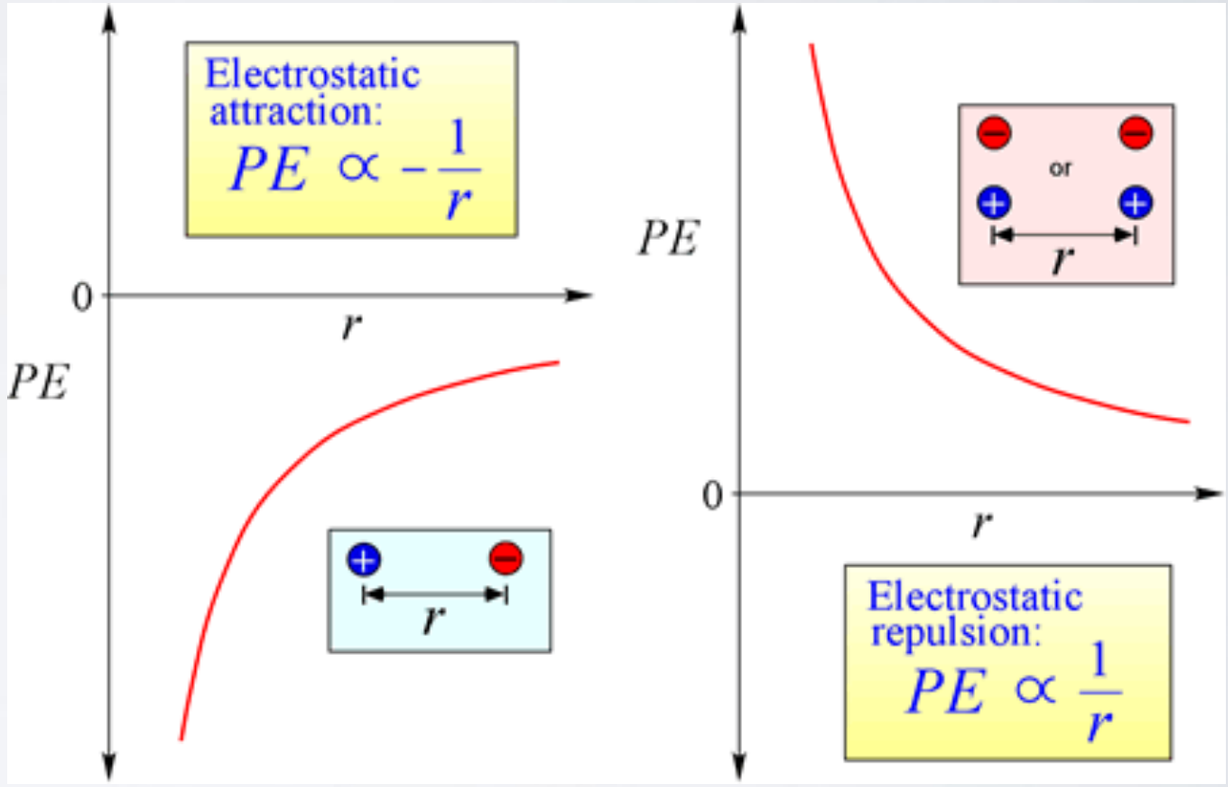
Non-bonded

- Approximate full *n*-body interactions as pairwise additive for simplicity and computational efficiency

- van der Waals



- Coulomb



$$V_{LJ}(r_{ij}) = 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right]$$

$$V_{Coul}(r_{ij}) = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}}$$

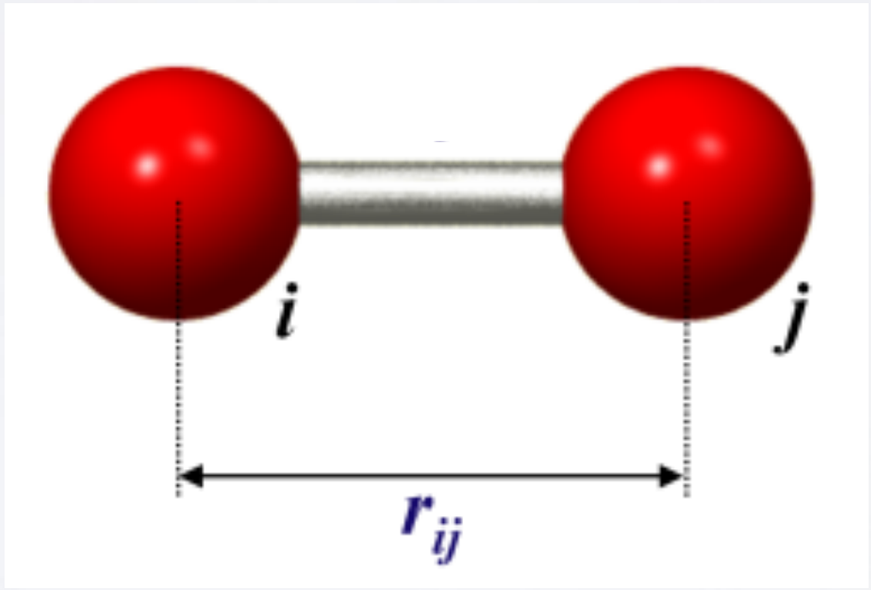
http://atomsinmotion.com/book/chapter5/md

http://guweb2.gonzaga.edu/faculty/cronk/chemistry/images/graph-electrostatic-PE-alt.gif

Restraints

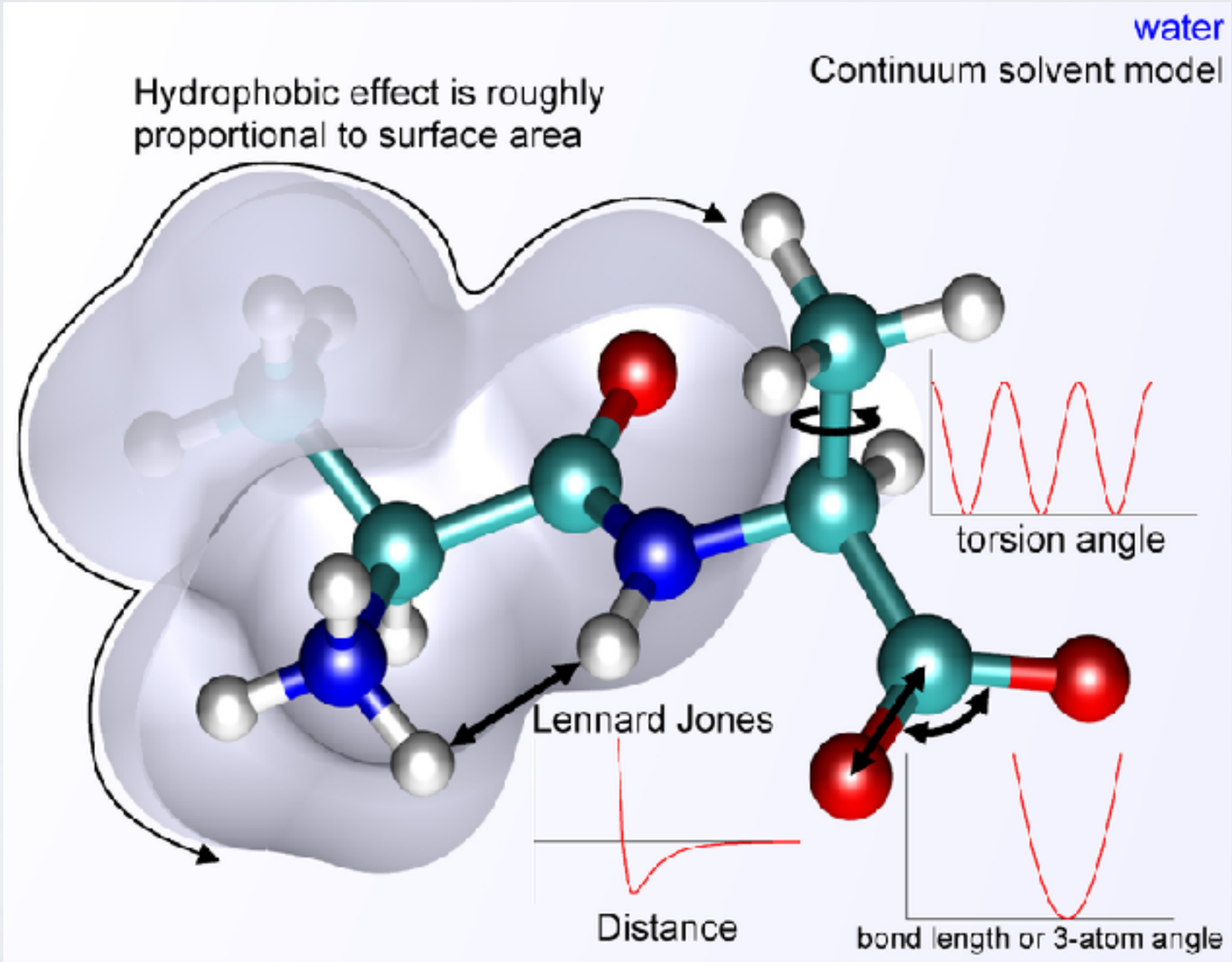
- Restraints can be part of, or supplemental, to a force field

- Many applications, common uses include:
  - fixed bond lengths and angles (esp. for light atoms)
  - artificially immobilize part of the system (e.g., rigid walls or boundary condition)



Fields

- Fields are commonly used to model:
  1. external potentials (e.g., electric field, flow field)
  2. continuum solvation (no explicit solvent molecules)



http://en.wikipedia.org/wiki/File:MM_PEF.png

Ingredient 3: Integrators

- [initial atomic coordinates and velocities] + [force field]
  ⇒ entire future (and past!) modeled by **F=ma**

- Analytical solutions for the dynamical evolution cannot be computed for all but the simplest systems (>2 body)

- Solve Newton's equations by numerical integration
  ⇒ computers ideally suited to rapid, repetitive calculations

- Solving by hand would require thousands of years

Verlet algorithm

- Many possible integration algorithms exist
  (e.g., explicit/implicit Euler, Gear predictor-corrector, n^{th} order Runge-Kutta, Beeman, Newmark-beta)

- The method of choice is the **Verlet algorithm**

  ✓ **fast**
  ✓ **simple**
  ✓ **low-memory**
  ✓ **stable**
  ✓ **time-reversible**
  ✓ **symplectic (phase space volume & E conserving)**

  ✗ **poor accuracy for large time steps (Δt must be small)**

- First recorded use by Delambre in 1791
  Popularized in MD by Loup Verlet in 1967

Verlet algorithm

- Derived from Taylor series:

$$r(t+\delta t) = r(t) + \dot{r}(t)\delta t + \frac{1}{2}\ddot{r}(t)\delta t^2 + ...$$

$$= r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + ...$$

$$r(t-\delta t) = r(t) - \dot{r}(t)\delta t + \frac{1}{2}\ddot{r}(t)\delta t^2 + ...$$

$$= r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + ...$$

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2 + \mathcal{O}\left(\delta t^4\right)$$

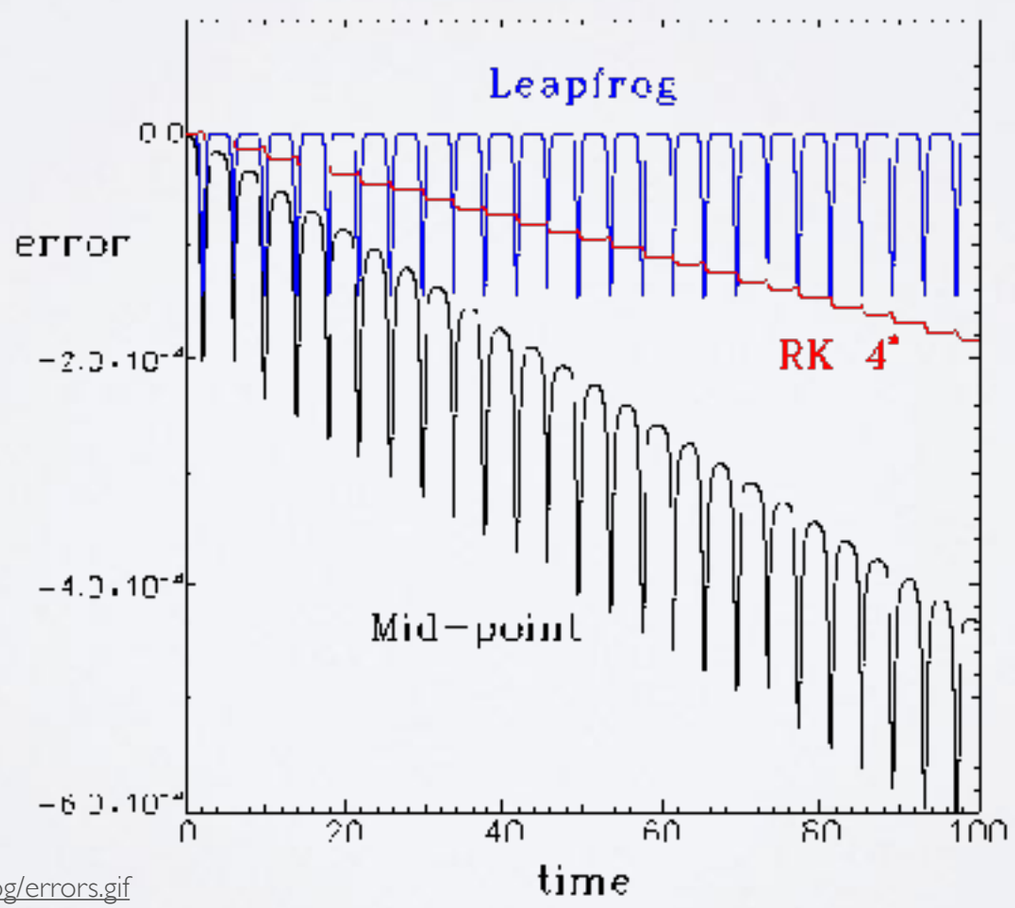
$$v(t) = \frac{r(t+\delta t) - r(t-\delta t)}{2\delta t} + \mathcal{O}\left(\delta t^2\right)$$

$$a_i = \frac{F_i}{m_i}$$

Time-reversibility

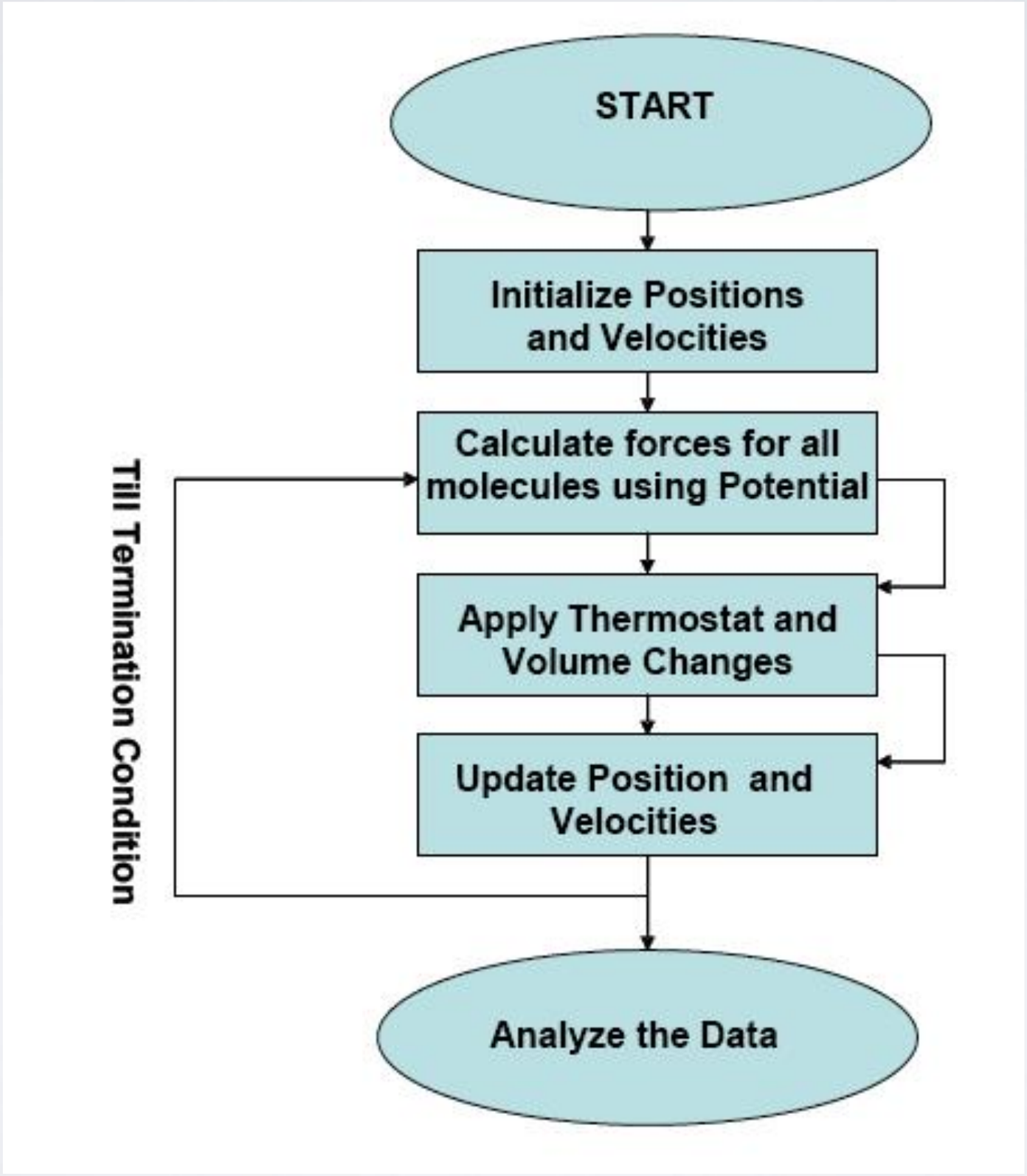
- Higher order integration algorithms have higher per step accuracy, enabling longer time steps and faster simulations
  (e.g., Runge-Kutta, Gear predictor-corrector)

- **But**, do not respect time reversibility of Newton's equations causing energy drift and error accumulation

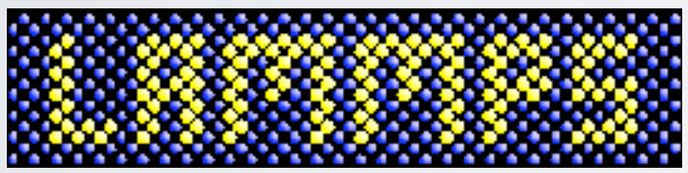


http://einstein.drexel.edu/courses/Comp_Phys/Integrators/leapfrog/errors.gif

Simulation overview



MD software

GROMACS FAST. FLEXIBLE. FREE.	U. Groningen www.gromacs.org	FREE
	Harvard www.charmm.org	$600
AMBER	Rutgers *et al.* www.ambermd.org	$400
NAMD Scalable Molecular Dynamics	UIUC www.ks.uiuc.edu	FREE
Desmond D E Shaw Research	D.E. Shaw Research www.deshawresearch.com	FREE
	Sandia National Lab http://lammps.sandia.gov	FREE
HOOMD-blue	U. Michigan http://codeblue.umich.edu/hoomd-blue/	FREE
Folding@home distributed computing	Folding@home http://folding.stanford.edu	FREE
OpenMM	OpenMM http://openmm.org	FREE

1. Classical molecular dynamics in 15 minutes

2. ANN accelerated sampling of molecular free energy landscapes **[ML-driven search of conformational space]**

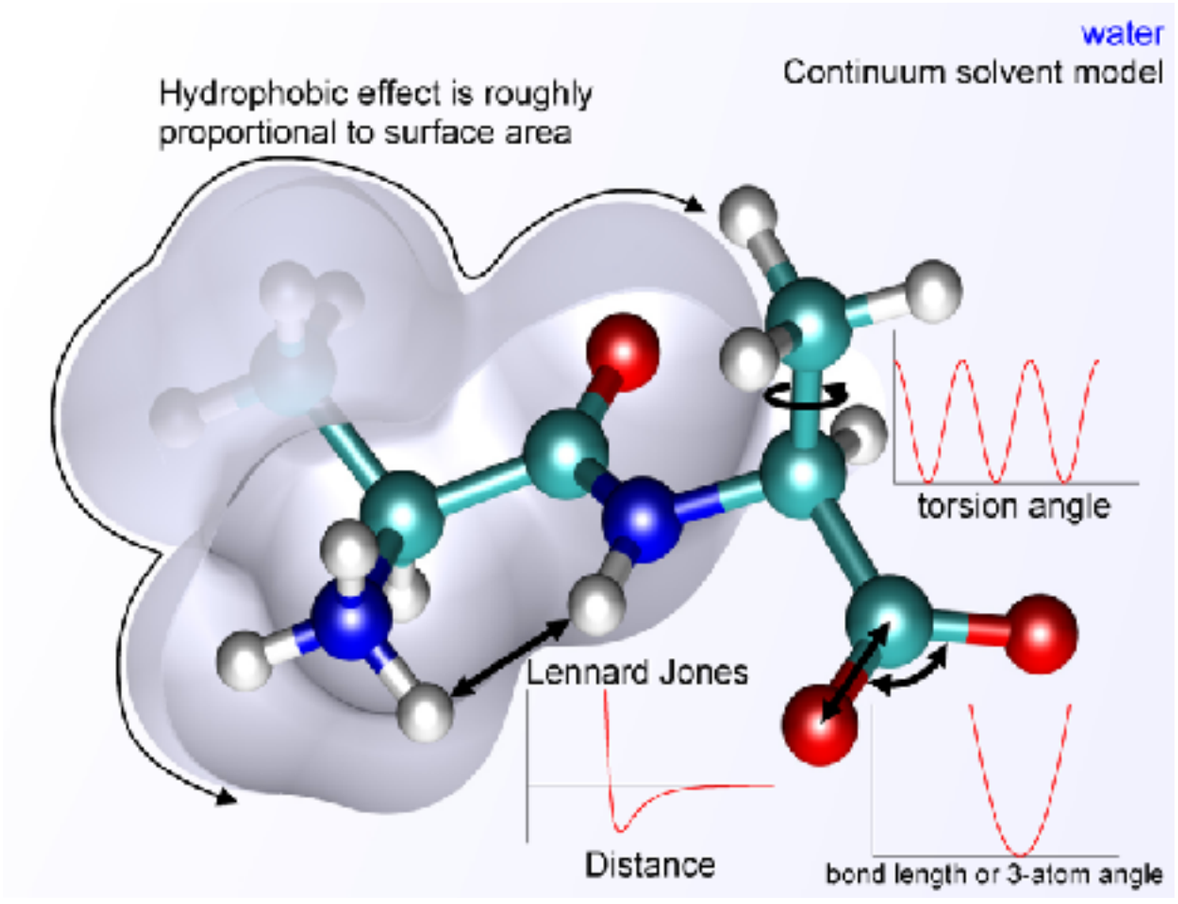
3. Data-driven design of π-conjugated oligopeptides **[DS-driven search of chemical space]**

Limitations of molecular simulation

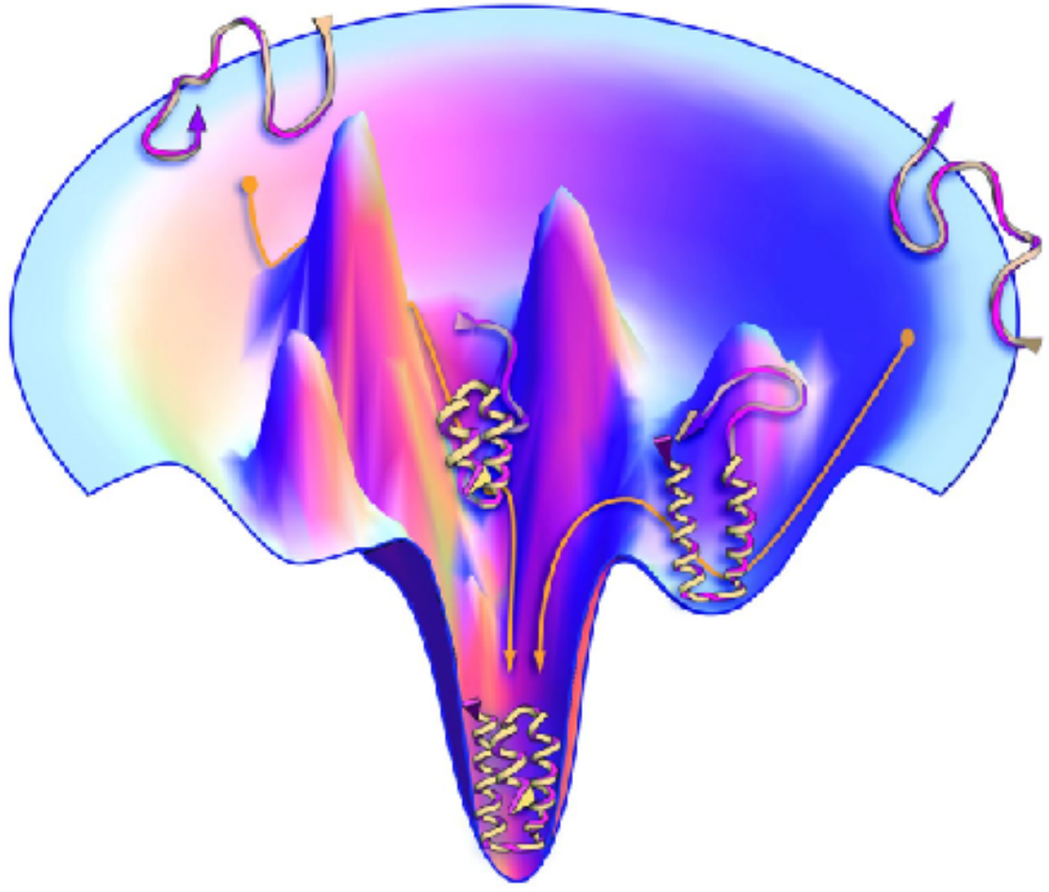
"Two limitations in existing simulations are the approximations in the potential energy functions and the lengths of the simulations. The first introduces systematic errors and the second statistical errors."

