Unsupervised Protein-Ligand Binding Energy Prediction via Neural Euler's Rotation Equation

Wengong Jin, Siranush Sarzikova, Xun Chen, Nir Hacohen, Caroline Uhler

Broad Institute of MIT and Harvard

Structure-based drug discovery

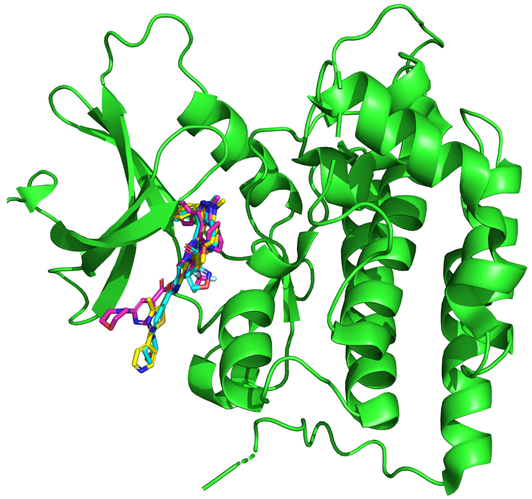
Binding energy prediction (small molecule/antibody ligands)

Protein structure

Ligand structure

Docking

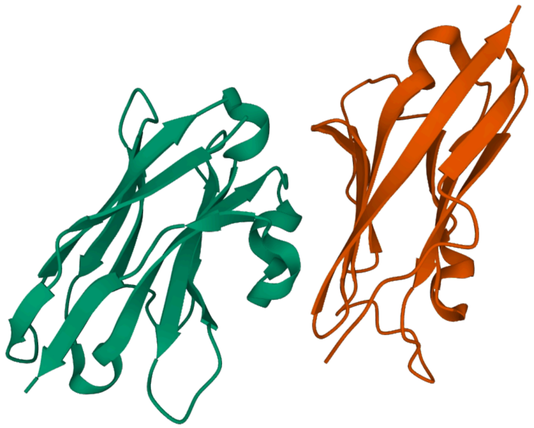
Lots of exciting progress in protein structure prediction and docking



Antigen structure

Antibody structure

Docking



Scoring function

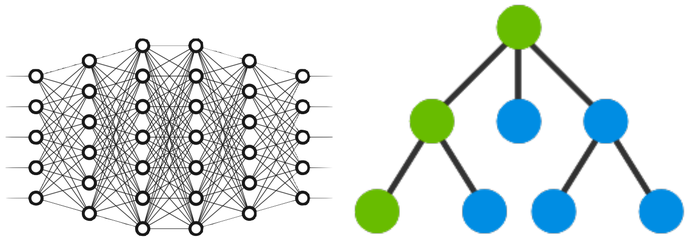
Today's focus

Binding affinity

Current binding energy prediction approach

Supervised and unsupervised approaches

Supervised models

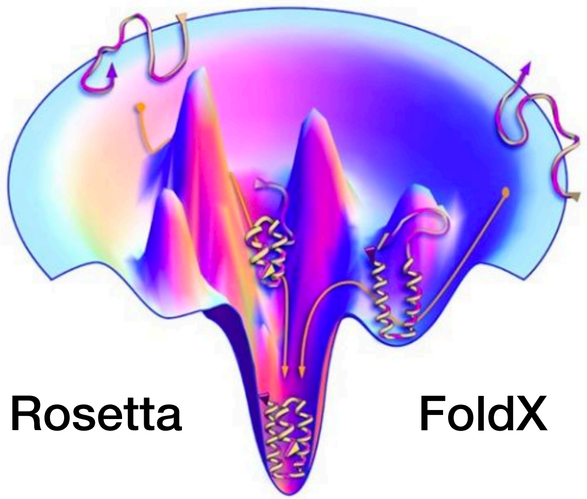


- High accuracy (provided enough training data)
- Need to collect binding data (e.g., phage-display library) for every protein
- Data collection is costly

Unsupervised models

?

