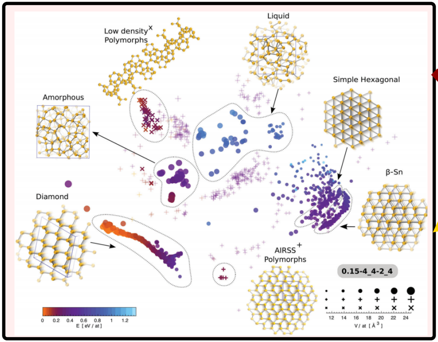
University of Basel

Deep Latent Variable Models in Materials Science

Volker Roth,
Department of Mathematics and Computer Science,
University of Basel

Machine Learning in Materials Science



~10^{60} ***Nature Insight*** **on chemical space**

Machine Learning

Properties

Energy, forces

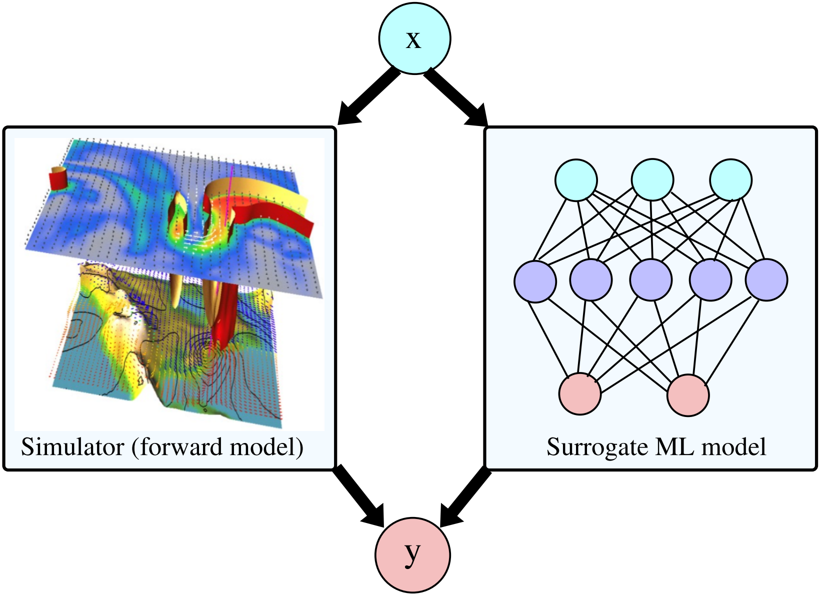
Free energy

Electrostatic, −dynamic

Magnetic moments

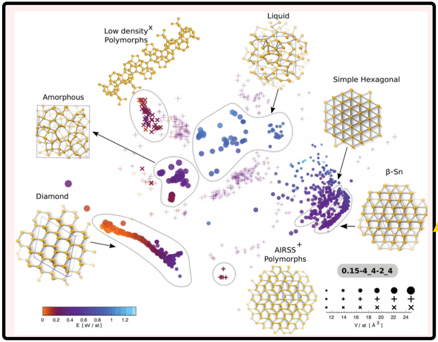
Optical properties

The Forward Path: Surrogate ML Models



The Inverse Path

- The **inverse** problem is challenging!
- We might first want to **better understand the structure** of the chemical space.
- Different views **highlighting the structure conditioned on desired properties.**



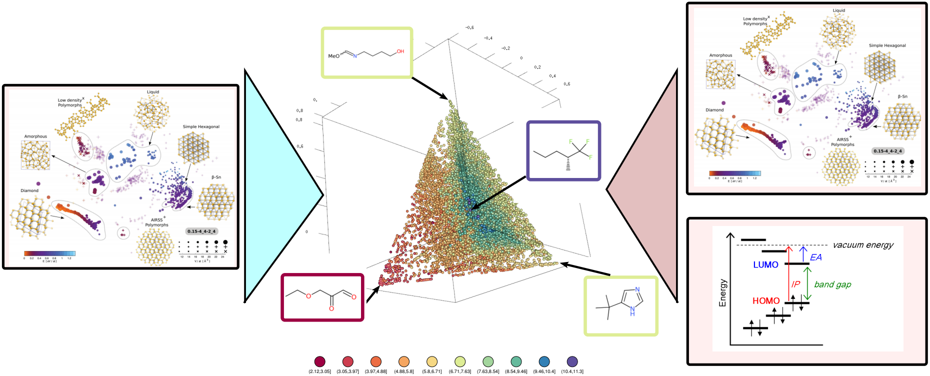
~10^{60} ***Nature Insight*** **on chemical space**