— M. Karplus & G.A. Petsko *Nature* (1990)

1. Accurate force fields



2. Sampling configurational space



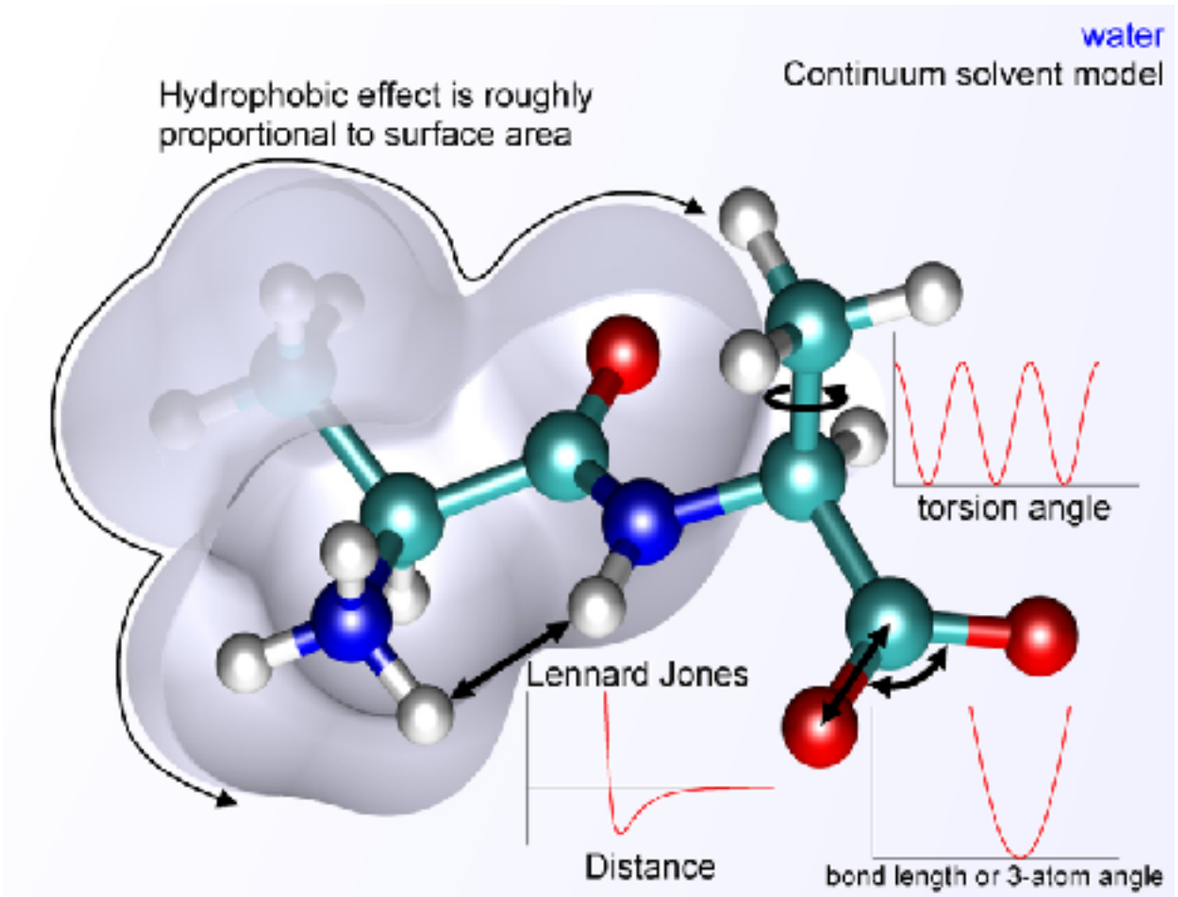
M. Karplus and G.A. Petsko Nature 347 631-639 (1990)
https://upload.wikimedia.org/wikipedia/commons/5/5c/MM_PEF.png http://jonlieffmd.com/wp-content/uploads/2012/12/F3.large_.jpg

Limitations of molecular simulation

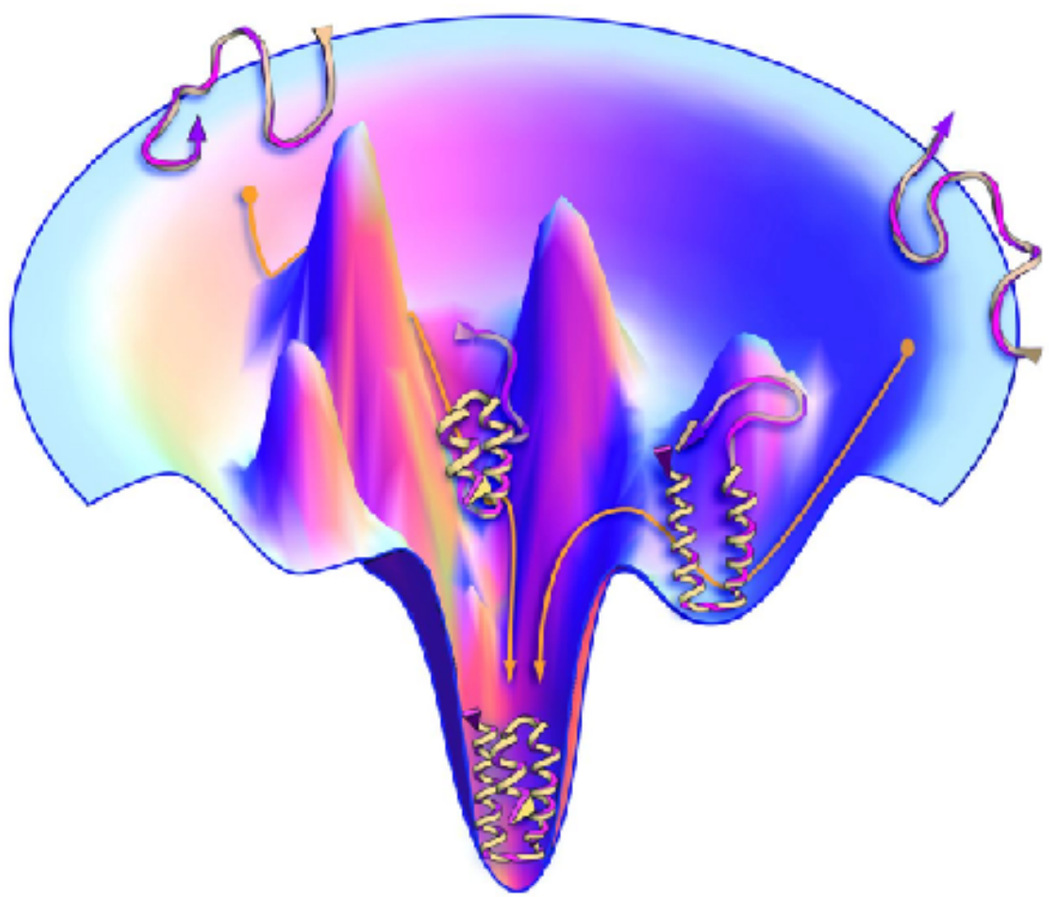
"Two limitations in existing simulations are the approximations in the potential energy functions and the lengths of the simulations. The first introduces systematic errors and the second statistical errors."

— M. Karplus & G.A. Petsko *Nature* (1990)

1. Accurate force fields



2. Sampling configurational space



M. Karplus and G.A. Petsko Nature 347 631-639 (1990)
https://upload.wikimedia.org/wikipedia/commons/5/5c/MM_PEF.png http://jonlieffmd.com/wp-content/uploads/2012/12/F3.large_.jpg

Accelerated sampling

- Accelerated sampling techniques partition largely into two classes:

Tempering techniques	**Collective variable biasing**
Simulated annealing	Umbrella sampling
Multicanonical algorithm	Hyperdynamics
Replica exchange	Metadynamics
Hamiltonian exchange	Adiabatic free energy dynamics (AFED)
Parallel tempering	Temperature accelerated dynamics (TAD)
…	Temperature accelerated MD (TAMD)
	Adaptive force biasing
	…

- **Tempering** modifies T or Hamiltonian to accelerate barrier crossing
  ➔ substantial CPU time expended on conditions not of direct interest

- **CV biasing** efficiently directs sampling along relevant order parameters
  ➔ presupposes *a priori* availability of "good" CVs

C. Abrams and G. Bussi Entropy 16 163–199 (2013) M.A. Rohrdanz, W. Zheng, and C. Clementi, C. Annu. Rev. Phys. Chem. 64 295–316 (2013)

Accelerated sampling

- Accelerated sampling techniques partition largely into two classes:

Tempering techniques

Simulated annealing

Multicanonical algorithm

Replica exchange

Hamiltonian exchange

Parallel tempering

…

Collective variable biasing

Umbrella sampling

Hyperdynamics

Metadynamics

Adiabatic free energy dynamics (AFED)

Temperature accelerated dynamics (TAD)

Temperature accelerated MD (TAMD)

Adaptive force biasing

…

- **Tempering** modifies T or Hamiltonian to accelerate barrier crossing
  ➔ substantial CPU time expended on conditions not of direct interest

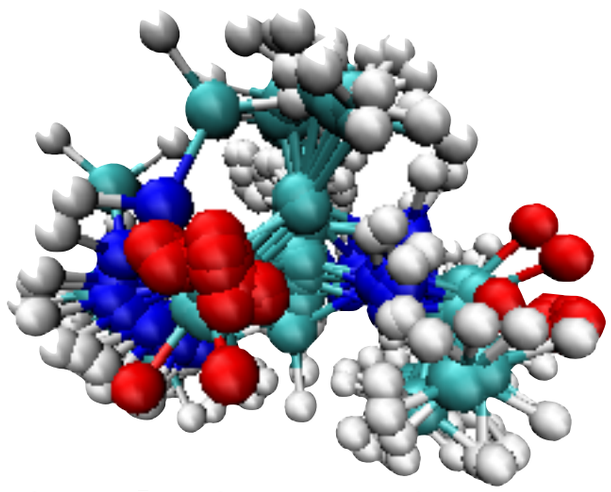
- **CV biasing** efficiently directs sampling along relevant order parameters
  ➔ presupposes *a priori* availability of "good" CVs

Automated CV discovery

- Given a simulation trajectory **data mining / dimensionality reduction** can discover "good" CVs that:

  (1) Separate metastable system states
  (2) Characterize important large-scale or slow conformational motions
  (3) Are explicit differentiable functions of atomic coordinates

- (3) is required to propagate CV biases to atomic forces $f_i^{\text{tot}} = f_i^U + f_i^B$



simulation trajectory

linear dim red
(e.g., PCA, MDS)

✗ Many required to separate states
✗ Poor descriptors of molecular motion
✔ Explicit function of atomic coords

nonlinear dim red
(e.g., LLE, Isomap, dMaps)

✔ Parsimonious state separation
✔ Coincident with large-scale motions
✗ Unknown mapping to atomic coords

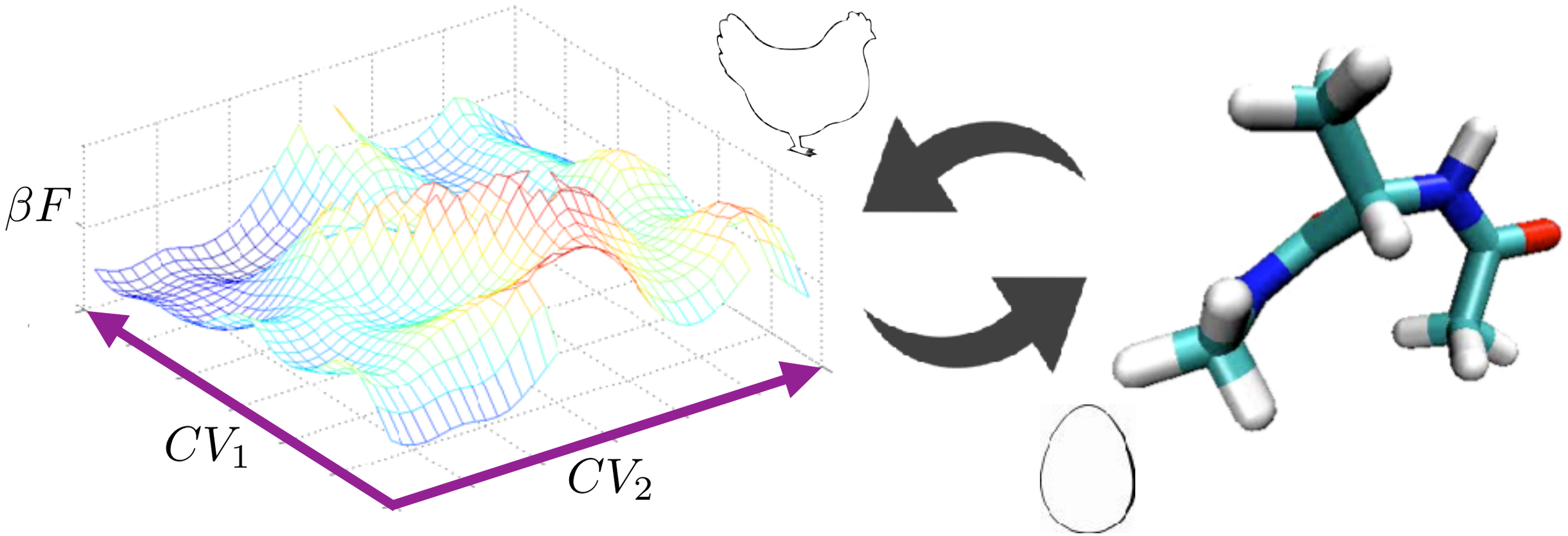
CV biasing: The chicken and the egg

"[n]o method can presently extract reaction coordinates on the fly during MD simulations and at the same time use them to enhance the sampling of the configurational space"

— M.A. Rohrdanz, W. Zheng, and C. Clementi *Annu. Rev. Phys. Chem.* (2013)

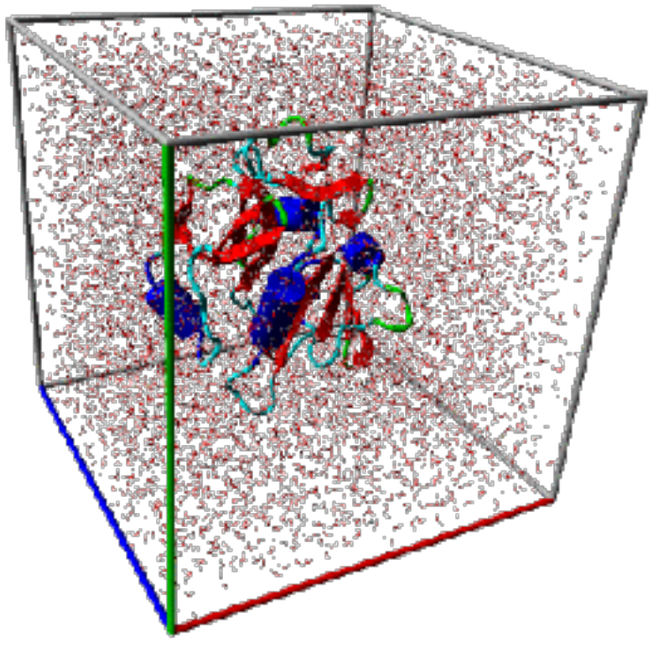
Good CVs required to drive sampling of configurational space (chicken)

Trajectories with good sampling needed to discover good CVs (egg)

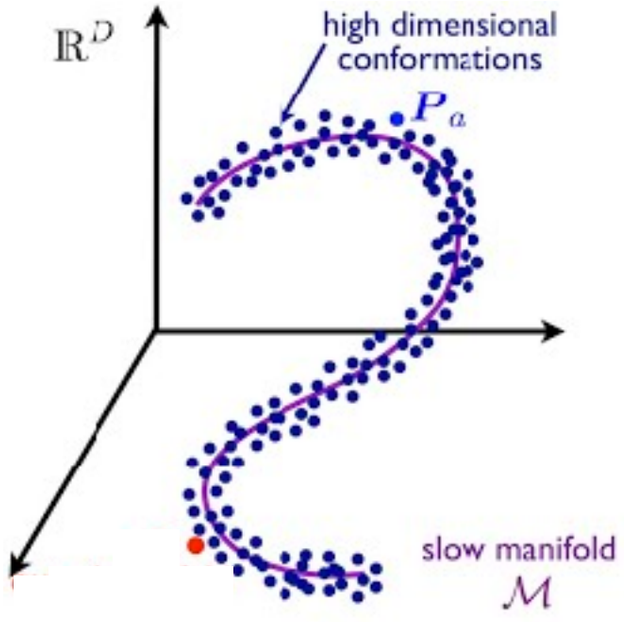


M.A. Rohrdanz, W. Zheng, and C. Clementi, C. Annu. Rev. Phys. Chem. 64 295–316 (2013)

Interleaved CV discovery and biased simulation



biased simulations in current CVs to expand exploration



nonlinear learning to update CV estimate

- **Biasing step frustrated by absence of CV mapping to atomic coords**

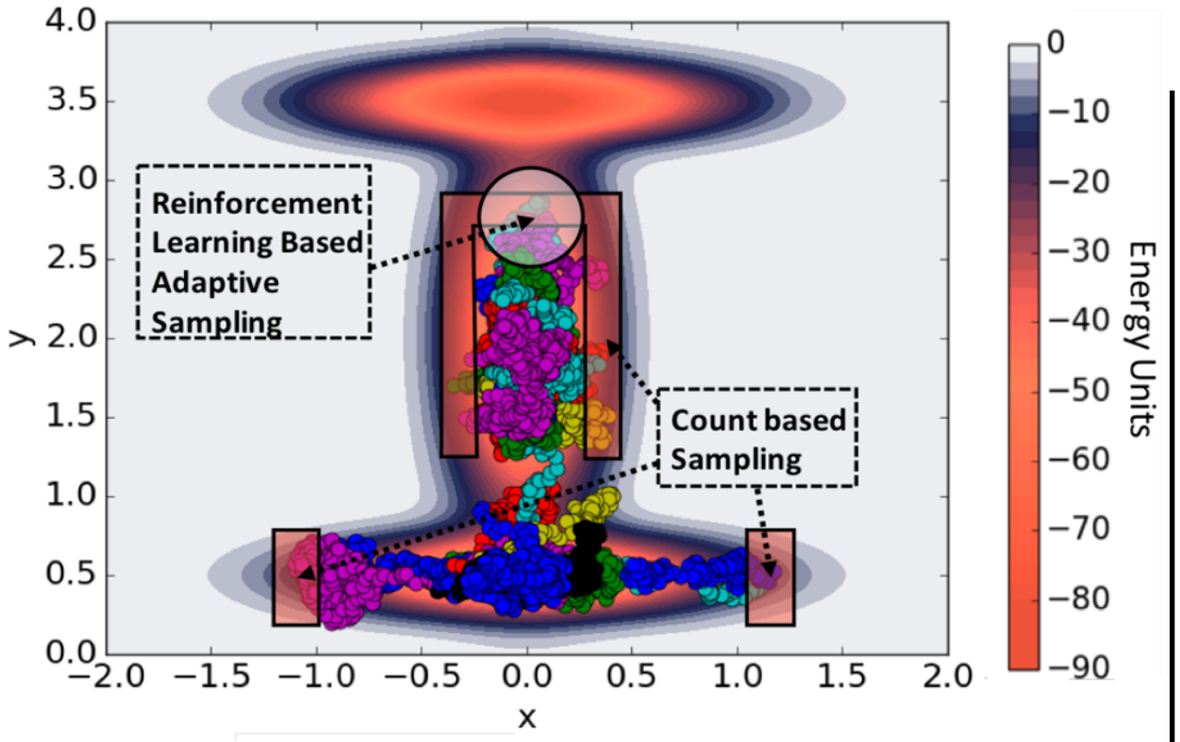
- Approximate solutions:
  - (i) correlate data-mined CVs with physical variables in which to do biasing [1]
  - (ii) select from (linear combinations of) known CVs [2]
  - (iii) use CVs not for biasing but smart initialization of new runs [3]
  - (iv) approximate CVs with functional fit or by "landmarks" in CV embedding [4]

http://openwetware.org/images/b/bc/UANLSimulation2.jpg https://sites.google.com/site/rdanielmillan/_/rsrc/1359727507229/publications/phd_thesis/slow_manifold_lme.jpg

1. ALF, A.Z. Panagiotopoulos, P.G. Debenedetti, and I.G. Kevrekidis J. Chem. Phys. 134 135103 (2011)
2. Z. Shamsi, K.J. Cheng, and D. Shukla J. Phys. Chem. B 122 35 8386–8395 (2018) A. Ma and A. Dinner J. Phys. Chem. B 109} 14 6769–6779 (2005) J. Marcelo, L. Ribeiro, P. Bravo, Y. Wang, and P. Tiwary J. Chem. Phys. 149, 072301 (2018)
3. J. Preto and C. Clementi Phys. Chem. Chem. Phys. 16, 19181–19191 (2014); W. Zheng, M.A. Rohrdanz, and C. Clementi J. Phys. Chem. B. 117 12769–12776 (2013) E. Chiavazzo, R. Covino, R.R. Coifman, C.W. Gear, A.S. Georgiou, G. Hummer, and I.G. Kevrekidis PNAS 114 28 E5494–E5503 (2017)
4. B. Hashemian, D. Millán, and M. Arroyo J. Chem. Phys. 139 214101 (2013); D. Branduardi, F.L. Gervasio, and M. Parrinello J. Chem. Phys. 126 054103 (2007); C.F. Abrams and E. Vanden-Eijnden, E. Chem. Phys. Lett. 547 114–119 (2012)

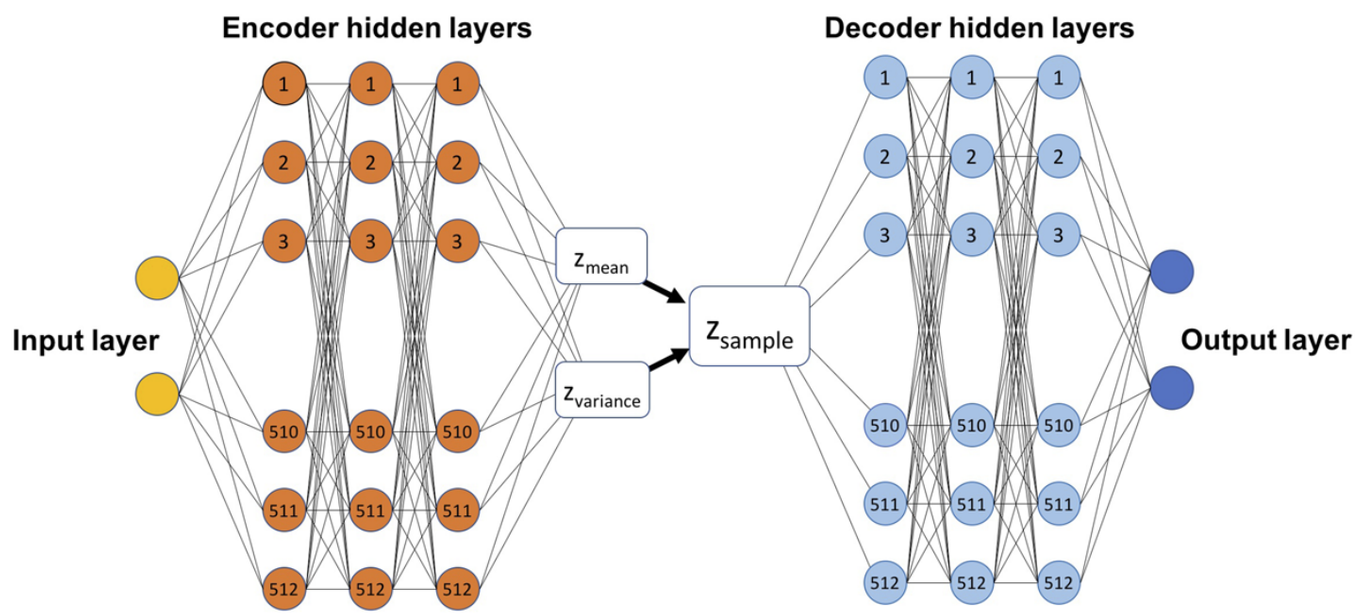
Selection among known CVs

REAP (Shukla et al.)



$$r^K(c_m) = \sum_{i=1}^{k} w_i^S \frac{|(\theta_i(c_m) - \langle\theta_i(C)\rangle|}{\sigma_i(C)}$$

RAVE (Tiwary et al.)