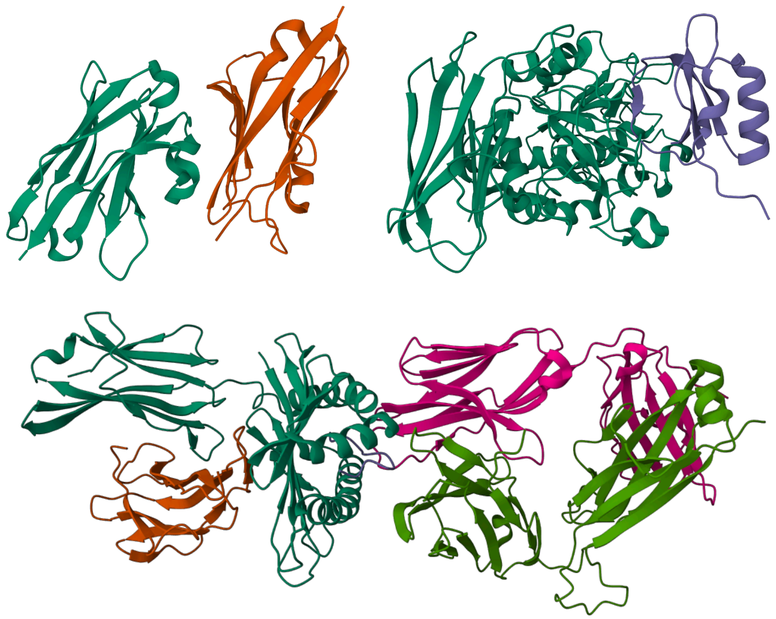
Physics-based models



- Based on first principles
- No training data required
- Extremely slow

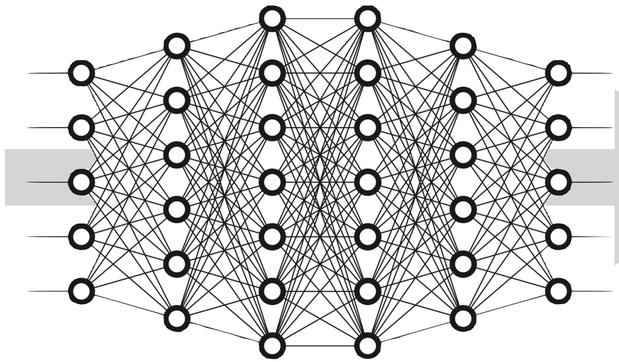
Our method: unsupervised energy-based model

Learn binding free energy from crystal structures (data-driven)

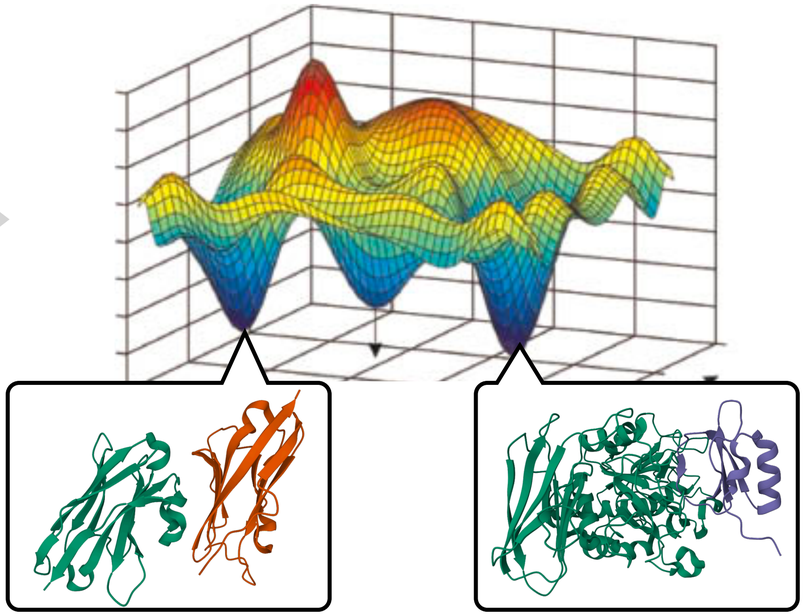


Training set: protein complexes
crystal structures from PDB
(No binding affinity labels)

Energy-based Model
(EBM)