Machine Learning

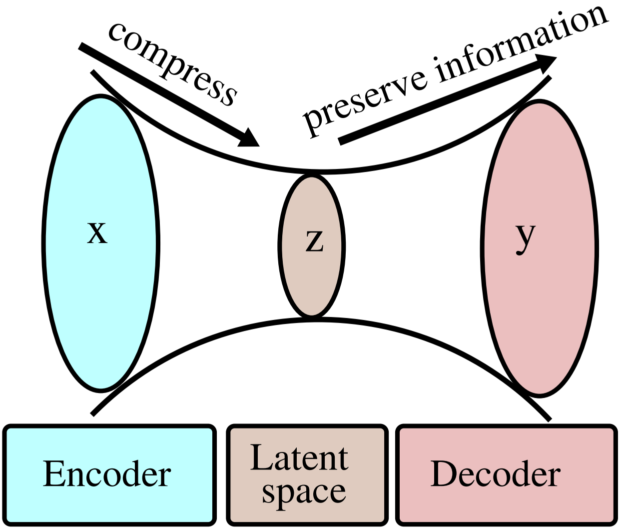
Properties

Energy, forces
Free energy
Electrostatic, −dynamic
Magnetic moments
Optical properties

Conditional Structure of Chemical Space



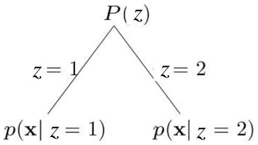
General Encoding-Decoding Scheme

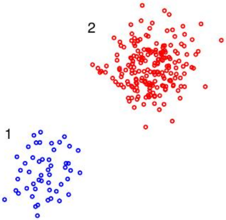


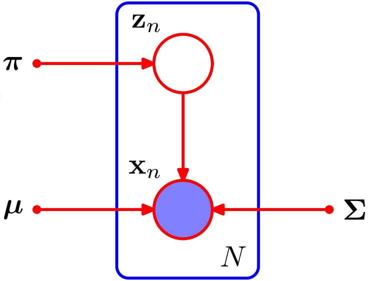
Contents

- **Linear Latent Variable Models**
  - Factor Analysis (with PCA and CCA as special cases).
  - The Information Bottleneck as a general latent variable model.
- **Nonlinear Latent Variable Models**
  - Nonlinearity through deep neural nets: Deep IB.
- **Structuring the Chemical Space**
  - Structuring the latent space.
  - Archetype analysis $\rightsquigarrow$ Deep Chemical Archetypes.

Latent Variable Models: Mixture Densities

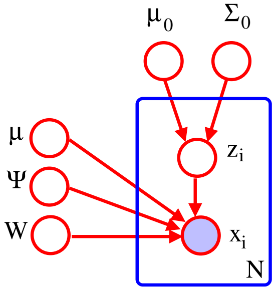






- Any data point $\boldsymbol{x}$ could have been generated in two ways; the component responsible for generating $\boldsymbol{x}$ needs to be **inferred**.
- We say, the class indicator variable z is **latent**.
- This is an example of a huge class of **latent variable models** (LVM)

Factor analysis



- One problem with mixture models: **only a single latent variable**. Each observation can only come from one of K prototypes.
- Alternative: $z_i \in \mathbb{R}^L$. Gaussian prior:

$$p(z_i) = \mathcal{N}(z_i | \mu_0, \Sigma_0)$$

$$p(x_i | z_i, \theta) = \mathcal{N}(Wz_i + \mu, \Psi),$$

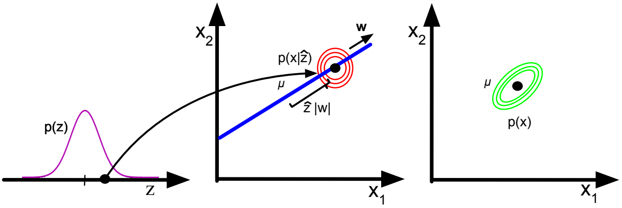
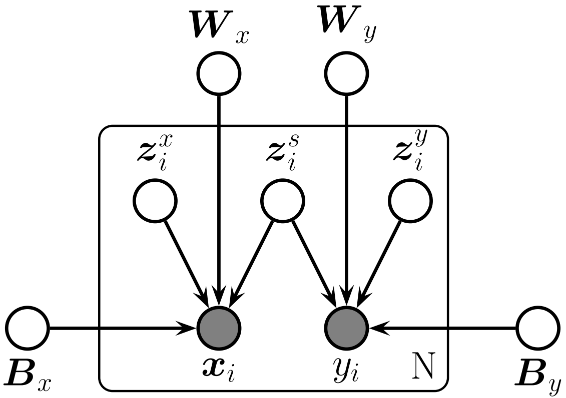