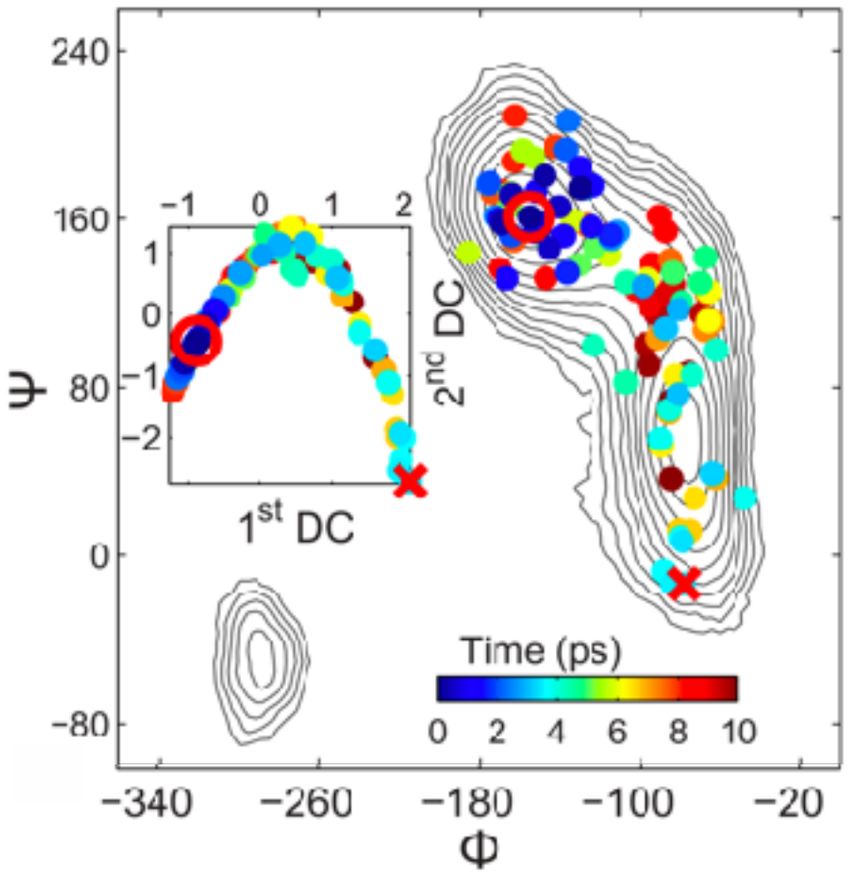


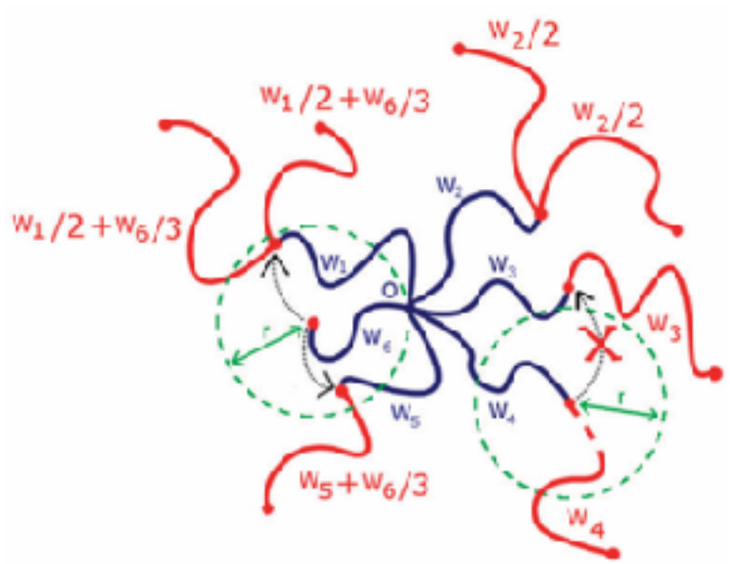
$$\mathcal{D}_{KL}(P(z)||P(\chi)) = \sum_i P^u(z_i) \log \frac{P^u(z_i)}{P^u(\chi_i)}$$

Z. Shamsi, K.J. Cheng, and D. Shukla, Diwakar J. Phys. Chem. B 122 35 8386–8395 (2018)
J. Marcelo, L. Ribeiro, P. Bravo, Y. Wang, and P. Tiwary J. Chem. Phys. 149, 072301 (2018)

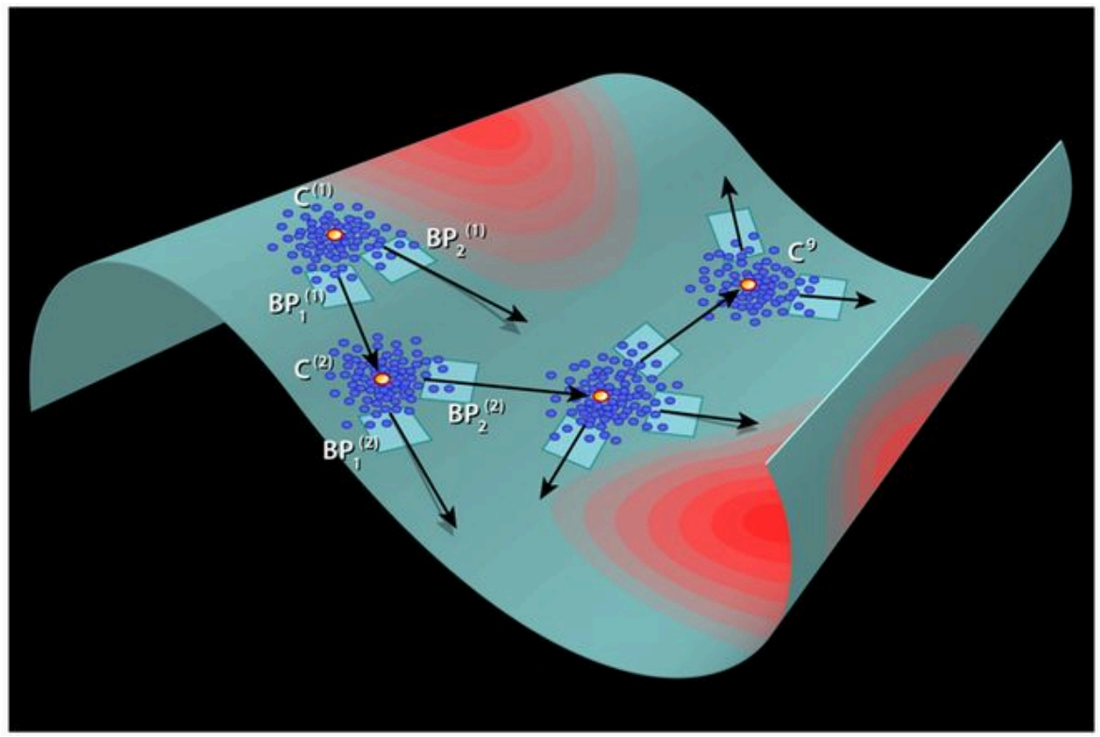
Nonlinear dim red + smart initialization

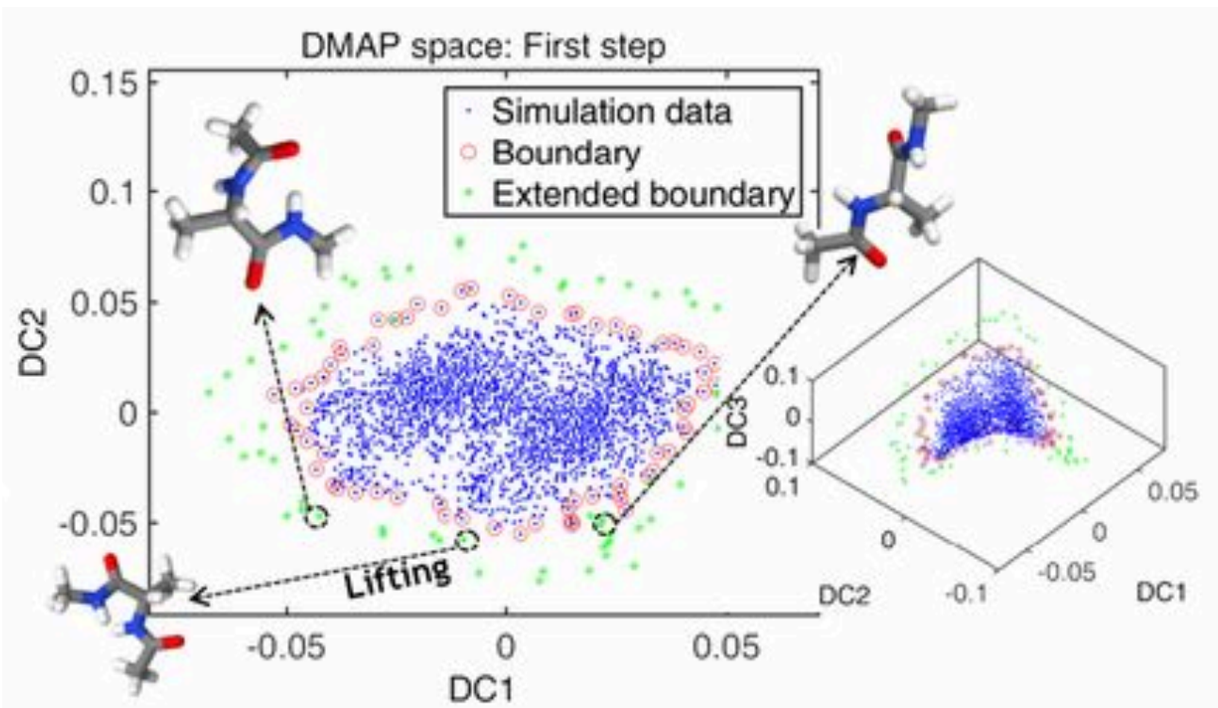
DM-d-MD (Clementi et al.)





iMapD (Kevrekidis et al.)

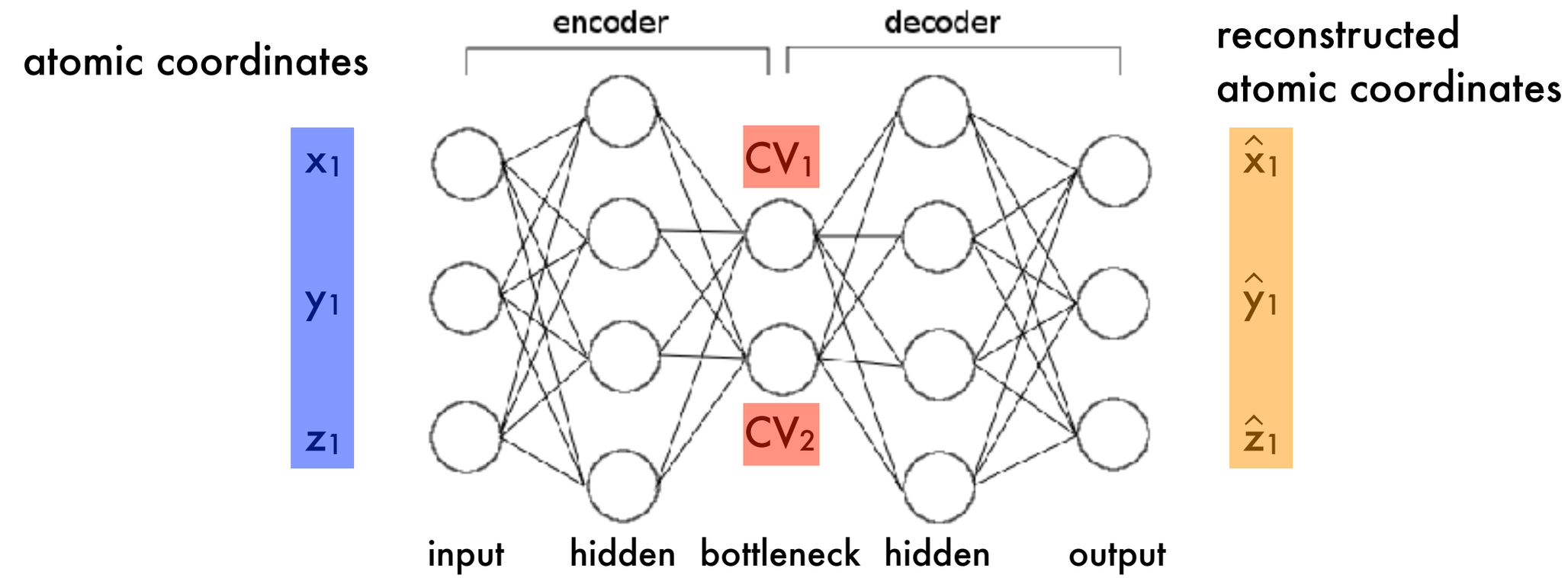




J. Preto and C. Clementi Phys. Chem. Chem. Phys. 16, 19181–19191 (2014); W. Zheng, M.A. Rohrdanz, and C. Clementi J. Phys. Chem. B. 117 12769–12776 (2013)
E. Chiavazzo, R. Covino, R.R. Coifman, C.W. Gear, A.S. Georgiou, G. Hummer, and I.G. Kevrekidis PNAS 114 28 E5494–E5503 (2017)

Auto-associative neural networks (autoencoders)

- **Autoencoders** unique among unsupervised nonlinear dimensionality reduction tools in furnishing **explicit** and **differentiable** latent space map



Neuron activation function

$$y_k^{(i)} = f^{(i)}\left(x_k^{(i)}\right)$$

output of node k in layer i

activation function (e.g., tanh)

input to node k in layer i

Weighted sums between layers

$$x_k^{(i)} = b_k^{(i)} + \sum_j w_{jk}^{(i-1)} y_j^{(i-1)}$$

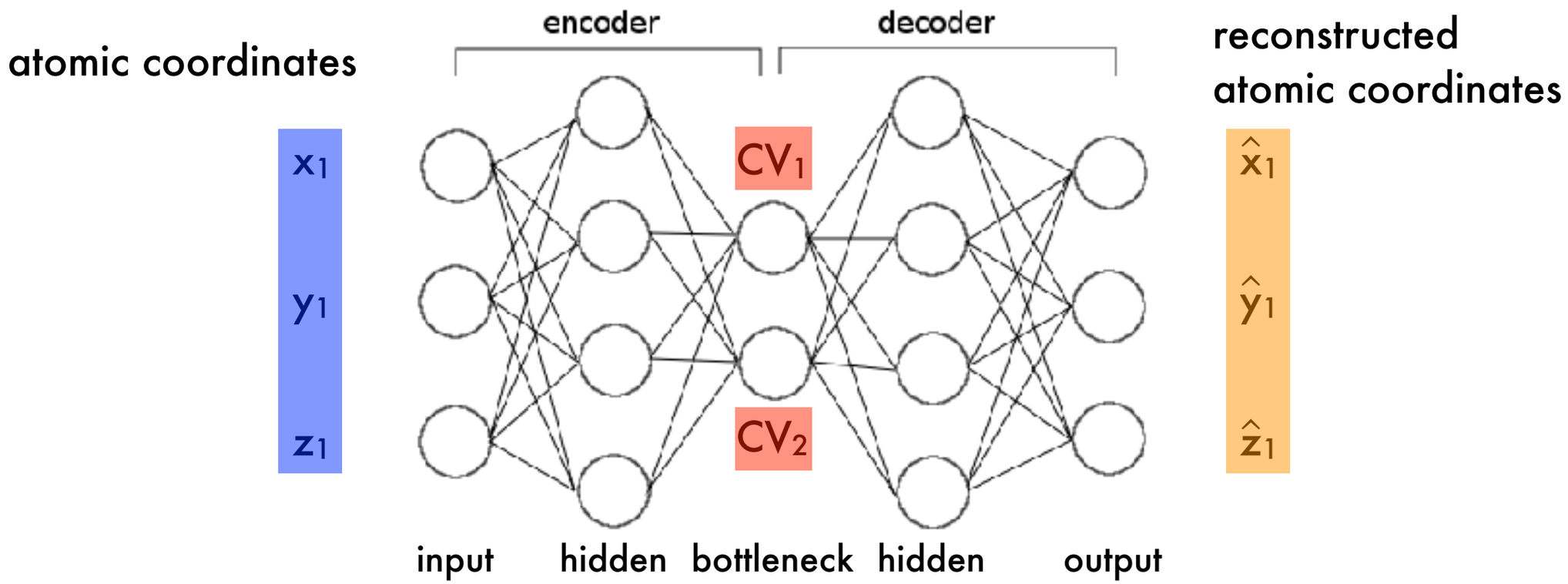
bias to node k in layer i

weight from node j in layer (i-1) and node k in layer i

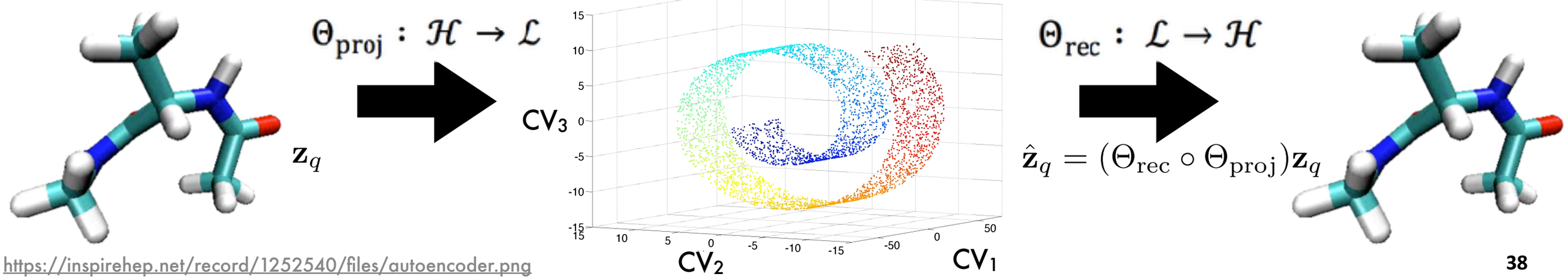
https://inspirehep.net/record/1252540/files/autoencoder.png

Auto-associative neural networks (autoencoders)

- **Autoencoders** unique among unsupervised nonlinear dimensionality reduction tools in furnishing **explicit** and **differentiable** latent space map



- Idea is to discover and parameterize with CVs a low-dim manifold from which atomic coordinates can be approximately reconstructed



https://inspirehep.net/record/1252540/files/autoencoder.png

Implementing the bias

- Generically apply bias through artificial potential in CVs

$$P(\mathbf{r}^N) = \frac{\mathrm{e}^{-\beta E(\mathbf{r}^N)}}{Z} = \frac{\mathrm{e}^{-\beta[H(\mathbf{r}^N)+V(\overrightarrow{CV}(\mathbf{r}^N))]}}{Z}$$

where CVs are explicit and differentiable functions of atomic coords

$$CV_i = CV_i(\mathbf{r}^N) \longleftarrow$$

ugly, but explicit, function of input atomic coords and autoencoder weights, biases, and activation functions

Implementing the bias

- Generically apply bias through artificial potential in CVs

$$P(\mathbf{r}^N) = \frac{\mathrm{e}^{-\beta E(\mathbf{r}^N)}}{Z} = \frac{\mathrm{e}^{-\beta[H(\mathbf{r}^N)+V(\overrightarrow{CV}(\mathbf{r}^N))]}}{Z}$$

where CVs are explicit and differentiable functions of atomic coords

$$CV_i = CV_i(\mathbf{r}^N)$$ ← ugly, but explicit, function of input atomic coords and autoencoder weights, biases, and activation functions

- Perform biased MD by analytically propagating CV bias into atomic forces

Energy $E(\mathbf{r}^N) = H(\mathbf{r}^N) + V(\overrightarrow{CV}(\mathbf{r}^N))$

total potential

unbiased Hamiltonian

biasing potential (e.g., harmonic umbrellas)

Force $f_i(\mathbf{r}^N) = \underbrace{-\nabla_{\mathbf{r_i}} H(\mathbf{r}^N)}_{f_i^U} \underbrace{- \frac{\partial V}{\partial \overrightarrow{CV}} \nabla_{\mathbf{r_i}} \overrightarrow{CV}(\mathbf{r}^N)}_{f_i^B}$

Computational implementation

- Autoencoders permit biased simulation directly in the discovered CVs

- **Interleaved on-the-fly learning and biasing:**
  Online biasing implemented in OpenMM as custom force plugin
  Offline autoencoder training over trajectory using Pytorch Python libraries