The space of all possible
protein complexes



Binding energy
= -13.6

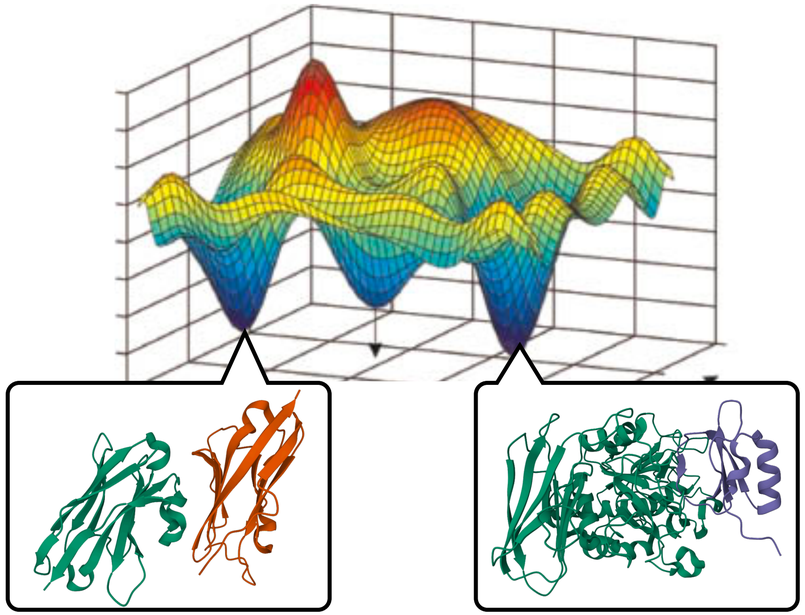
Binding energy
= -15.2

Our method: unsupervised energy-based model

Learn binding free energy from crystal structures (data-driven)

- **Intuition**: crystal structures should be the local minimum of the energy landscape
- Suppose $E_\theta(x)$ is the energy of a protein complex
- Its likelihood is $p(x) \propto \exp(-E_\theta(x))$
- Minimizing the energy of crystal structures = maximizing their likelihood (standard objective in generative models / protein language models)
- Q1: How to parameterize $E_\theta(x)$?
- Q2: What's the training objective?