Figure 12.1 in K. Murphy

Special Cases: PCA and CCA

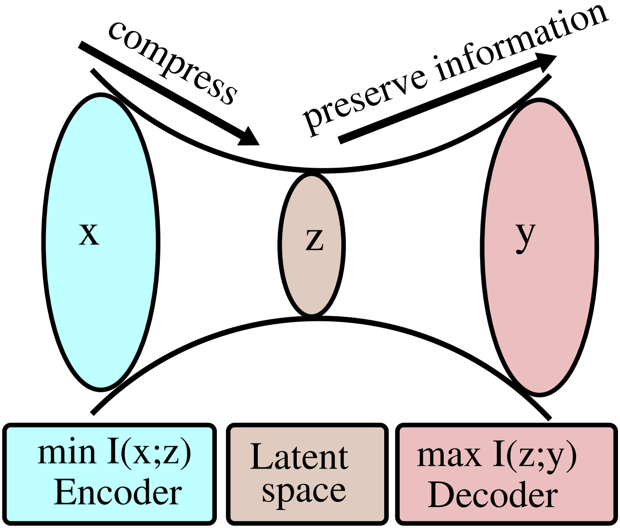
- Factor loading matrix $W = \sigma^2 I \rightsquigarrow$ (probabilistic) **PCA**.
- Multi-view version involving $\boldsymbol{x}$ and $\boldsymbol{y} \rightsquigarrow$ **CCA**.



From figure 12.19 in K. Murphy

The Information Bottleneck (Tishby et al., 1999)

FA is powerful, but still limited (Gaussian assumptions etc.). Alternatives?



Mutual Information

- A measure of **mutual dependence** between two random variables: reduction of uncertainty by knowing one variable.
- For continuous RVs:

$$\begin{aligned} I(\boldsymbol{x}; \boldsymbol{y}) &= \int\int p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) dx\, dy \\ &= D_{KL}(p(x,y) \| p(x)\, p(y)) \end{aligned}$$

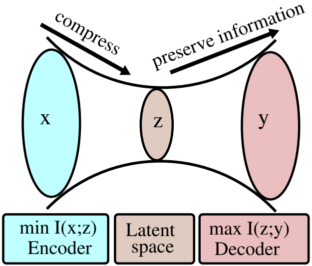
- $\boldsymbol{x}$ and $\boldsymbol{y}$ independent $\rightsquigarrow$ knowing $\boldsymbol{x}$ does not give any information about $\boldsymbol{y}$ $\rightsquigarrow$ $I(\boldsymbol{x}; \boldsymbol{y}) = 0$.

Information Bottleneck

- **The IB principle:** compress x into z, keep information about y.
- Assume y and z are conditionally independent given x an solve:

$$\min_{p(z|x)} I(x;z) - \lambda I(z;y).$$

- The original IB formulation is **not a generative model**:x, y are only used for estimating $p(x, y)$.

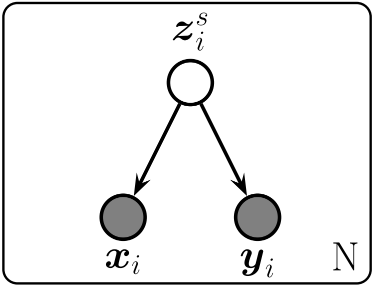


IB as a latent variable model

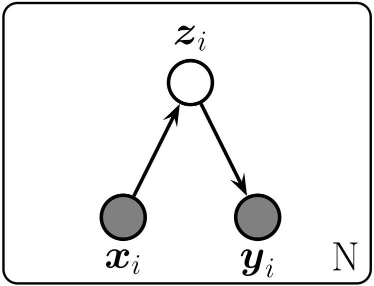
Assume $\boldsymbol{z} = f(\boldsymbol{x}) + \boldsymbol{\xi}$ captures all relevant information about $\boldsymbol{y}$.
Then $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z}) \Rightarrow \boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}|\boldsymbol{z}$

$\rightsquigarrow$ **latent version IB $^{\textbf{(lat)}}$**, basically an **asymmetric CCA model.**