`openmm.org`

`AEForce` plugin to OpenMM molecular dynamics package

`pytorch.org`

autoencoder training over simulation trajectory using Pytorch Python libraries

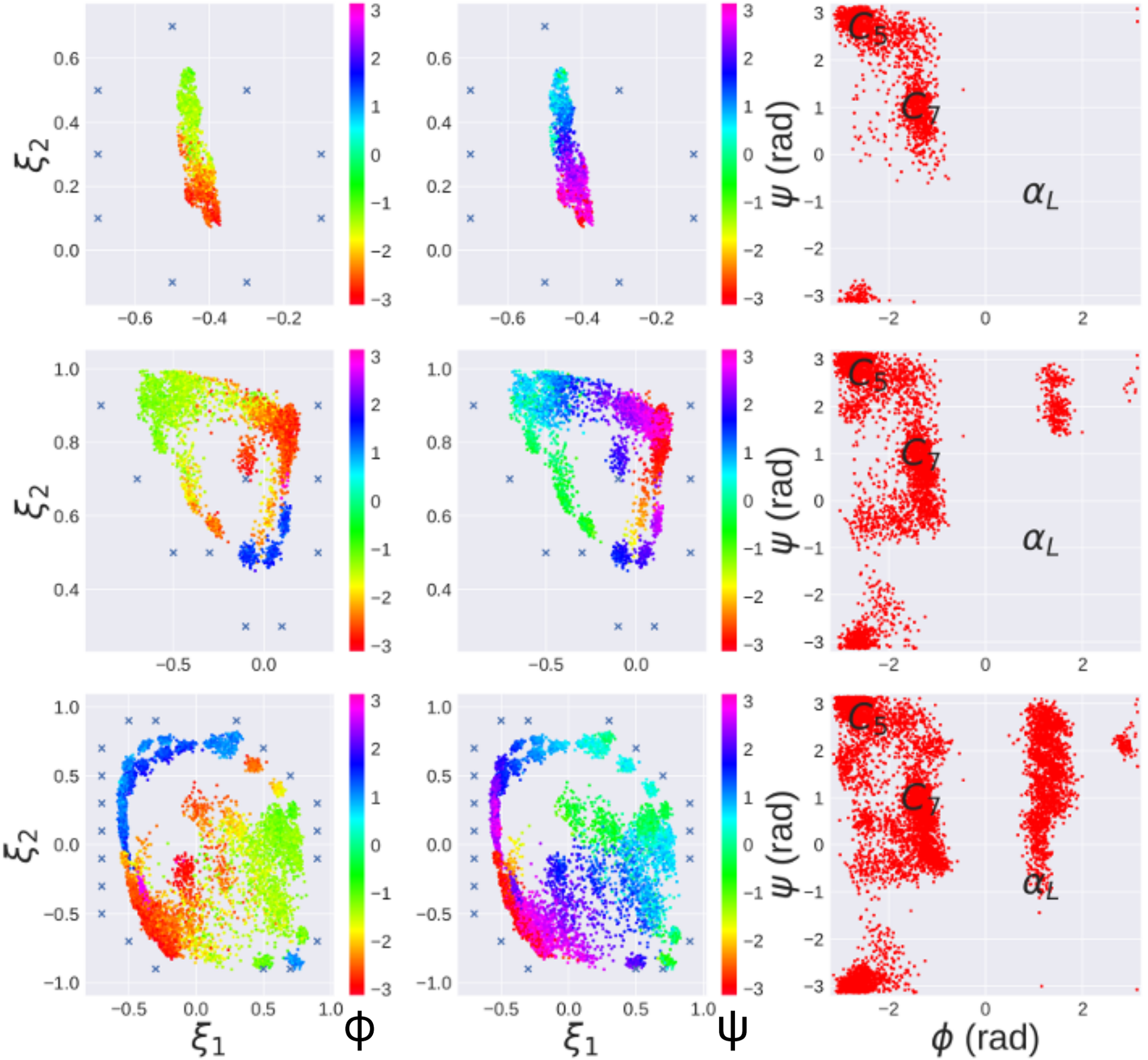
W. Chen and ALF J. Comput. Chem. J. Comput. Chem. 39 25 2079-2102 (2018) W. Chen, A.-R. Tan, and ALF J. Chem. Phys. 149 072312 (2018)
W. Chen, H. Sidky, and ALF J. Chem. Phys. 150 214114 (2019) ∘ Selected as J. Chem. Phys. "Editor's Pick"

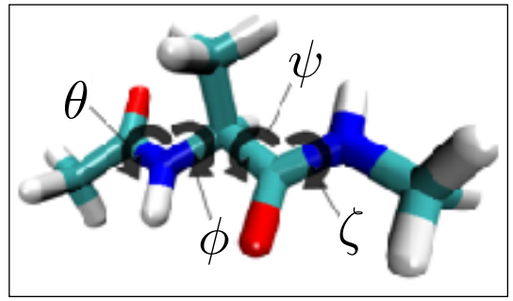
Alanine dipeptide in vacuum (Amber99sb)

ITER 1

ITER 3

ITER 5



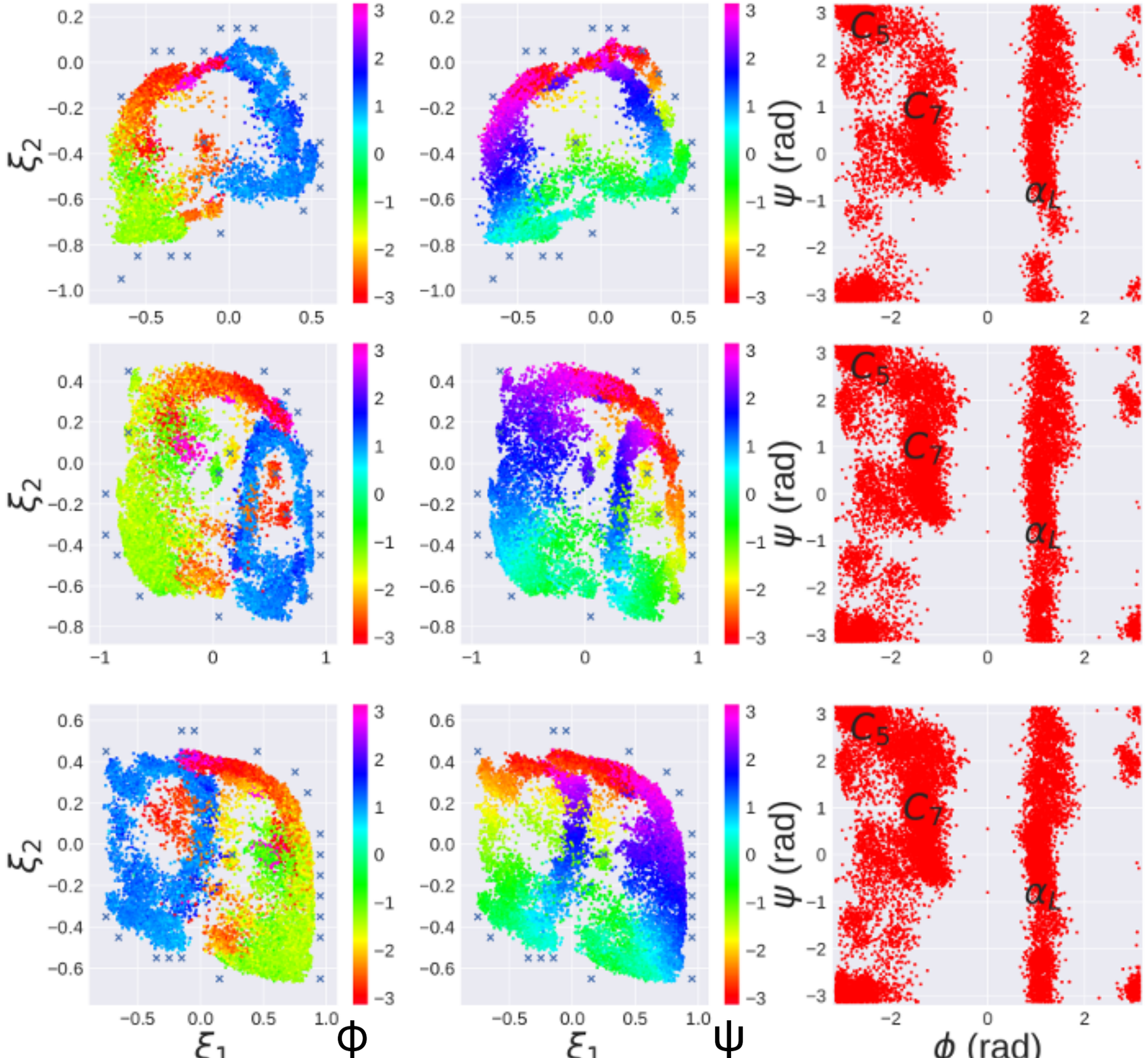


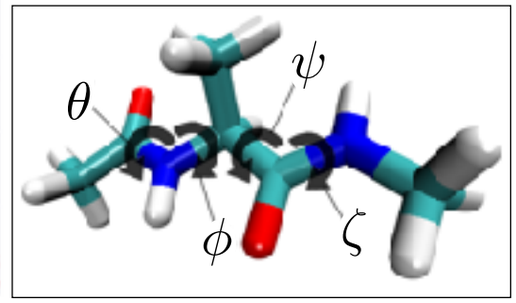
Alanine dipeptide in vacuum (Amber99sb)

ITER 7

ITER 9

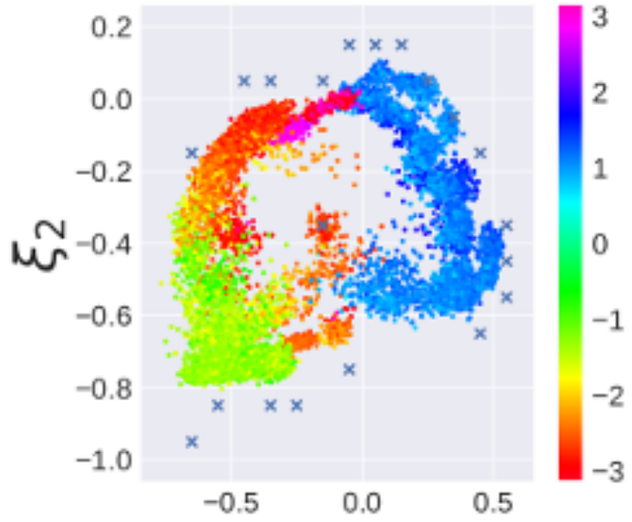
ITER 10

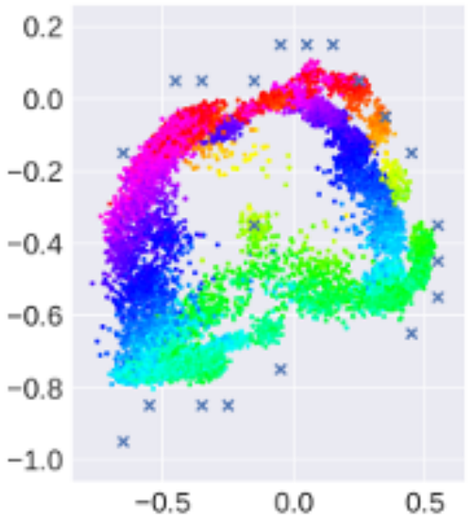


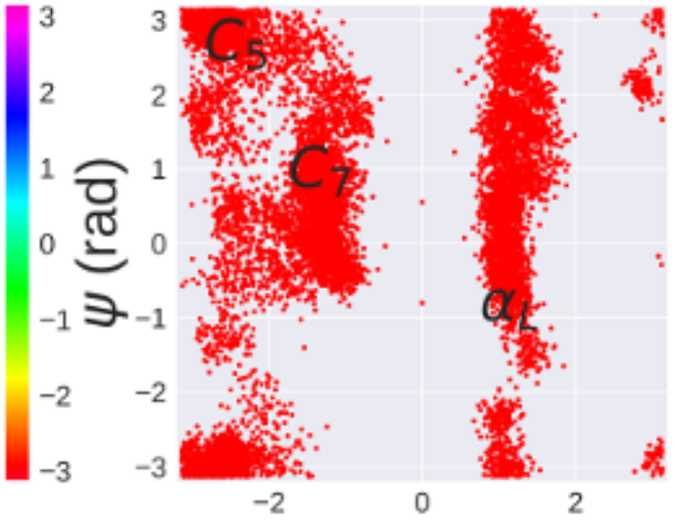


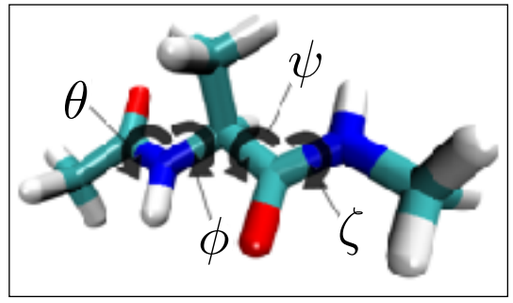
Alanine dipeptide in vacuum (Amber99sb)

ITER 7

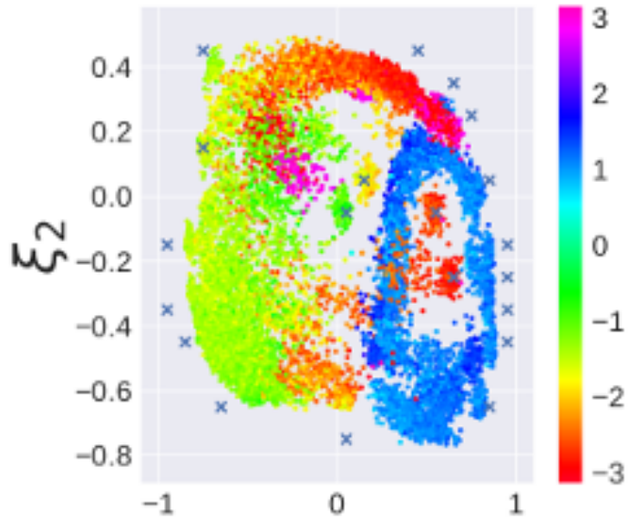


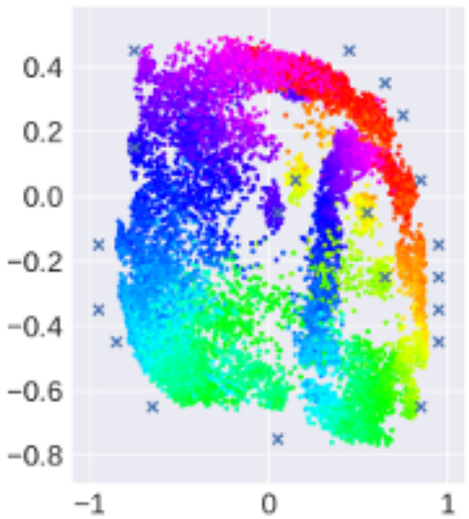


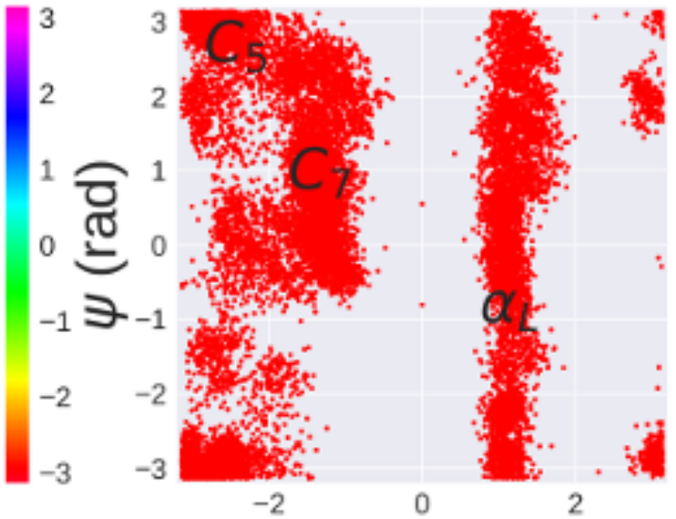




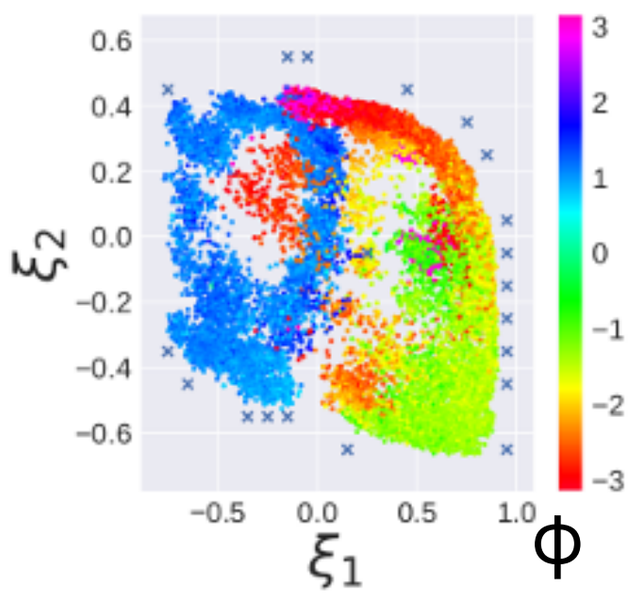
ITER 9

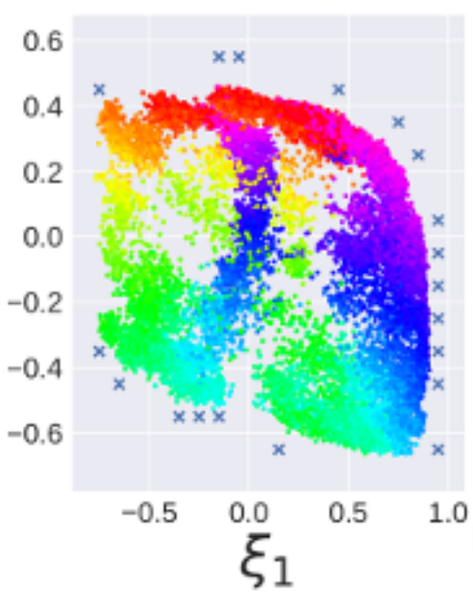


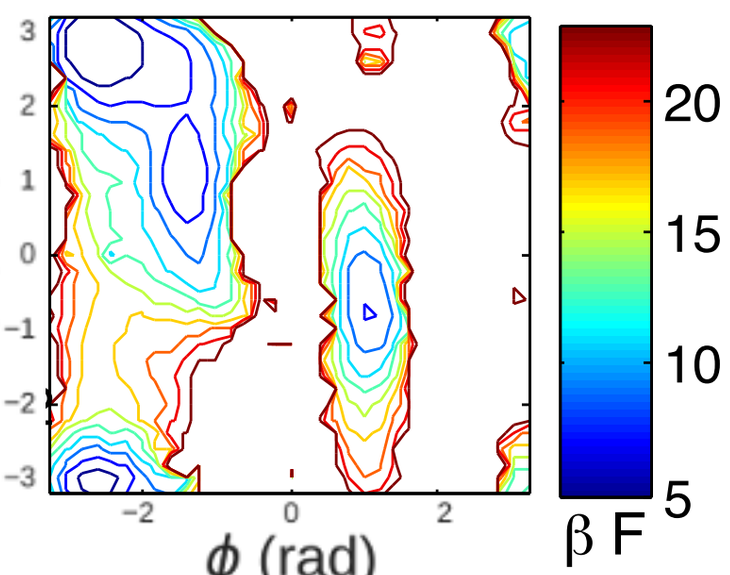




ITER 10





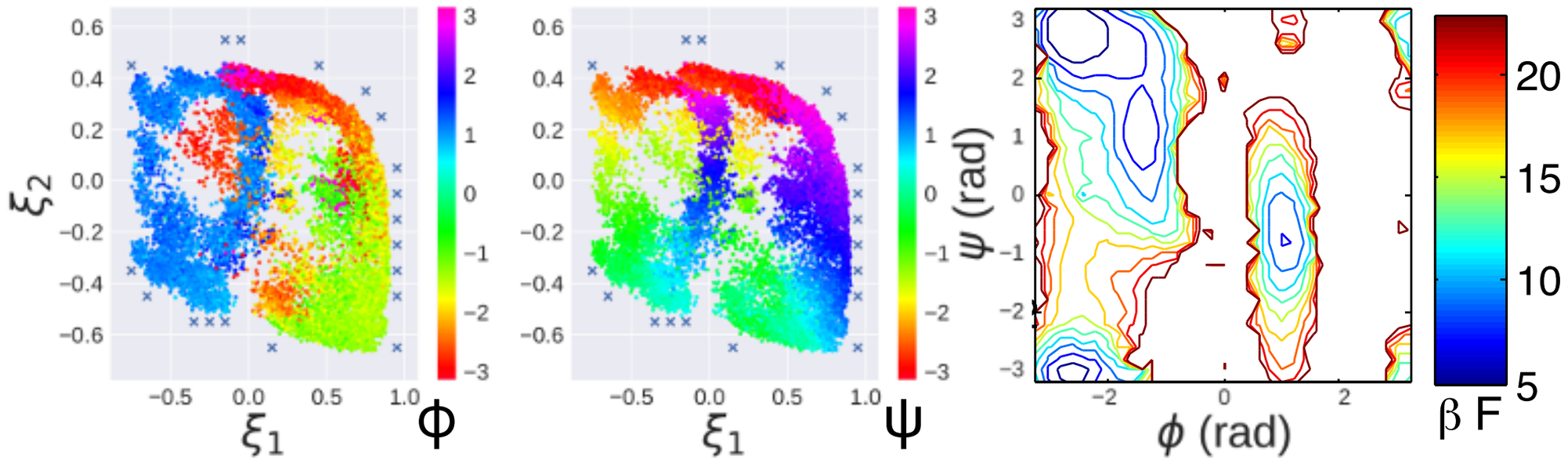


Alanine dipeptide in vacuum (Amber99sb)

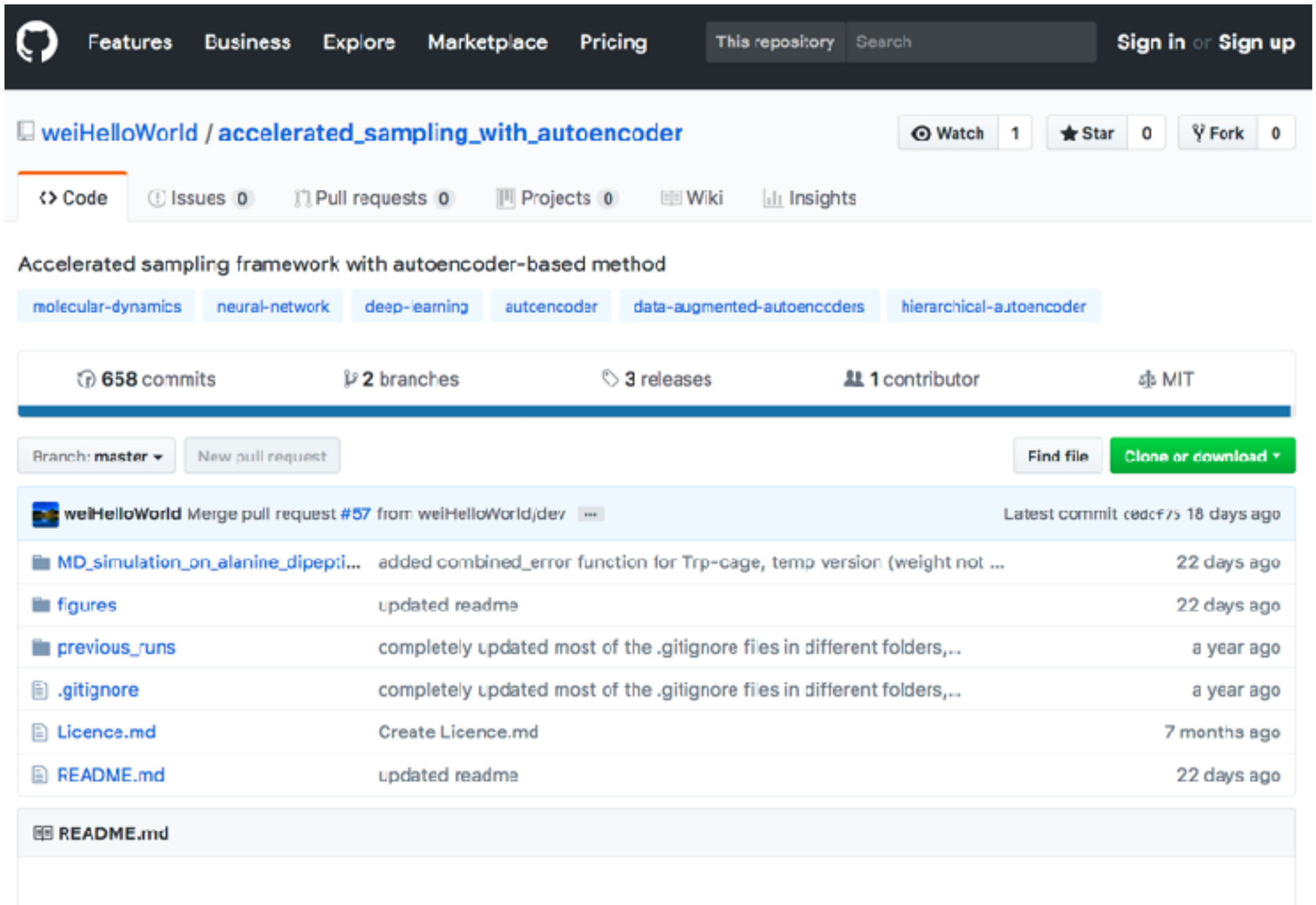
- MESA **converges within 10 iterations** to quantitatively accurate FES
- Autoencoder discovers correct 4D flat torus topology with two periodic collective variables {Φ,Ψ}
- Timings on single Intel i7-5820K CPU core:

10 × training 21-40-2-40-21 networks w/ Q=1500 & N=16	**1200 s**
1 × 800 ps unbiased simulation	**12 s**
75 × 10 ps biased simulations	**130 s**
	22 CPU-mins

ITER 10



Open-source availability



https://github.com/weiHelloWorld/accelerated_sampling_with_autoencoder
https://github.com/weiHelloWorld/ANN_Force

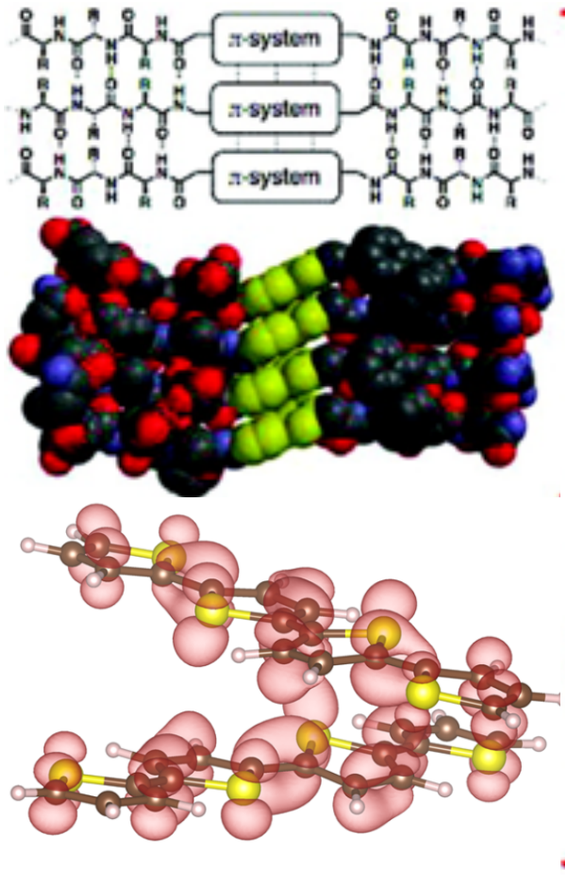
1. Classical molecular dynamics in 15 minutes

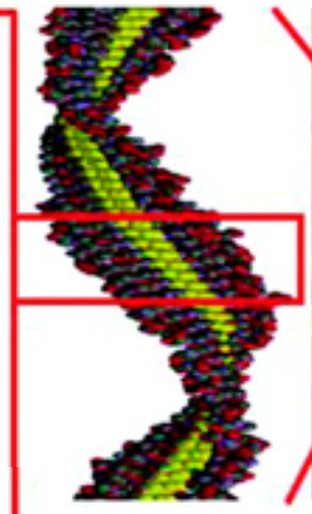
2. ANN accelerated sampling of molecular free energy landscapes [ML-driven search of conformational space]

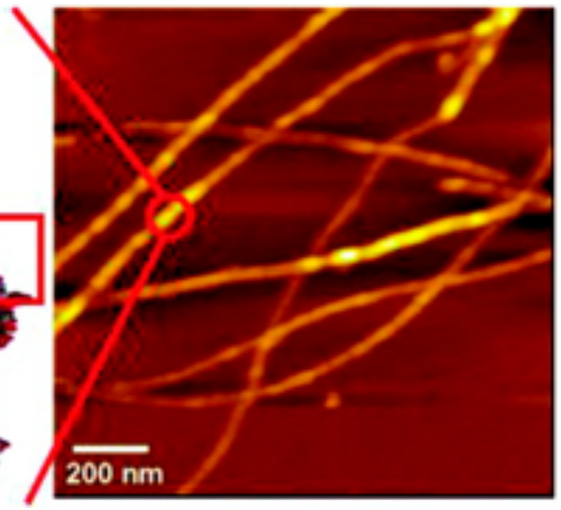
3. Data-driven design of π-conjugated oligopeptides [DS-driven search of chemical space]

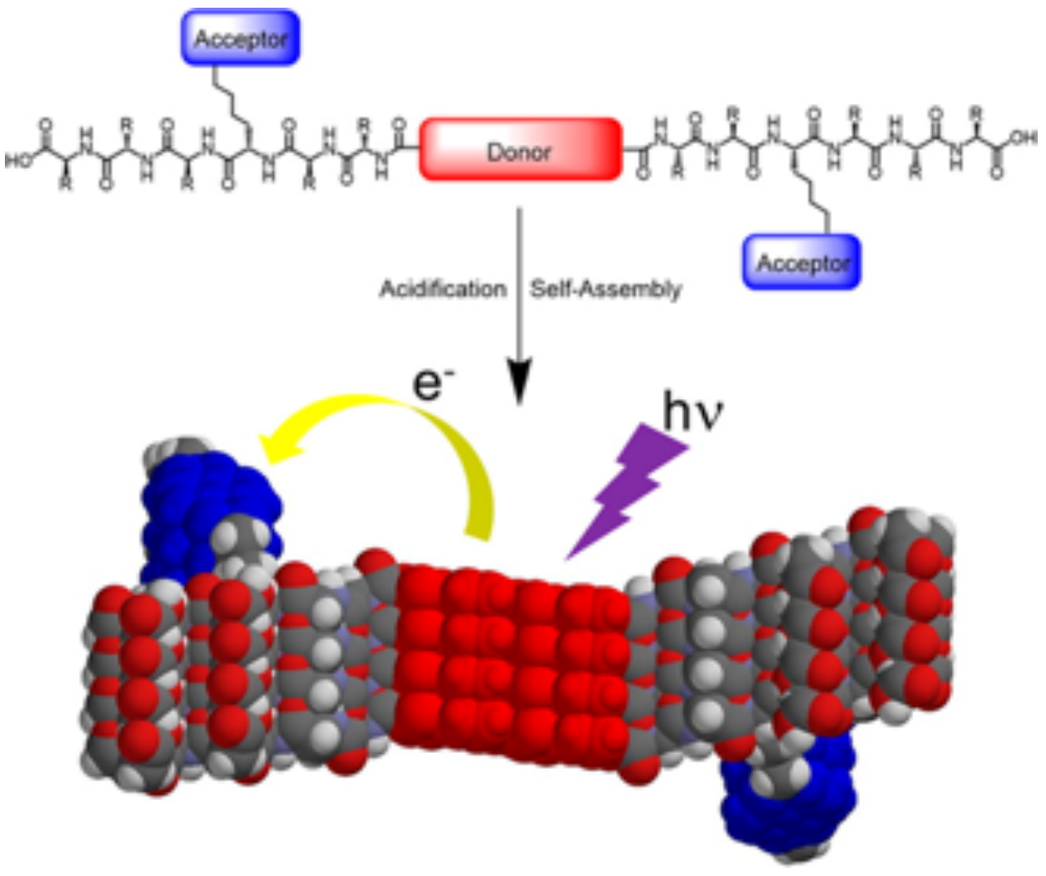
Supramolecular biocompatible optoelectronics

- Synthetic π-conjugated peptides can self-assemble into 10-100 nm fibers
- Fibers possess emergent optical and electronic functionality due to e^- delocalization along overlapping *p* orbitals
- Absorption of UV light produces transient electric fields, exciton generation, and organic photovoltaic activity





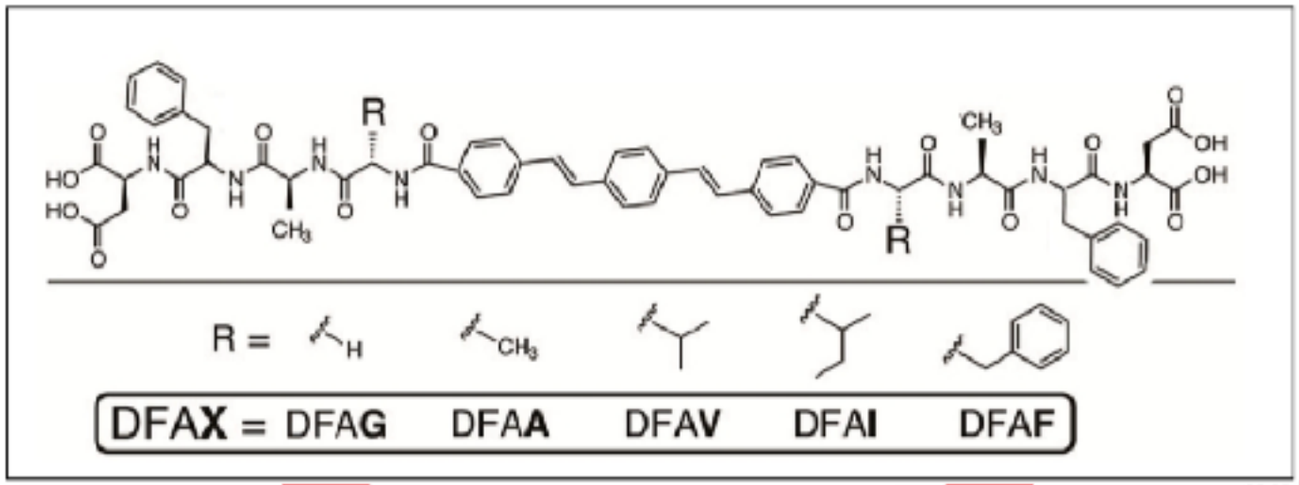


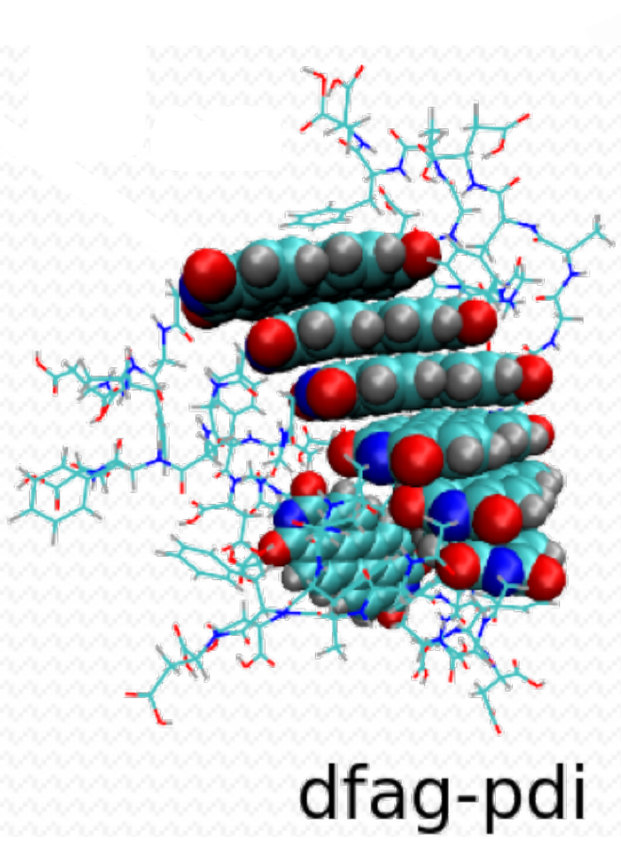


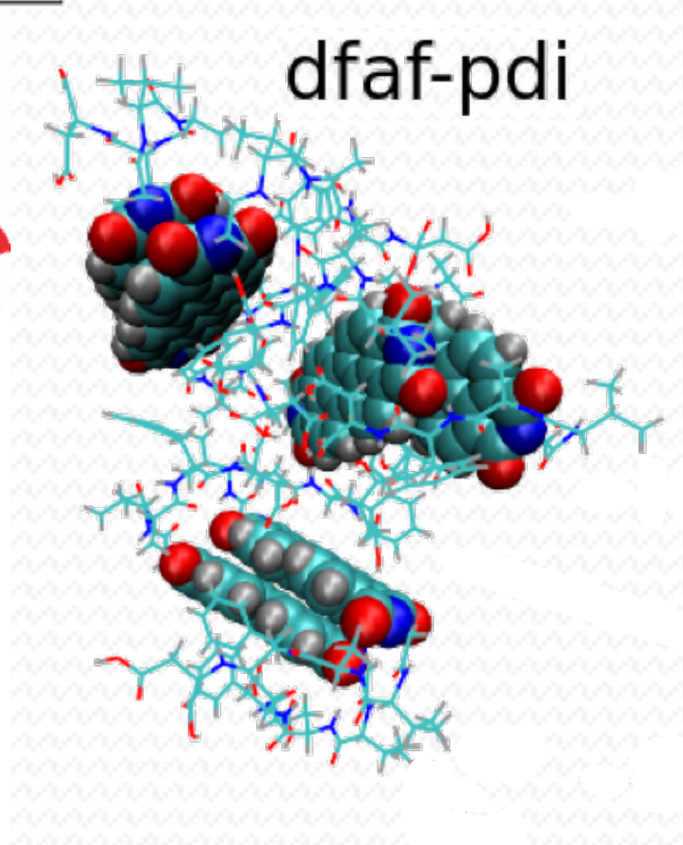
S.R. Diegelmann, J.M. Gorham, and J.D. Tovar JACS 130 42 13840-13841 (2008)
B.A. Thurston, E.P. Shapera, J.D. Tovar, A. Schleife, and ALF (submitted, 2019)
A.M. Sanders, T.J. Magnanelli, A.E. Bragg, and J.D. Tovar JACS 138 10 3362-3370 (2016)
https://phys.org/news/2011-08-smart-skin-electronics-temporary-tattoo.html

Sequence–structure–function relation

- Peptide-wing and π-core sequence programs self-assembly behavior
- Self-assembled structure governs optical and electronic function



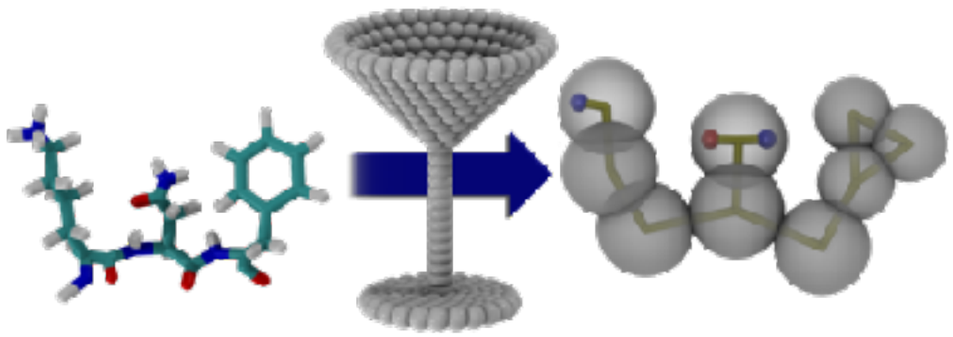




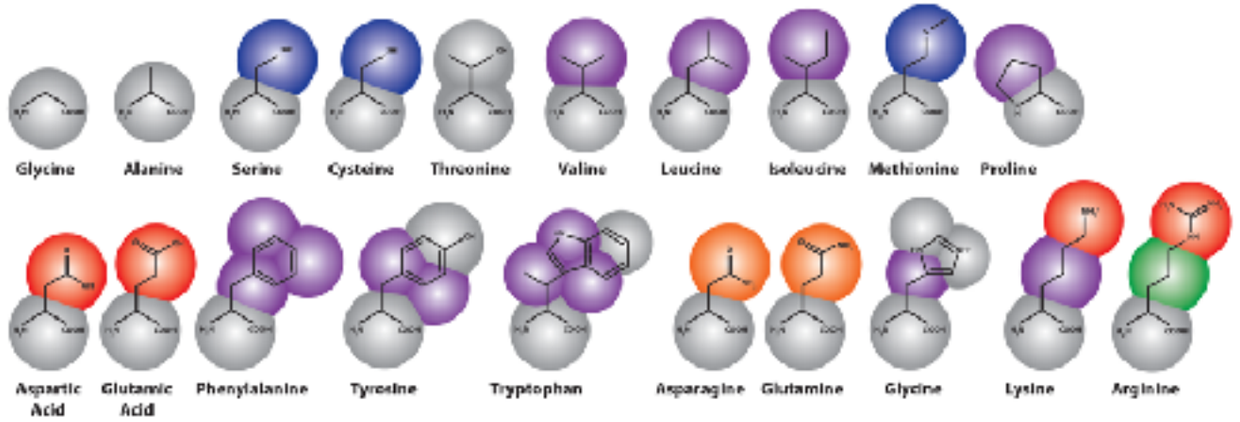
B.A. Thurston and ALF Mol. Sim. 44 11 930-945 (2018)

Coarse-grained MD of oligopeptide assembly

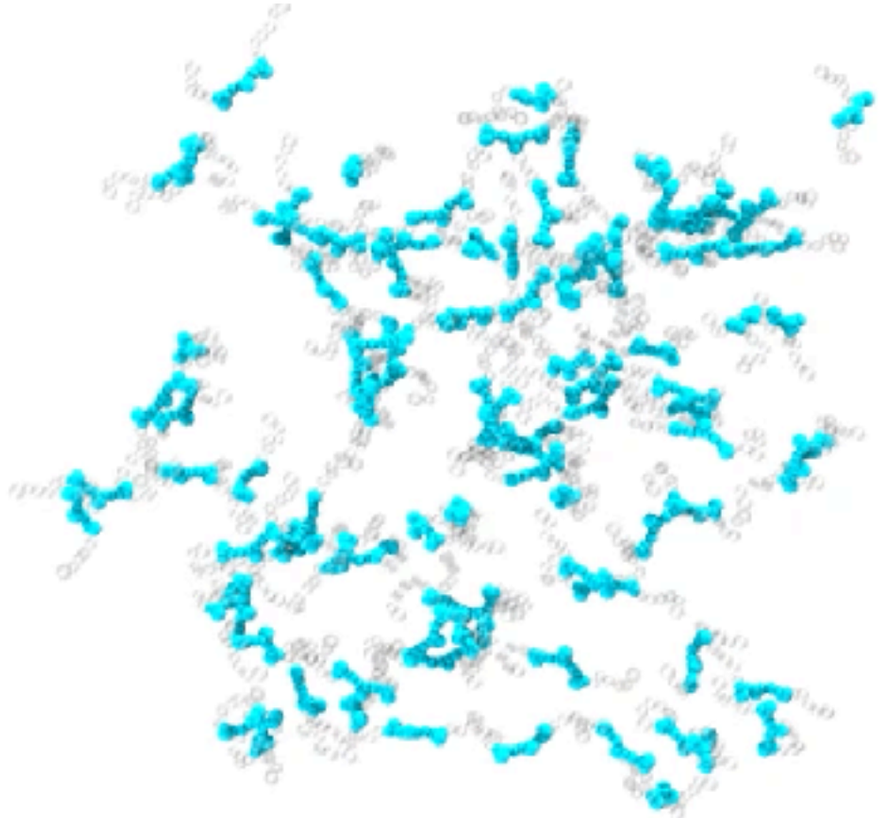
- Coarse-grained MARTINI bead-level representation of oligopeptides
- Compromise between accuracy and speed — can predict aggregation of hundreds of oligopeptides over microseconds



MARTINI



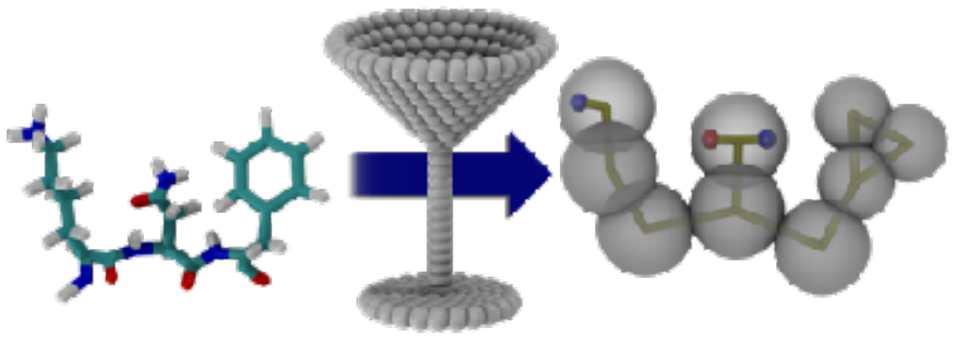
96 × DGAG-OPV3-GAGD
GROMACS 2018
Martini w/ explicit non-polarizable water
T = 298 K, P = 1 bar
t = 3,000 ns (100 h wall time)



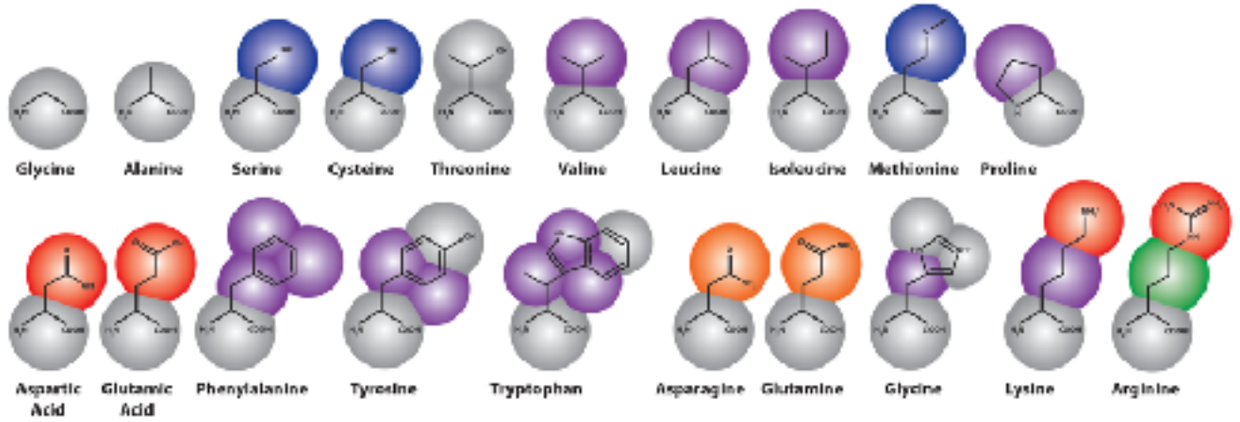
D.H. de Jong, G. Singh, W.F.D. Bennett, C. Arnarez, T.A. Wassenaar, L.V. Schäfer, X. Periole, D.P. Tieleman, S.J. Marrink J. Chem. Th. Comp, 9:687–697 (2013)

Coarse-grained MD of oligopeptide assembly

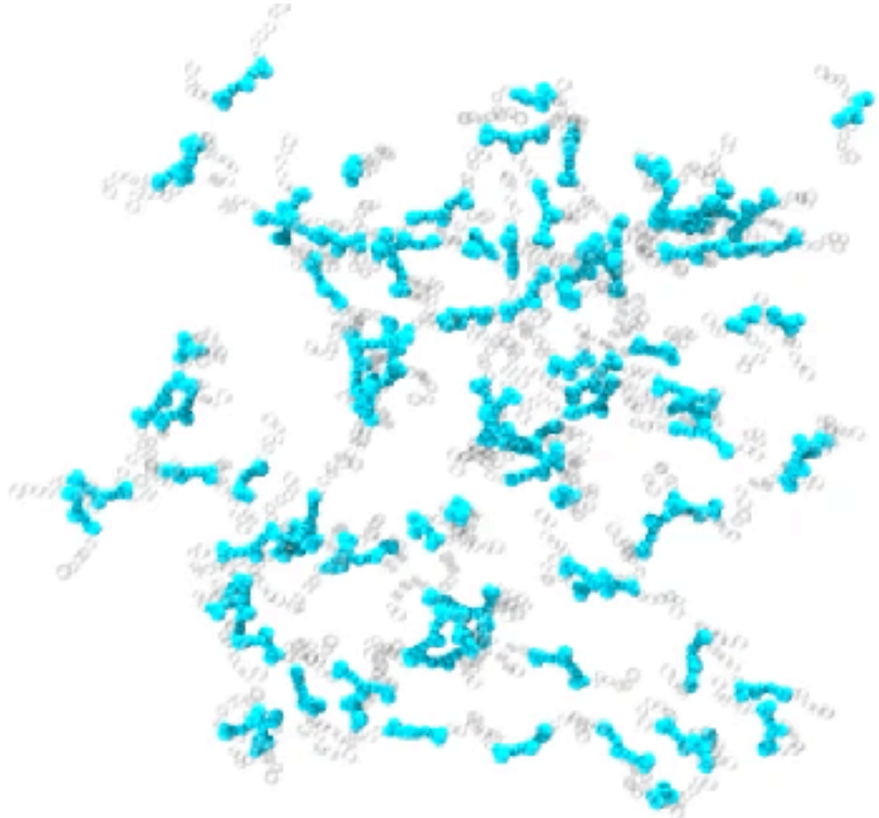
- Coarse-grained MARTINI bead-level representation of oligopeptides
- Compromise between accuracy and speed — can predict aggregation of hundreds of oligopeptides over microseconds