The space of all possible protein complexes

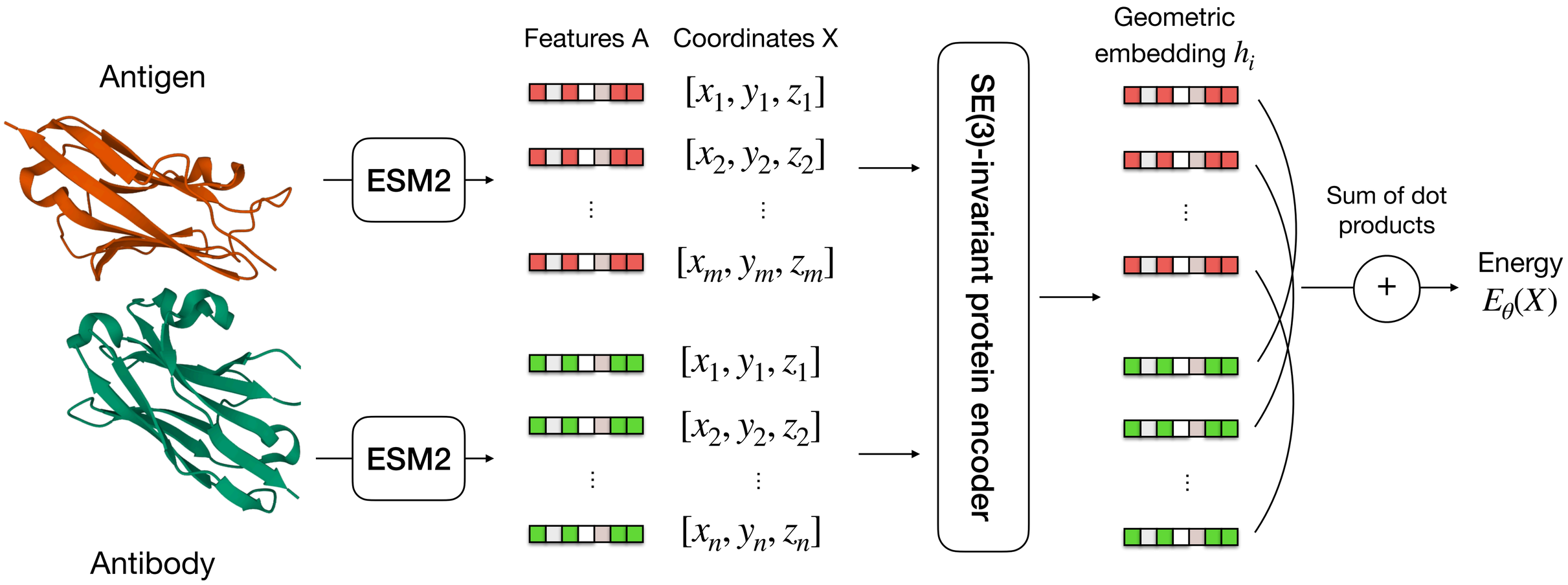


Binding energy = -13.6

Binding energy = -15.2

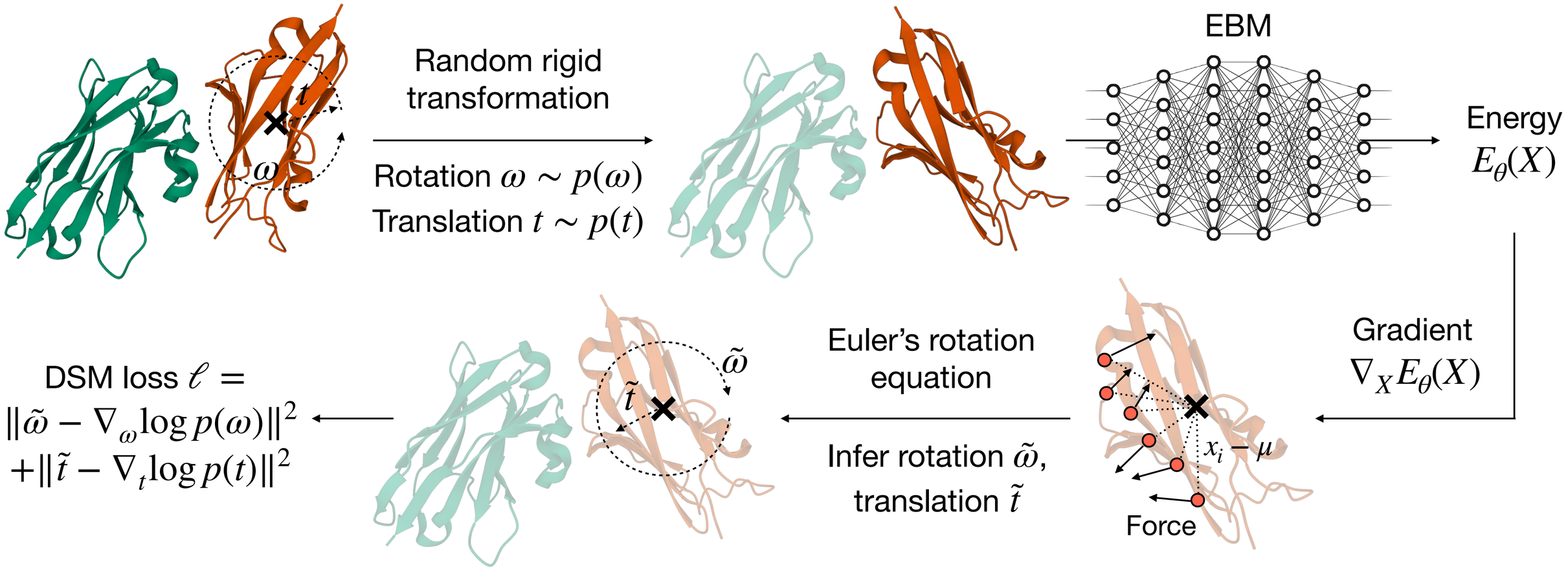
Energy-based model (EBM) architecture

Requirement: $E(X)$ is SE(3)-invariant and differentiable w.r.t. X



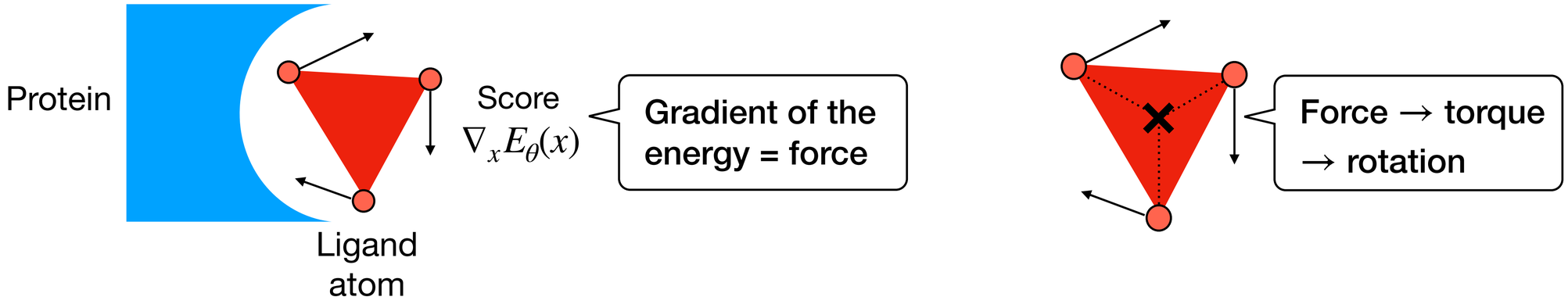
Training EBMs with SE(3) denoising score matching