CCA: $p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z})$
$\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \,|\, \boldsymbol{z}$



IB$^{\textbf{(lat)}}$: $p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{x})$
$\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \,|\, \boldsymbol{z}$



Gaussian IB (Chechnik et al. 2003))

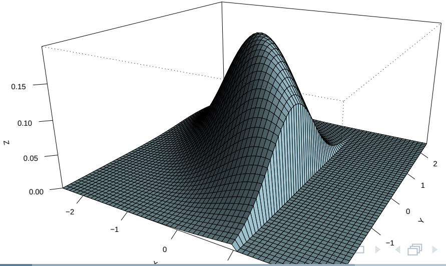
- Assume $\boldsymbol{x}$ and $\boldsymbol{y}$ are **jointly Gaussian-distributed**.

$$(x, y) \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma_x & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_y \end{pmatrix}\right),$$

- The optimal $\boldsymbol{z}$ is a **noisy projection of $\boldsymbol{x}$**:

$$z = Ax + \xi, \quad \xi \sim \mathcal{N}(0, I) \Rightarrow z|x \sim \mathcal{N}(Ax, I), \; z \sim \mathcal{N}(0, A\Sigma_x A^\top + I).$$

- **Analytic form of mutual information:**
  $I(\boldsymbol{x}; \boldsymbol{z}) = \frac{1}{2} \log |A\Sigma_x A^\top + I|,$
  $I(\boldsymbol{z}; \boldsymbol{y}) = I(\boldsymbol{x}; \boldsymbol{z}) - \frac{1}{2} \log |A\Sigma_{x|y} A^\top + I|.$



Gaussian IB as a "universal" latent variable model

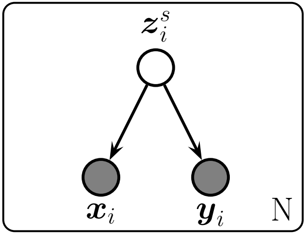
- general $\boldsymbol{y}$: rows of A are eigenvectors of $\Sigma_x^{-1}\Sigma_{x|y}$ $\rightsquigarrow$ **CCA**
- one-dimensional $\boldsymbol{y}$: $\rightsquigarrow$ **least squares regression**
- $\boldsymbol{y}$ is noisy version of $\boldsymbol{x}$: $\rightsquigarrow$ **PCA**

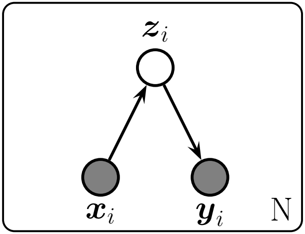
CCA: $p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z})$

$\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{z}$

IB$^{\text{(lat)}}$: $p(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{x})$

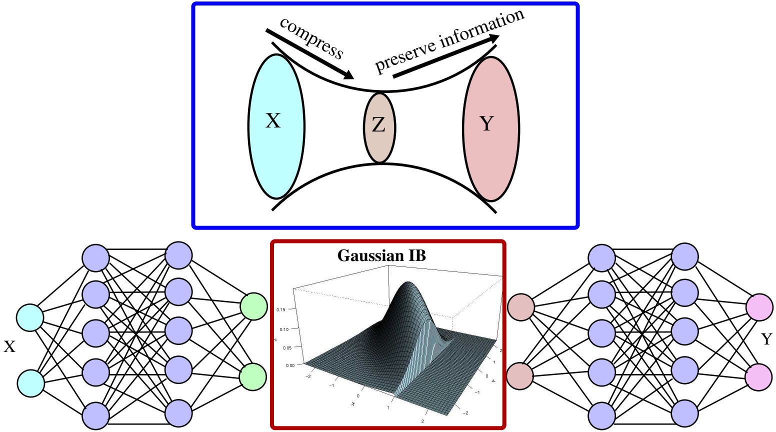
$\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y} \mid \boldsymbol{z}$





Nonlinear Latent Variable Models

- Nonlinearity through deep neural nets: Deep IB.

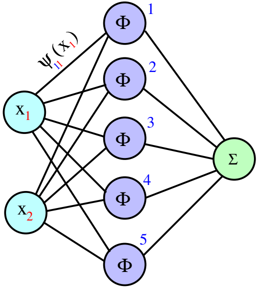


Expressive power of Neural Nets

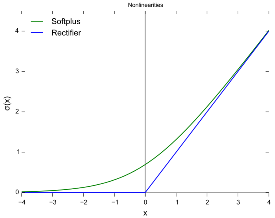
Theorem (Kolmogorov 61, Arnold 57, Lorentz 62): every continuous function on the hypercube has the form

$$f(x) = \sum_{j=1}^{2d+1} \Phi \left(\sum_{i=1}^{d} \psi_{ji}(x_i) \right),$$

for properly chosen functions Φ, ψ_{ji}.