MARTINI



96 × DGAG-OPV3-GAGD
GROMACS 2018
Martini w/ explicit non-polarizable water
T = 298 K, P = 1 bar
t = 3,000 ns (100 h wall time)



D.H. de Jong, G. Singh, W.F.D. Bennett, C. Arnarez, T.A. Wassenaar, L.V. Schäfer, X. Periole, D.P. Tieleman, S.J. Marrink J. Chem. Th. Comp, 9:687–697 (2013)

Coarse-grained MD of oligopeptide assembly

- Coarse-grained MARTINI bead-level representation of oligopeptides
- Compromise between accuracy and speed — can predict aggregation of hundreds of oligopeptides over microseconds

The curse of dimensionality:

The DXXX-Π-XXXD family comprises 20^3 = 8,000 sequences for each Π core

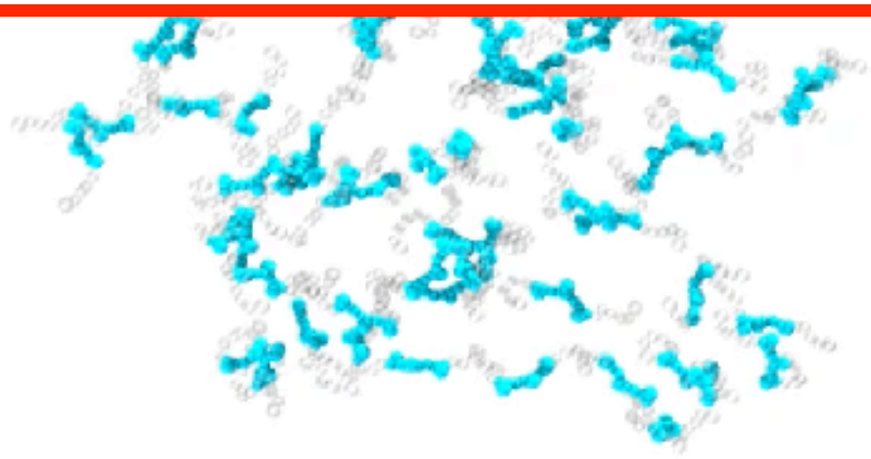
DXXXX-Π-XXXXD $\Longrightarrow$ 20^4 = 160,000

DXXXXX-Π-XXXXXD $\Longrightarrow$ 20^5 = 3,200,000

…

Trial-and-improvement AA or CG simulation too slow for high-throughput virtual screening and rational design

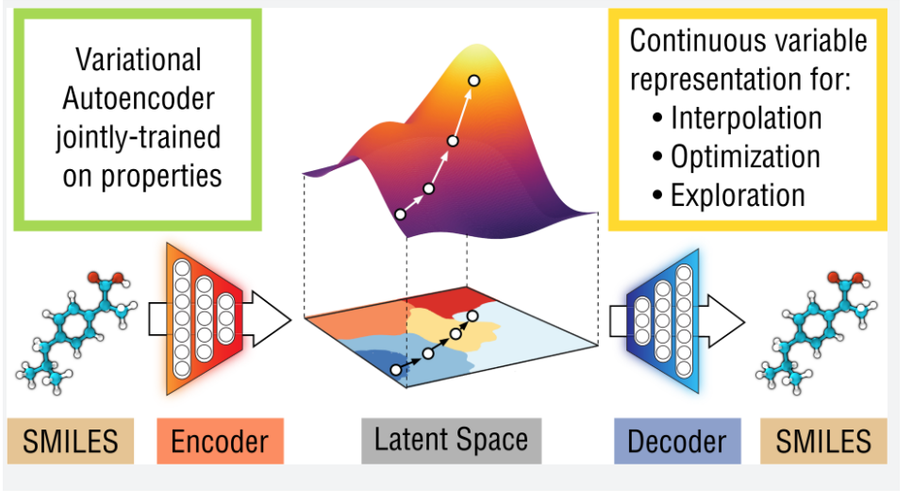
Martini w/ explicit non-polarizable water
T = 298 K, P = 1 bar
t = 3,000 ns (100 h wall time)



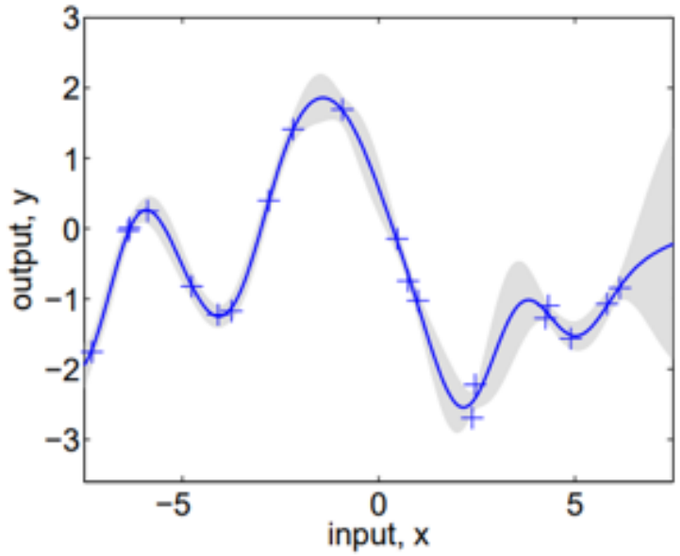
D.H. de Jong, G. Singh, W.F.D. Bennett, C. Arnarez, T.A. Wassenaar, L.V. Schäfer, X. Periole, D.P. Tieleman, S.J. Marrink J. Chem. Th. Comp, 9:687–697 (2013)

Machine learning can help

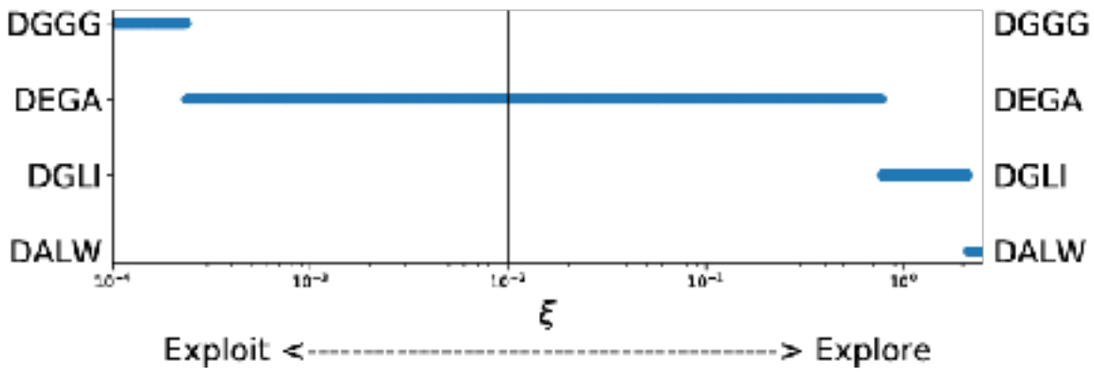
1 **Unsupervised** nonlinear learning of low-dimensional oligopeptides representations using **variational autoencoders**



2 **Supervised** learning of sequence—morphology relation using **Gaussian process regression**



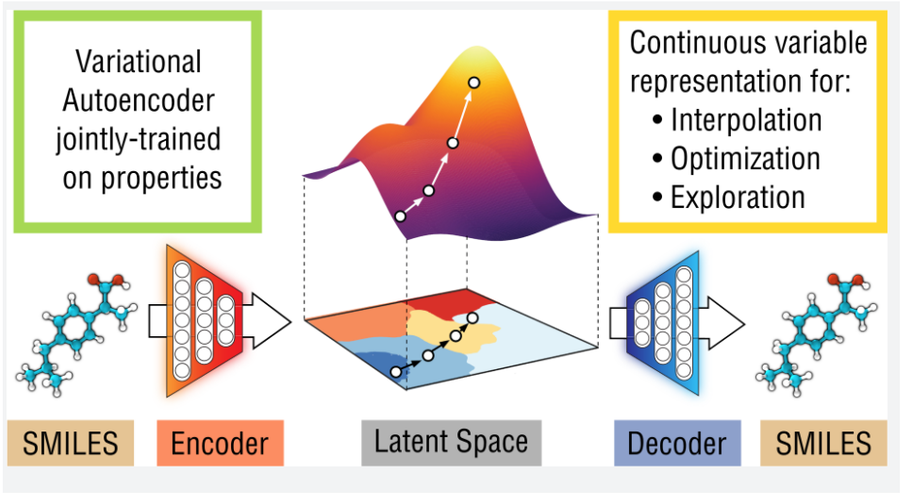
3 **Active learning** to optimally deploy computational effort to explore oligopeptide sequence space



R. Gomez-Bombnarelli, J.N. Wei, ..., and A. Aspuru-Guzik ACS Central Science 4 268-276 (2018) C.E. Rasmussen "Gaussian processes in machine learning" Springer, 2003.

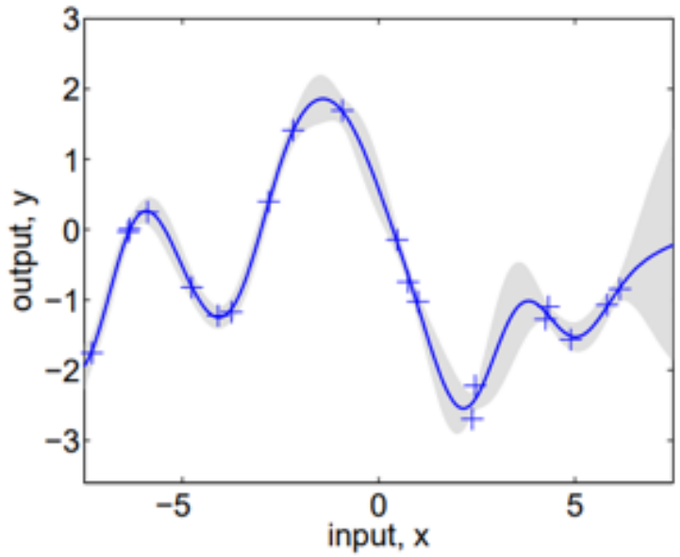
Machine learning can help

1 **Unsupervised** nonlinear learning of low-dimensional oligopeptides representations using **variational autoencoders**



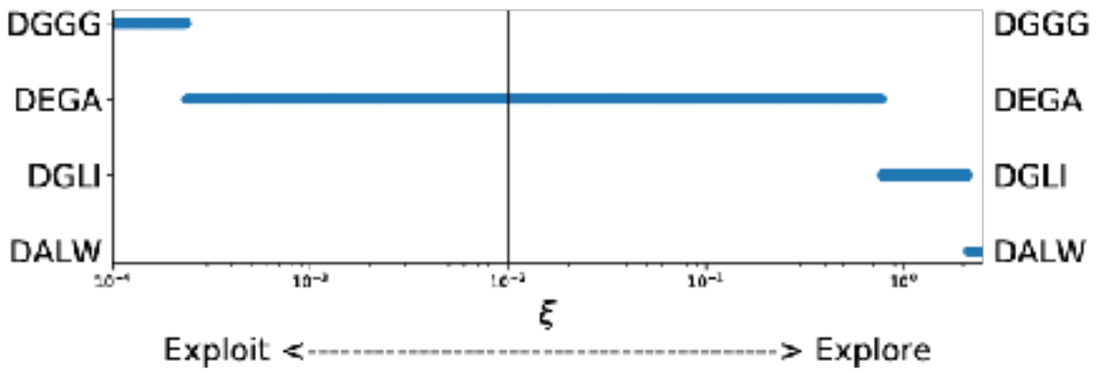
Learn featurization

2 **Supervised** learning of sequence—morphology relation using **Gaussian process regression**



Estimate fitness

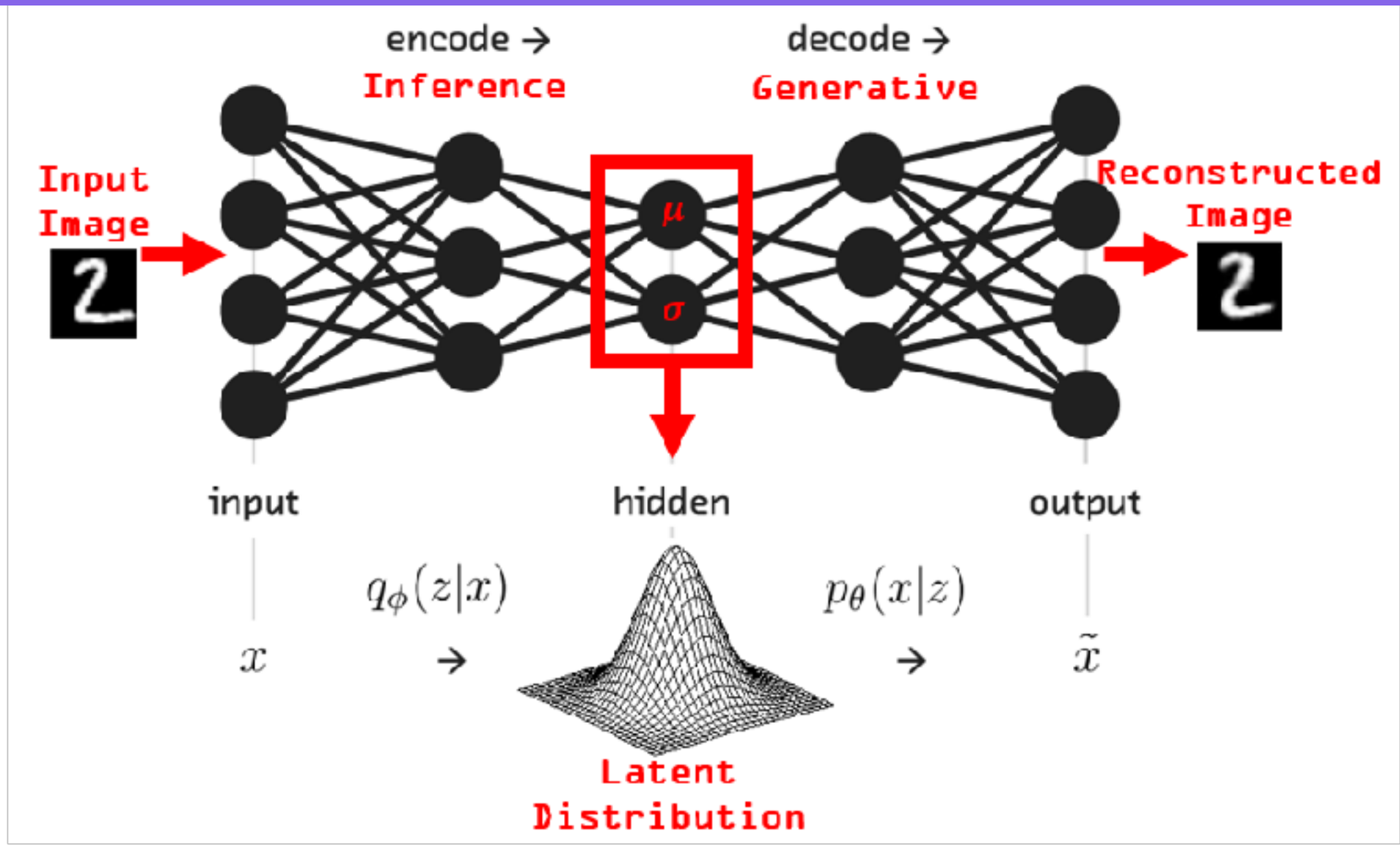
3 **Active learning** to optimally deploy computational effort to explore oligopeptide sequence space



Explore sequence space

R. Gomez-Bombnarelli, J.N. Wei, ..., and A. Aspuru-Guzik ACS Central Science 4 268-276 (2018) C.E. Rasmussen "Gaussian processes in machine learning" Springer, 2003.

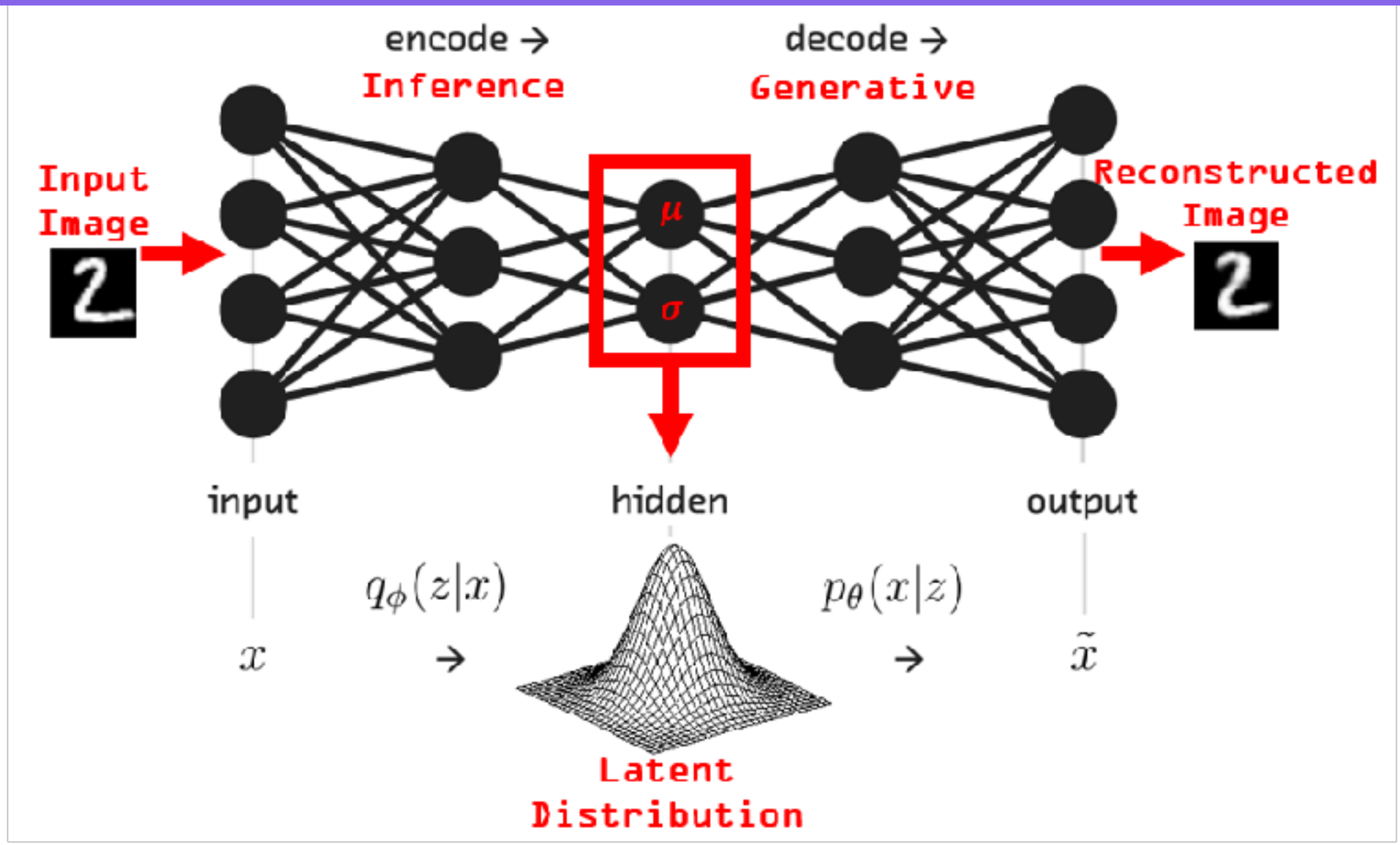
Learn oligopeptide featurization



- Variational autoencoders comprise two linked deep neural networks
  — The **encoder** Φ learns to project samples **x** into a low-dim latent space **z**
  — The **decoder** θ reconstructs samples **x** from latent space vectors **z**
- Trained to reconstruct its own inputs (i.e., auto-encode) the VAE performs **unsupervised nonlinear dimensionality reduction**

https://i.imgur.com/ZN6MyTx.png

Learn oligopeptide featurization



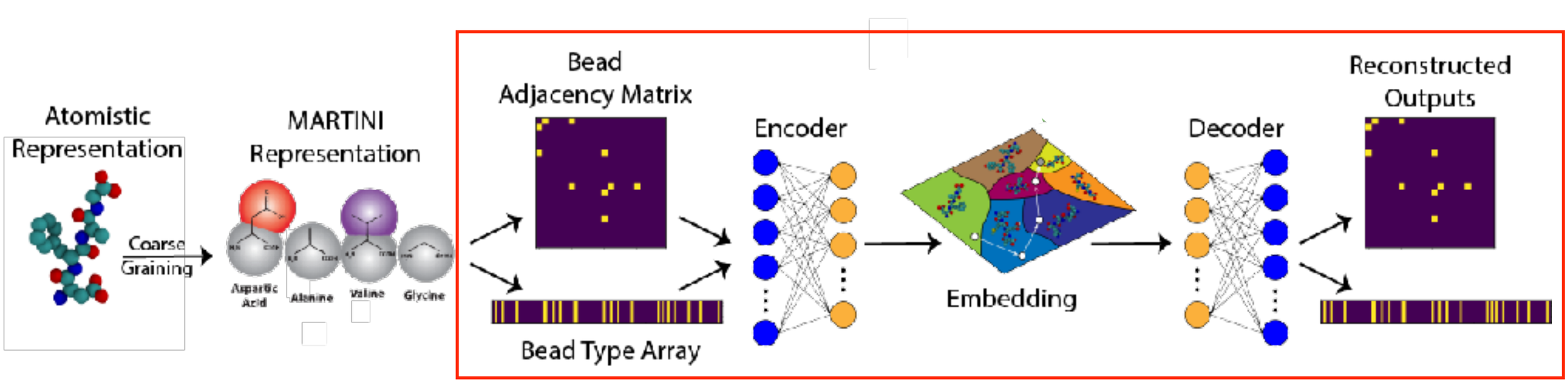
- The latent space is regularized to a Gaussian for mathematical convenience — the encoder infers (μ,σ) for each input
- A trained VAE is **generative** — decoder can hallucinate new samples from arbitrary latent space vectors sampled from the latent space

https://i.imgur.com/ZN6MyTx.png

Learn oligopeptide featurization

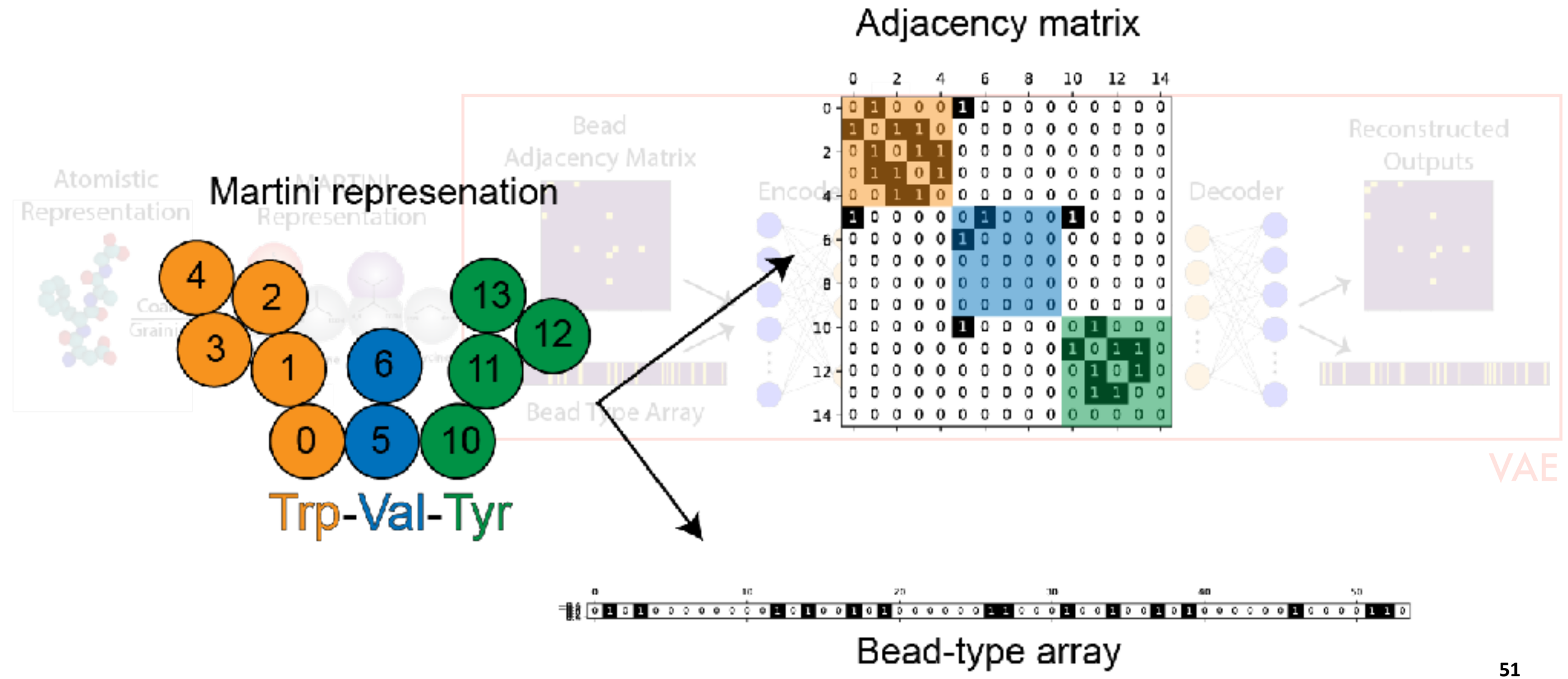
1

- We train VAEs to learn latent space embeddings = essential featurizations of all 20^n oligopeptides for a given Π core
- Represent oligopeptides to VAE as:
  (i) vector of Martini bead types (composition)
  (ii) bead adjacency matrix (molecular topology)