Perturbing input complexes with rigid transformation noises



Neural Euler's rotation equation (NERE)

Infer rotation R from gradient $\nabla_x E_\theta(x)$ (force)

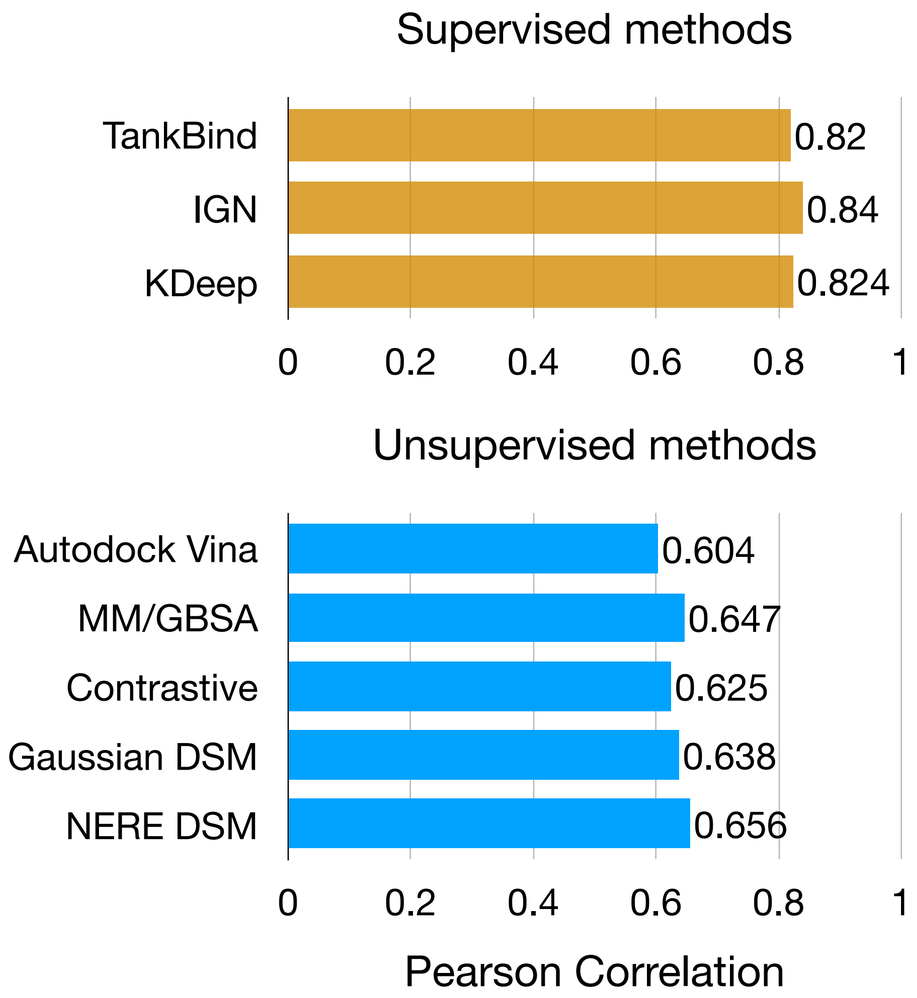


- The torque applied to the ligand $\tau = \sum_i (x_i - \mu) \times \nabla_{x_i} E_\theta(x)$
- (**Euler's rotation equation**) Angular acceleration of the ligand $\alpha = I^{-1}\tau$, where I is the inertia matrix
- Angular velocity $\omega = I^{-1}\tau\Delta t$ for an infinitesimal time Δt
- Rotation matrix R is the exponential of the following matrix $W(\omega) = \begin{pmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{pmatrix}$