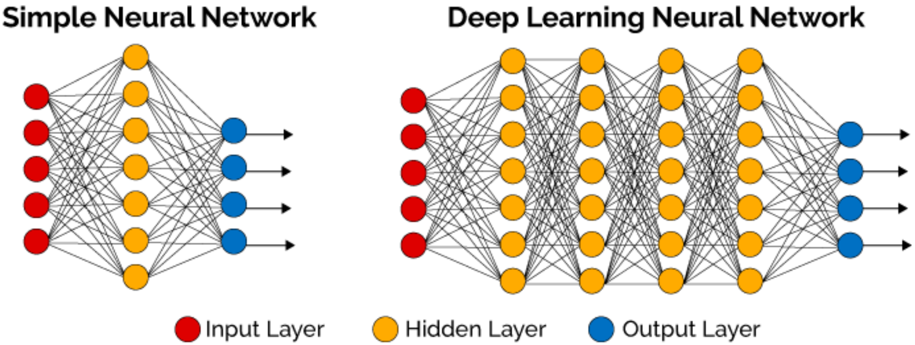


Universal function approximators can be built from "simple" neurons using only **one hidden layer** (Cybenko 89, Hornik 91,Pinkus 99).



Wikipedia

Why Deep Architectures ?



(Montufar et al, 2014): The complexity of **Deep Rectifier Models** grows **exponentially in the number of layers L** and only **polynomially in the width of the layers m.**

The deep IB

- Neural nets are trained **"end-to-end"** using **stochastic gradient descent.**
- Consider **parametric IB**, with conditionals $p_\phi(z|x)$ and $p_\theta(y|z)$.
- **Assumption:** complex joint distribution, but **simple conditionals.**

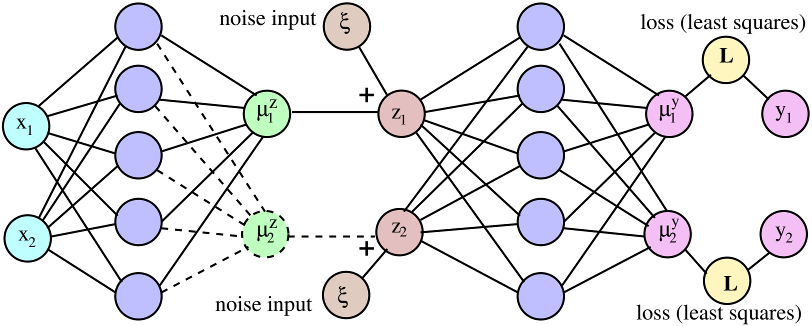
$$\boxed{\max_{\phi,\theta} -I_\phi(z;x) + \lambda I_{\phi,\theta}(z;y)}$$

$$I_\phi(z;x) \approx \frac{1}{n}\sum_i \underbrace{D_{KL}(p_\phi(z|x_i)\|p(z))}_{\text{assume analytic form available}}$$

$$I_{\phi,\theta}(z;y) \approx \frac{1}{n}\sum_i \underbrace{\log p_\theta(y_i|z_i)}_{\text{log likelihood}} + c.$$

Towards the deep IB: the decoder side

- Deep IB: $z = f(x) + \xi$, $\xi \sim \mathcal{N}(0, I)$,
  $f(x)$ implemented by deep neural net.
  $\rightsquigarrow$ **add stochastic input $\xi \sim \mathcal{N}(0, I)$.**
- This is sometimes called the **reparameterization trick.**
  ...basically just the law of transformations of random variables.

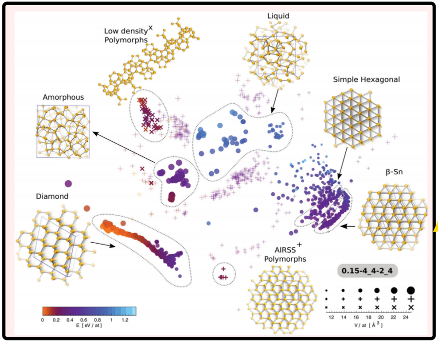


Structuring the Chemical Space

- A general problem of Deep IBs
  $\rightsquigarrow$ need more structure in the latent space.
- Solution: archetype analysis $\rightsquigarrow$ Deep Chemical Archetypes.