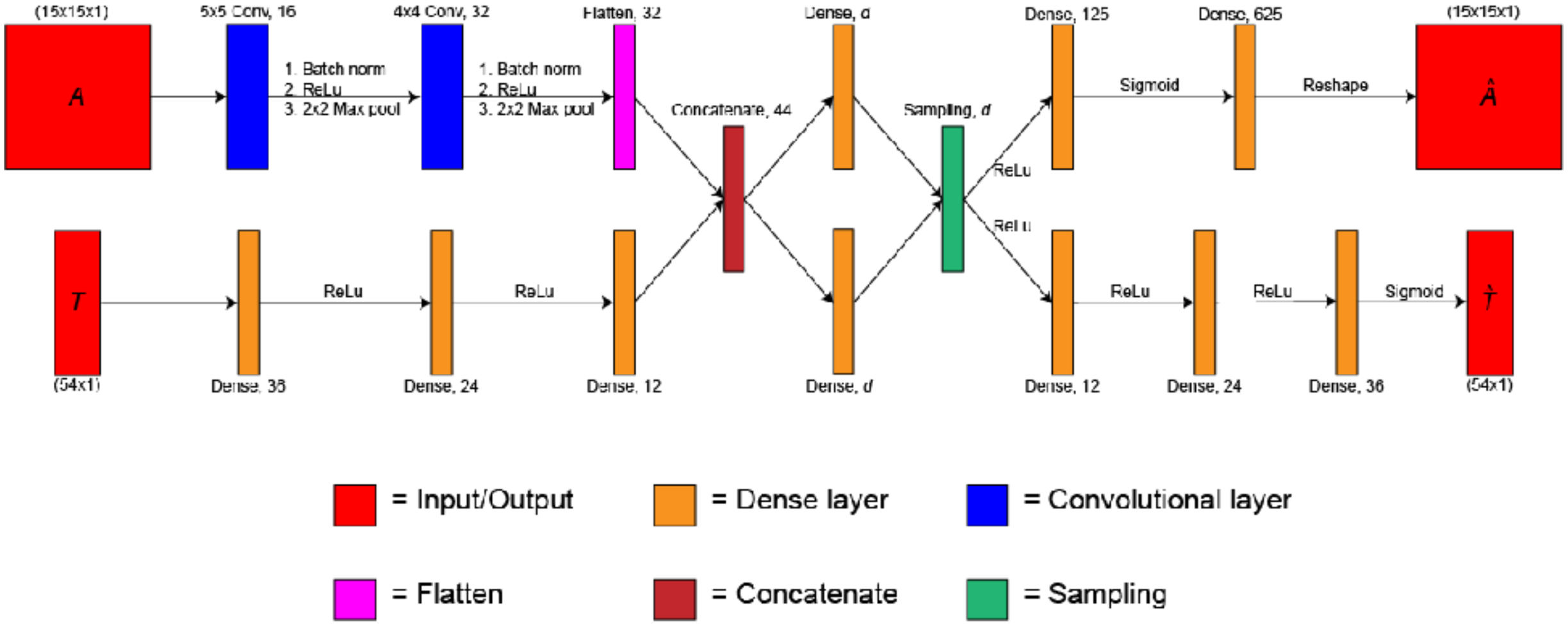
Learn oligopeptide featurization

- We train VAEs to learn latent space embeddings = essential featurizations of all 20^n oligopeptides for a given Π core
- Represent oligopeptides to VAE as:
  (i) vector of Martini bead types (composition)
  (ii) bead adjacency matrix (molecular topology)



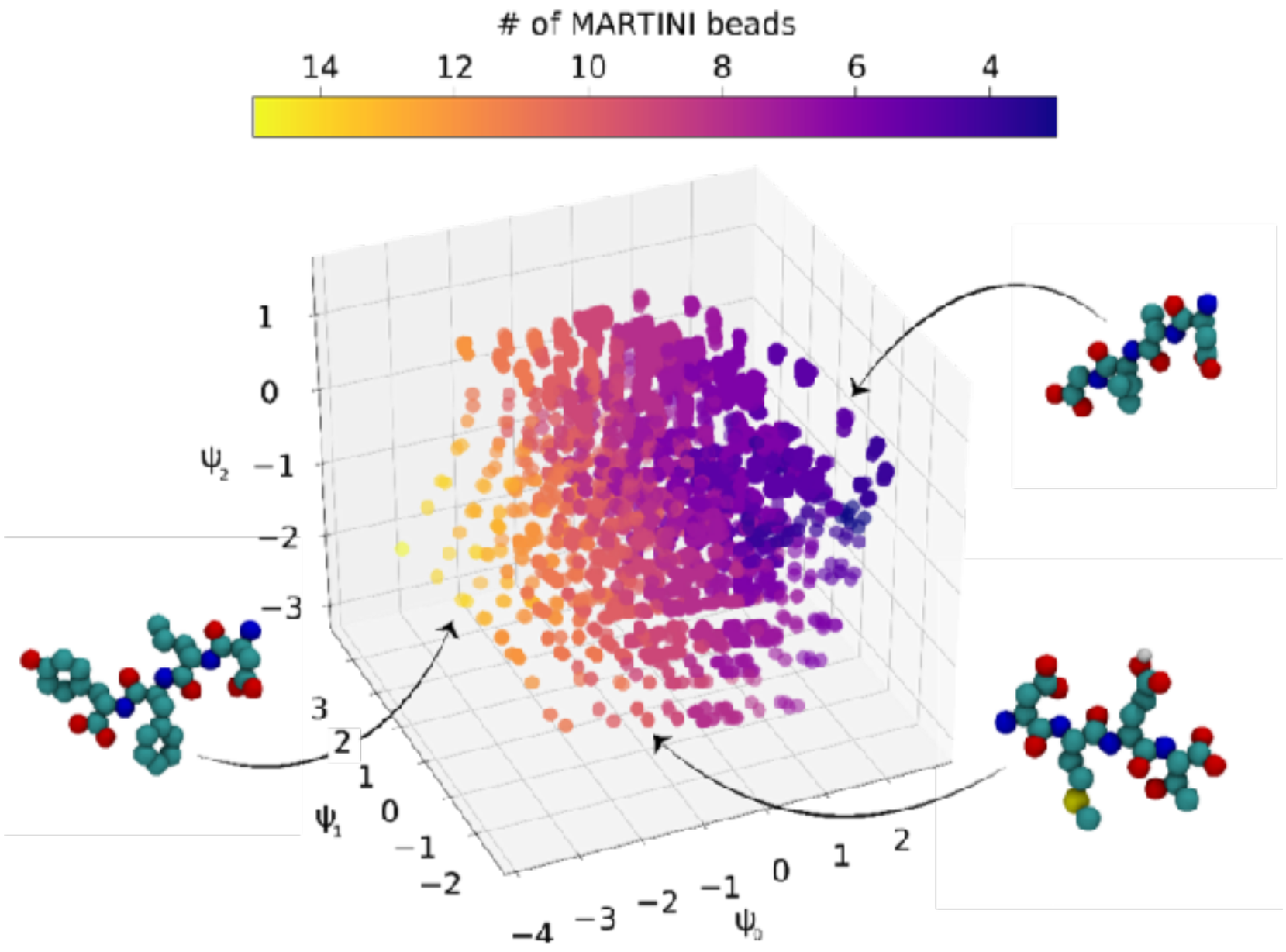
Learn oligopeptide featurization

- We train VAEs to learn latent space embeddings = essential featurizations of all 20^n oligopeptides for a given Π core
- Construct and train VAEs in TensorFlow to minimize loss function error on cross-validation partitions



Regress oligopeptide fitness over latent space 2

- VAE latent space provides 3D featurization of oligopeptides
  — leading order description of oligopeptide composition and structure
  — embeds similar oligopeptides close together



Regress oligopeptide fitness over latent space 2

- Run Martini CG simulations for O(10) randomly selected oligopeptides
- "Fitness" is **number of inter-core contacts** in self-assembled aggregate
  — more core contacts ⇨ better *p* orbital overlap and e^-/h^+ paths
- Construct supervised learning of **Gaussian process regression** model
  — fitness = *f*(VAE latent space)

GPR assumes data {**y**} can be represented as a sample from a multivariate Gaussian distribution over **x**

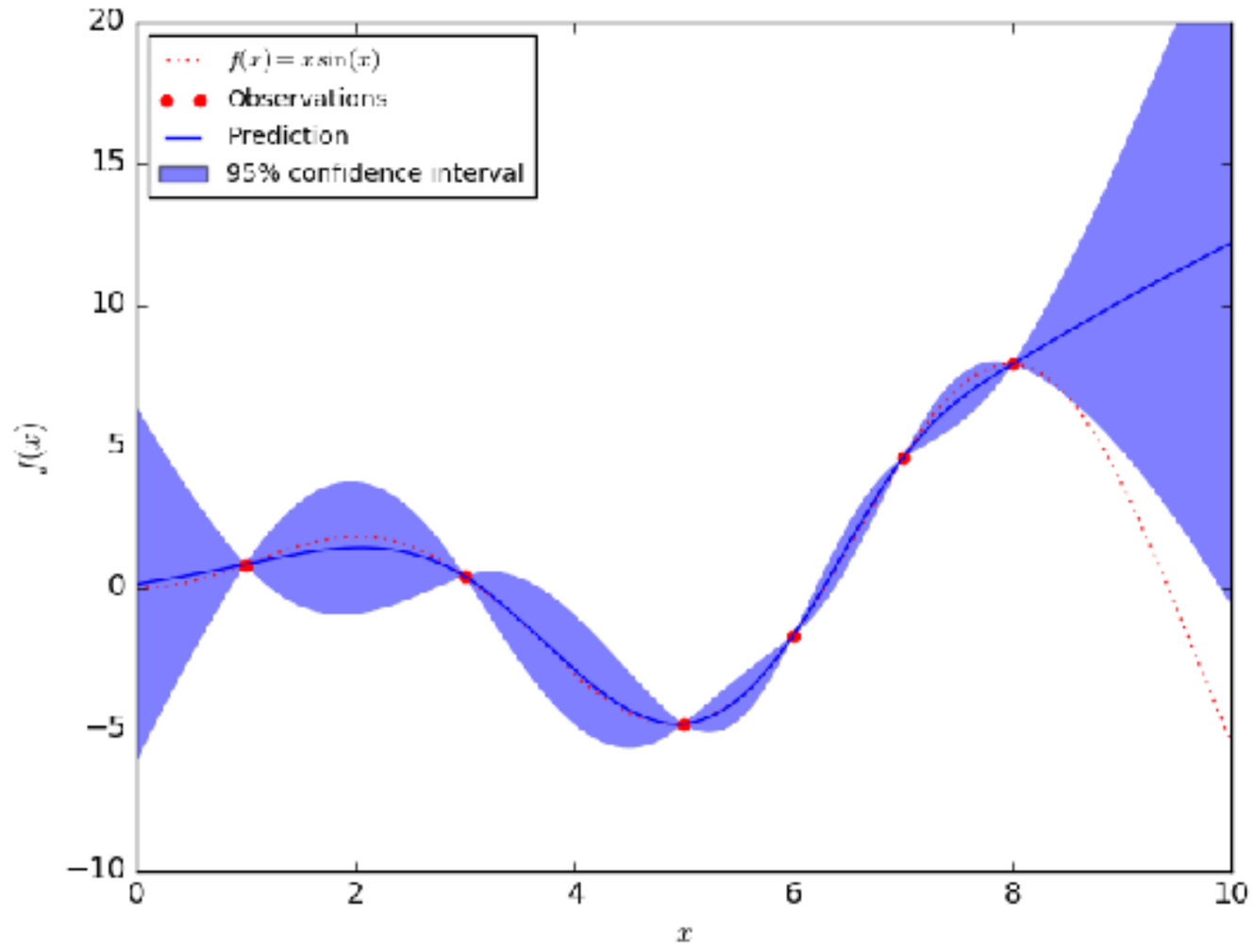
training data → $\begin{bmatrix} \mathbf{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K & K_*^{\mathrm{T}} \\ K_* & K_{**} \end{bmatrix}\right)$ ← covariance

predictions →

Conditional probability of new datum y* given training {**y**} follows Gaussian

$$\bar{y}_* = K_* K^{-1} \mathbf{y}$$

$$\mathrm{var}(y_*) = K_{**} - K_* K^{-1} K_*^{\mathrm{T}}$$

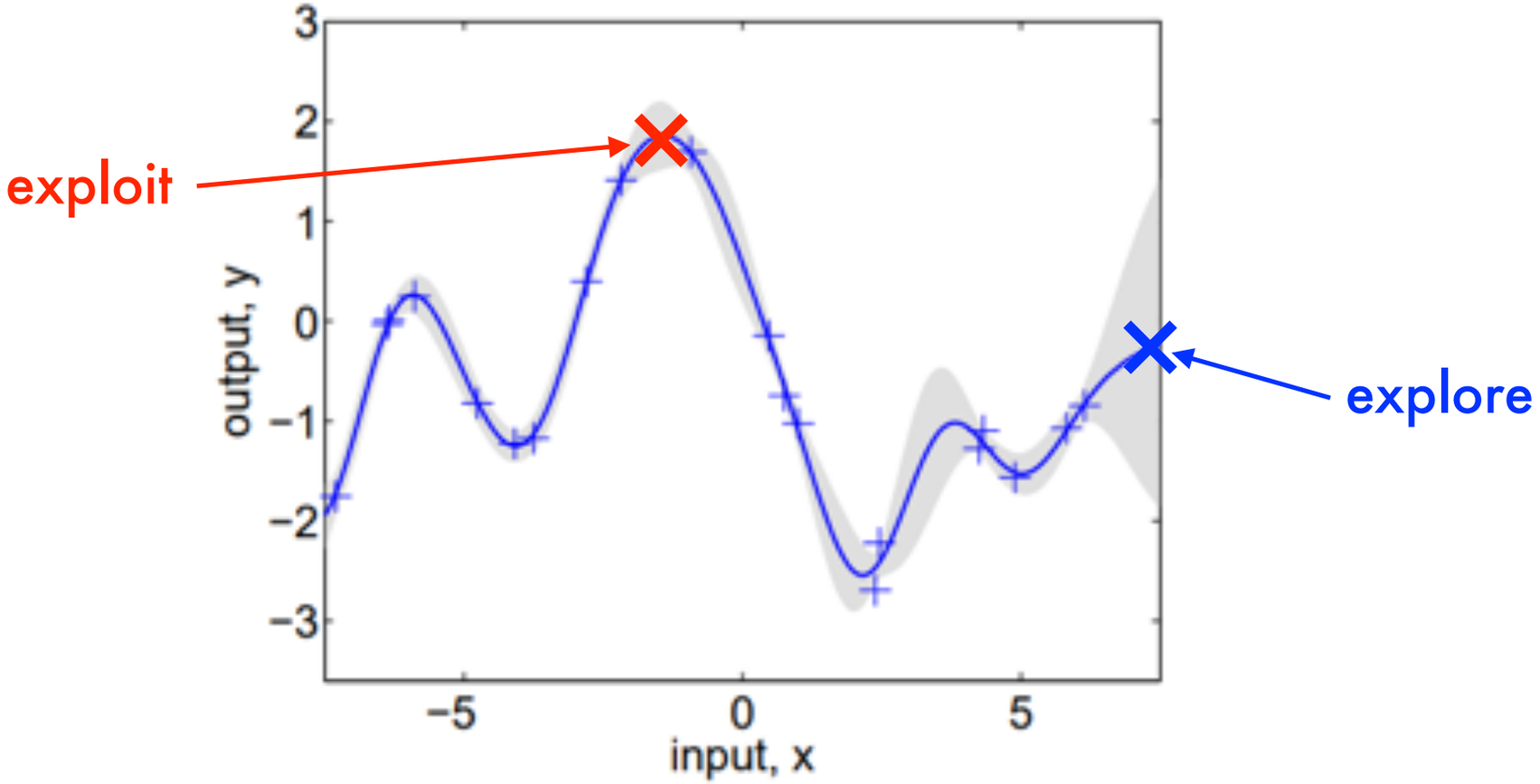


C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006
M. Ebden "Gaussian Processes for Regression: A Quick Introduction", August 2008 arXiv:1505.02965v2

Optimally sample sequence space

- Use GPR to inform **optimal traversal of sequence space** in virtuous cycle
  — GPR strengthens with samples ⇔ better guidance from GPR

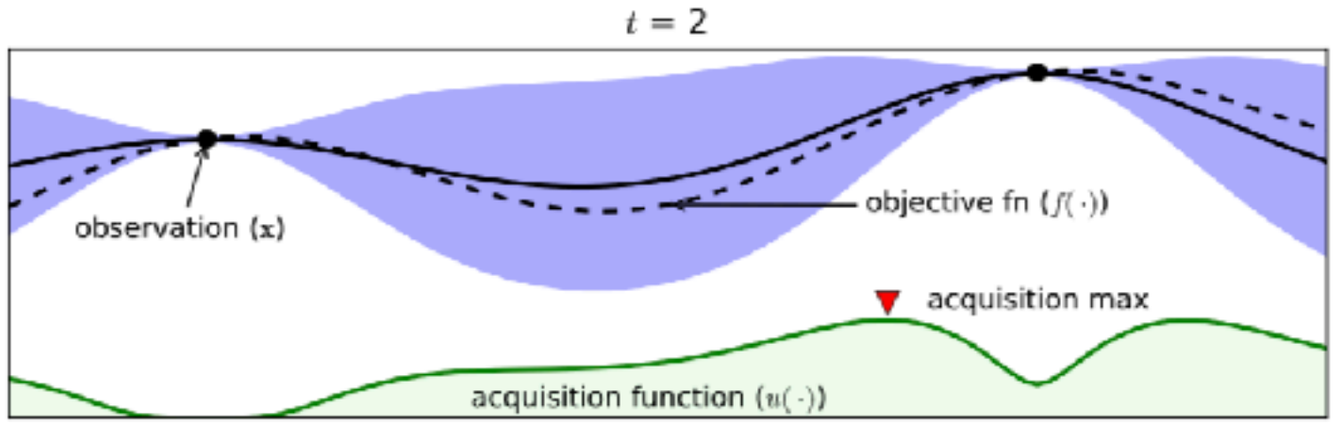
- **Active learning** paradigm to select new oligopeptides to simulate from GPR
  — **exploit** : best candidate picked by GPR
  — **explore** : sample candidates where GPR has maximum uncertainty



C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006

Optimally sample sequence space

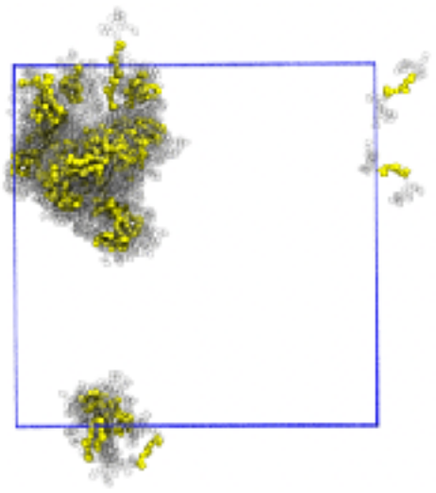
- **Expected improvement** (EI) acquisition function balances exploit / explore
- Select next oligopeptide to simulate as that which maximizes EI



E. Brochu, V.M. Cora, and N. De Freitas arXiv preprint arXiv:1012.2599 (2010)

Putting it all together…

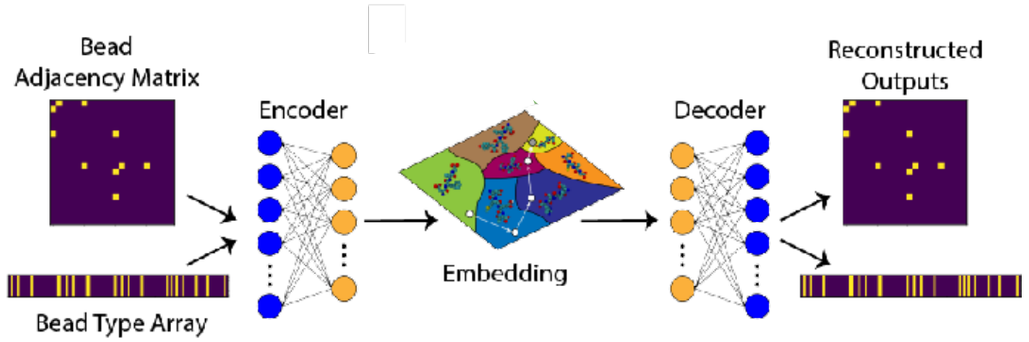
GOAL: Computationally identify optimal assembling DXXX-OPV3-XXXD oligopeptides