Results: protein-ligand binding

Log-likelihood is strongly correlated with binding affinity

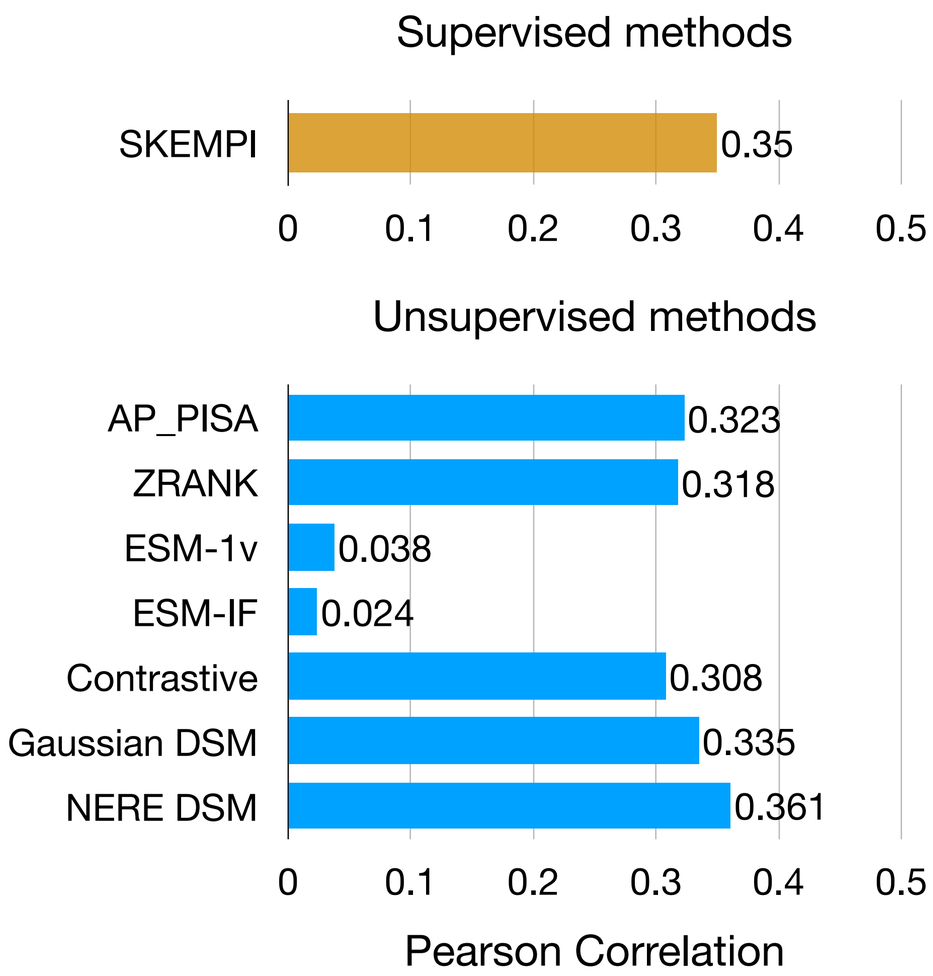
- Training set: 5237 protein-ligand complexes in PDBBind refined set (without using binding affinity data)
- Test set: 285 complexes from PDBBind (core) Measure the Pearson correlation between predicted and true affinity
- Supervised models (TankBind, IGN, KDeep) are trained on ~18000 binding affinity data in PDBBind (general subset)
- SE(3) DSM outperforms MM/GBSA and other unsupervised models like Gaussian DSM and contrastive learning



Results: antibody-antigen binding

Unsupervised models outperform supervised methods

- Training set: 3416 complexes from Structure Antibody Database (SAbDab).
- Test set: 566 complexes from SAbDab that have binding affinity labels
- We compare with physical potentials (AP_PISA, ZRANK), protein language models (ESM-IF, ESM-1v), and a supervised neural network trained on SKEMPI binding affinity data
- We outperform supervised baseline because we can leverage more unlabeled antibody-antigen complexes



Conclusion & acknowledgements

Towards unsupervised models for protein-ligand binding

Main contribution

1. Formulate binding affinity prediction as a generative modeling problem
   - Train the generative model using SE(3) denoising score matching (DSM)
2. Propose a **simple equivariant** rotation prediction module for SE(3) DSM
   - Embed Euler's rotation equation into neural networks (adding physical prior)

Acknowledgements:

Funding: Eric and Wendy Schmidt Center