0 CG MD simulations

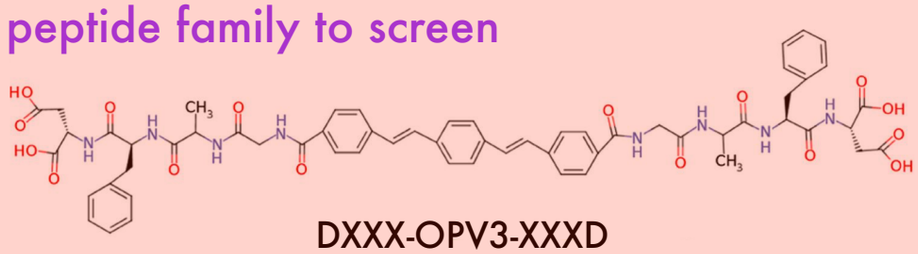
– measure core-core contacts

– R1: O(10); R2+: 3-4

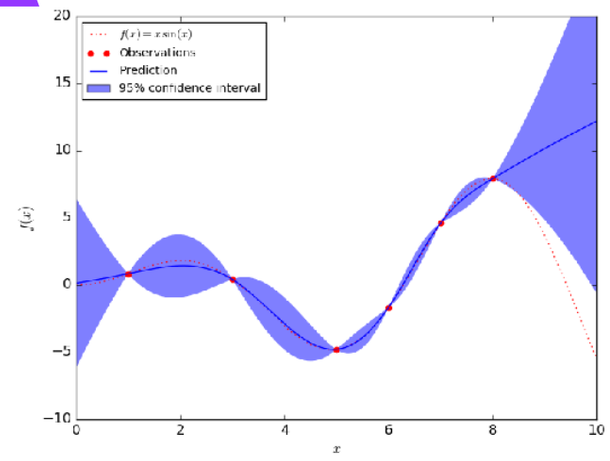


1 unsupervised VAE (re)training and latent space embedding

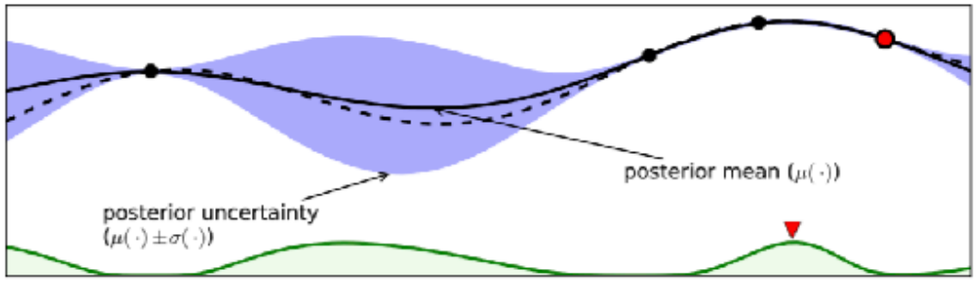
peptide family to screen



DXXX-OPV3-XXXD



2 supervised GPR (re)training over VAE latent space

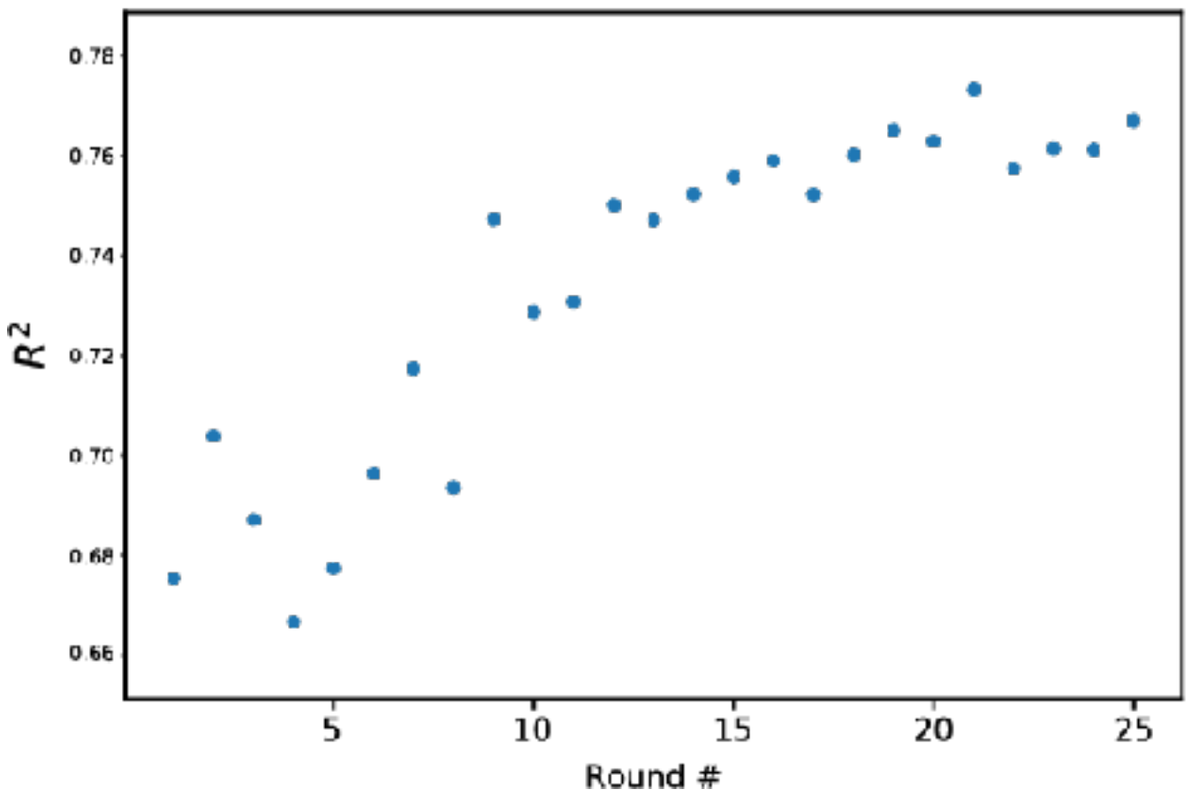


3 active learning of next best oligopeptides to simulate

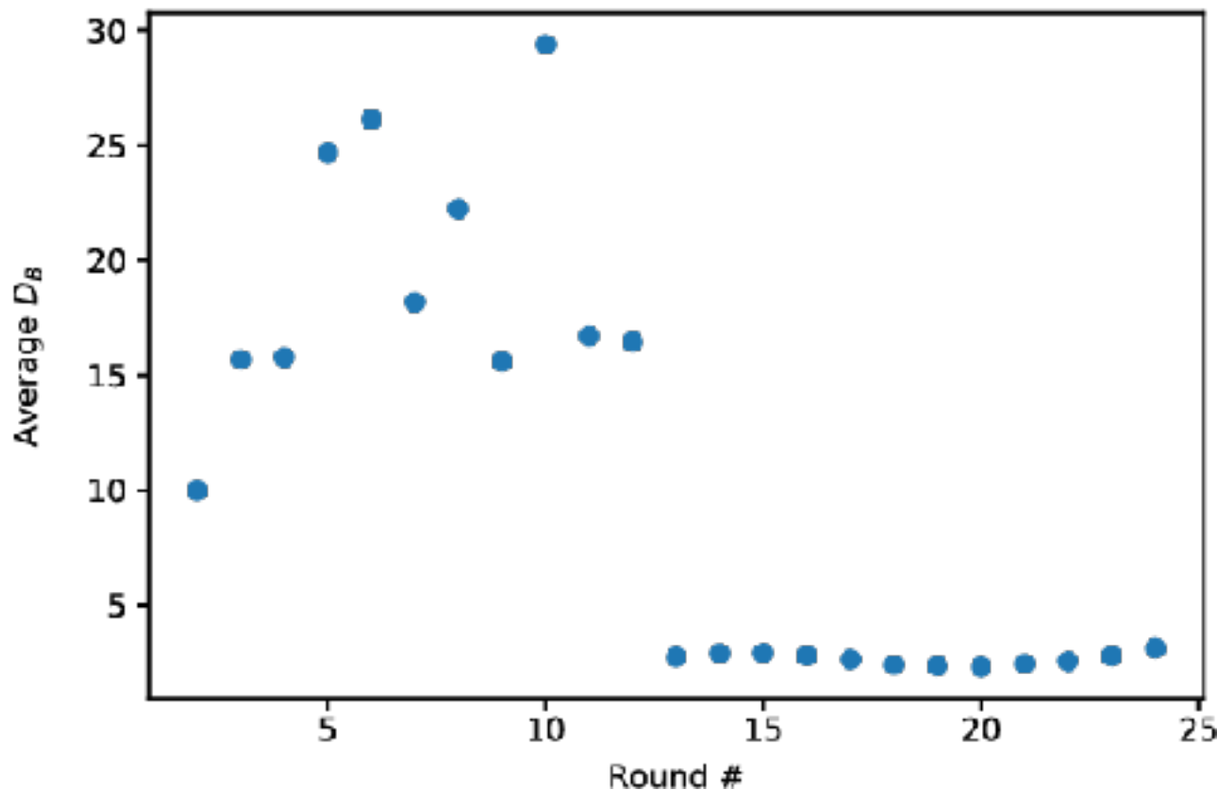
Stopping criteria

- Determine convergences by monitoring GPR model performance
- Terminate when model predictions stop changing with additional samples
- Model converges after 24 rounds, 186 chemistries, 558 μs of simulation

Cross validated R2 score on observed data



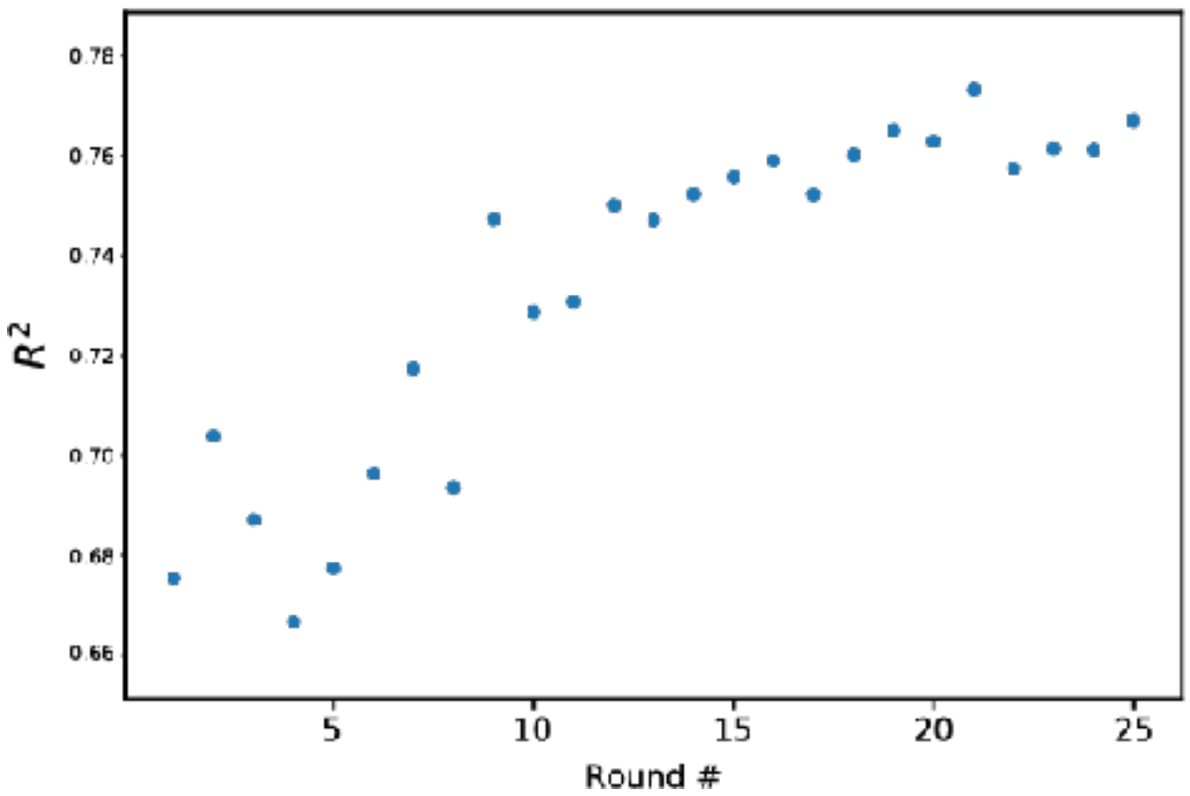
Bhattacharya distance D_B between GPR posteriors (change in posterior with added samples)



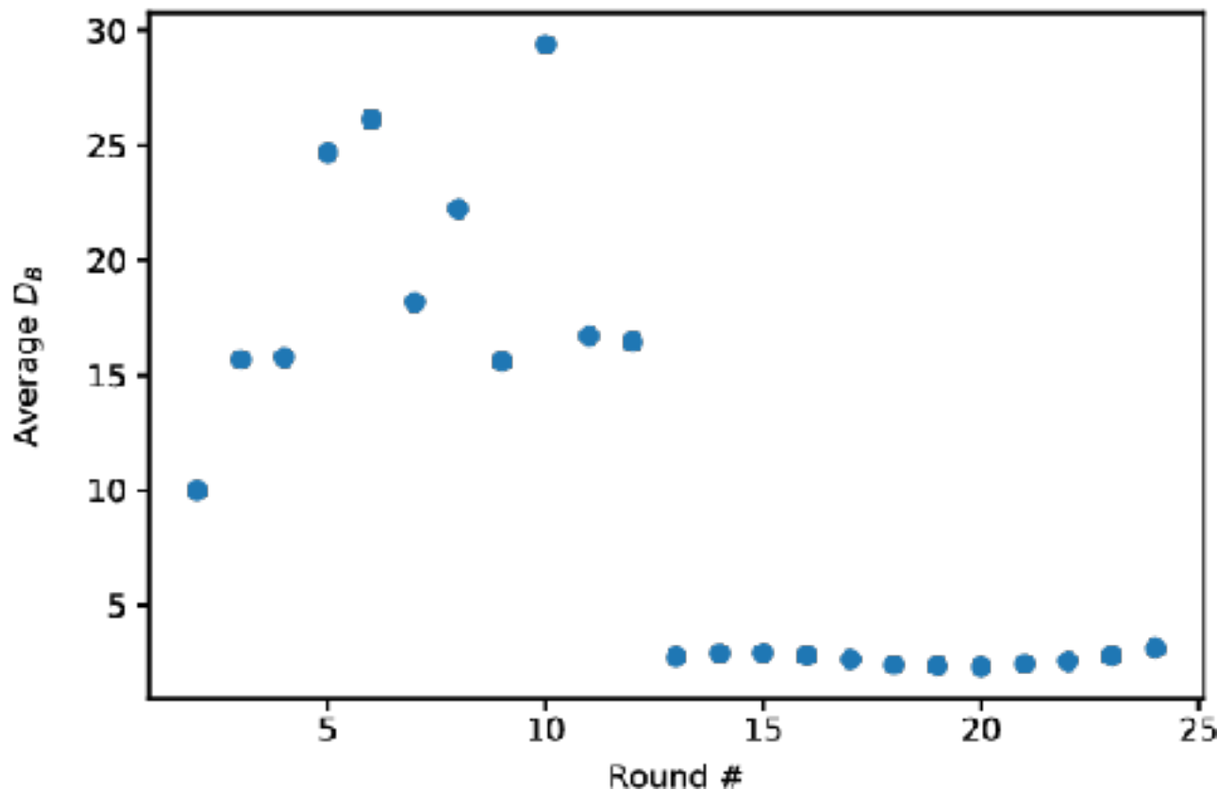
Stopping criteria

- Determine convergences by monitoring GPR model performance
- Terminate when model predictions stop changing with additional samples
- Model converges after 24 rounds, 186 chemistries, 558 μs of simulation

Cross validated R2 score on observed data



Bhattacharya distance D_B between GPR posteriors (change in posterior with added samples)



ML model (GPR) capable of identifying optimal oligopeptides after active learning sampling of only 2.3% of accessible sequence space

Optimal candidates

- Top 10 candidates identified and validated throughout 25 rounds of active learning protocol

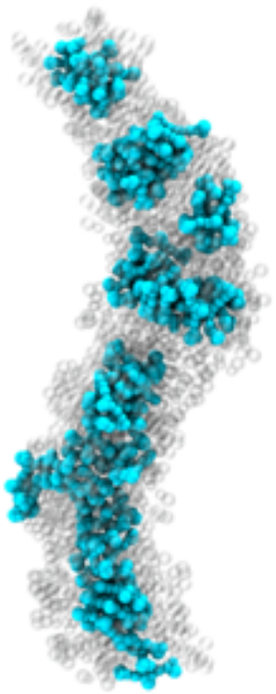
average degree of core-core contacts interaction graph (higher is better)

Chemistry (DXXX)	$\langle \kappa_l \rangle$	Round #
DEAA	6.067	1
DDAI	6.034	0
DIAM	6.008	17
DVAA	5.950	9
DAAV	5.925	19
DGLG	5.922	20
DAEA	5.920	25
DAGI	5.900	21
DGIG	5.883	25
DEAL	5.880	23
⋮	⋮	⋮
DGAG	5.540	0
⋮	⋮	⋮
DFAG	4.984	0

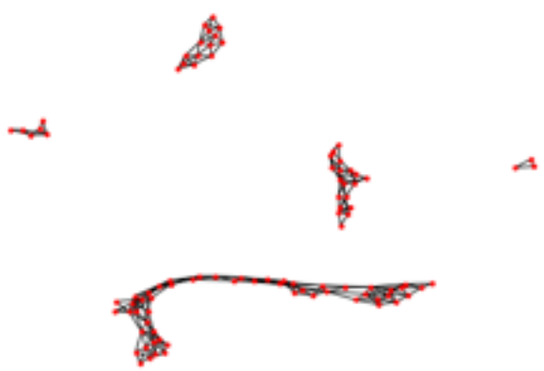
Three classes of oligopeptides

TRAPPERS

DFAG

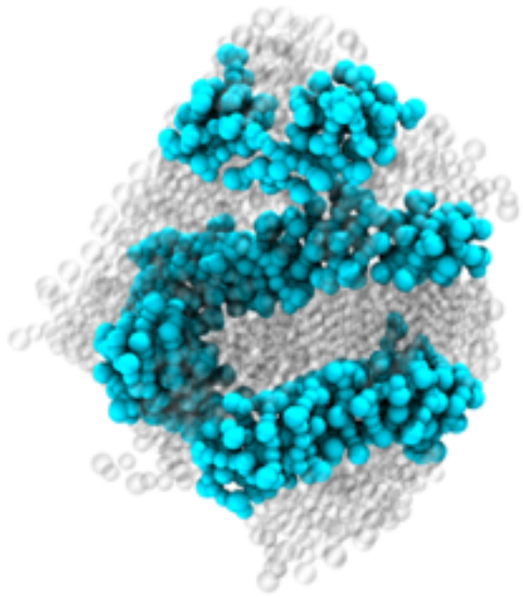


5 clusters formed

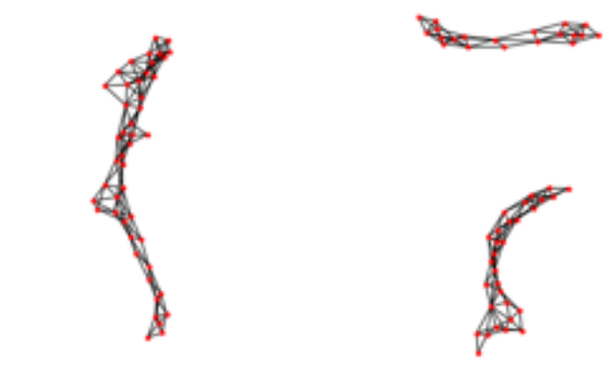


ASSEMBLERS

DGAG

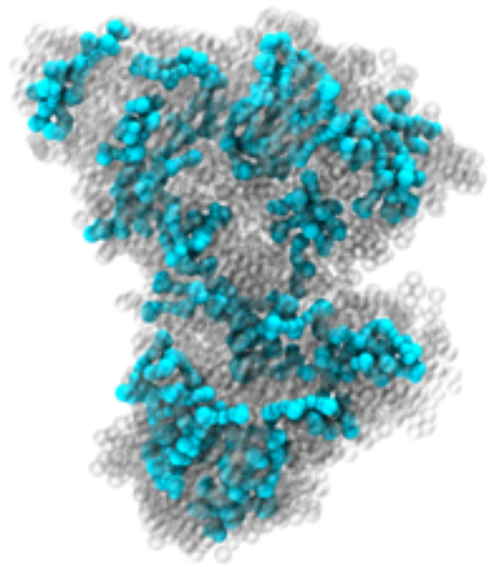


3 clusters formed

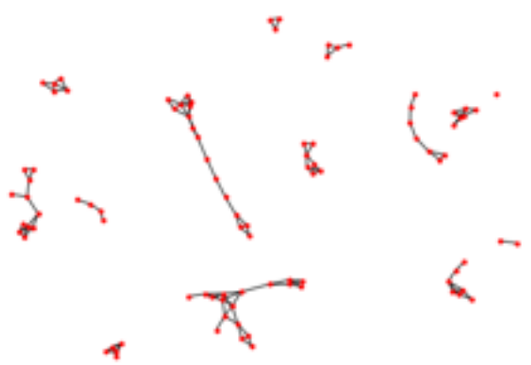


FRAGMENTORS

DWWW

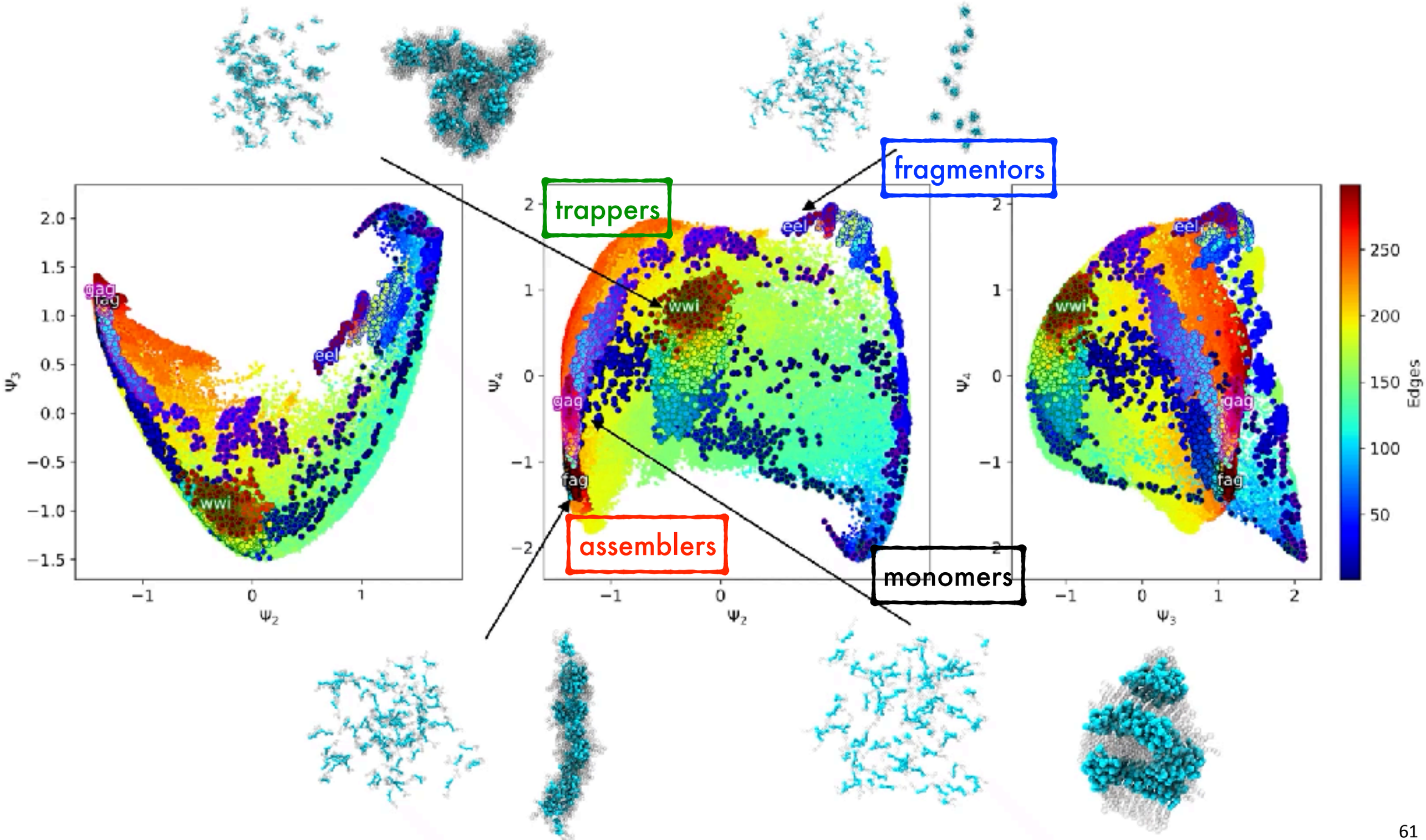


14 clusters formed



Lo-dim viz of assembly pathways

- Diffusion map dimensionality reduction over interaction graphs reveals mechanistic partitioning of the three classes identified by active learning



Lo-dim viz of assembly pathways

- Diffusion map dimensionality reduction over interaction graphs reveals mechanistic partitioning of the three classes identified by active learning