The Inverse Path

- The **inverse** problem is challenging!
- We might first want to **better understand the structure** of the chemical space.
- Different views **highlighting the structure conditioned on desired properties.**



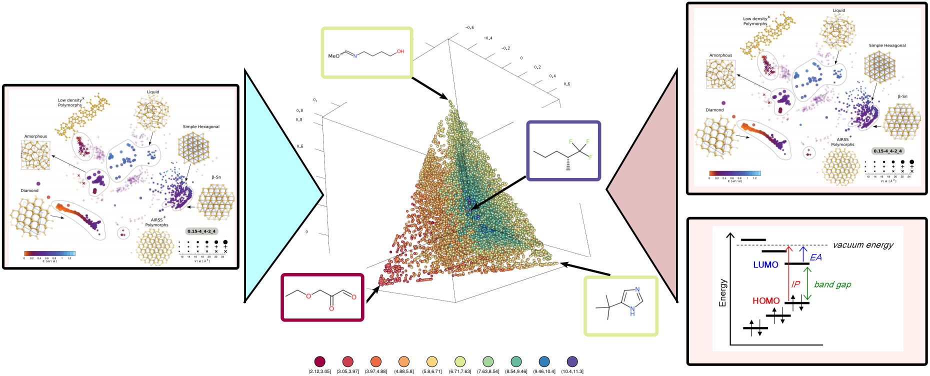
~10^{60} ***Nature Insight*** **on chemical space**

Machine Learning

Properties

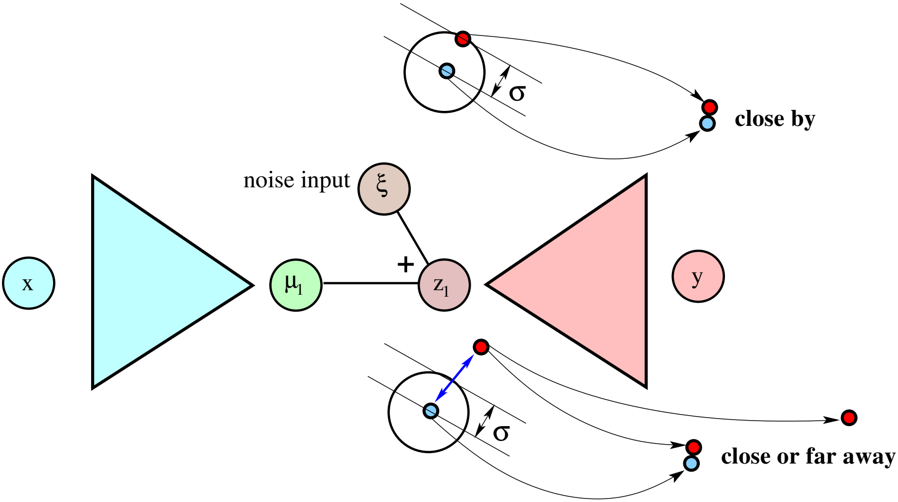
Energy, forces
Free energy
Electrostatic, −dynamic
Magnetic moments
Optical properties

Conditional Structure of Chemical Space



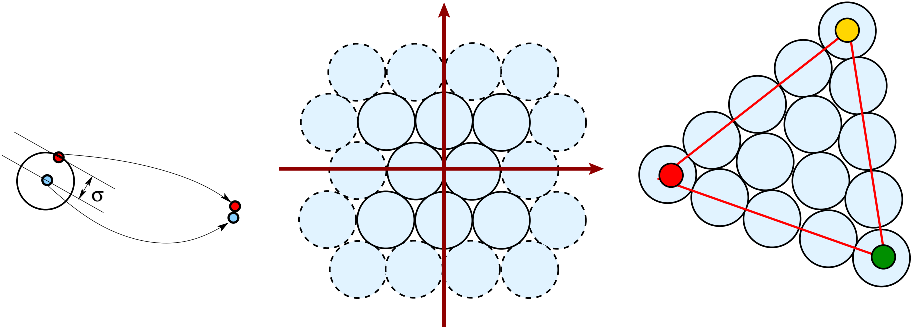
One problem with Autoencoders/IBs

Local similarity in the latent space is translated to **local** similarity in the output space...but no "global" structure.



Archetypes

Idea: enforce structure in the latent space. Objects must be **convex mixtures** of "extreme" objects ⇝ **archetypes.**



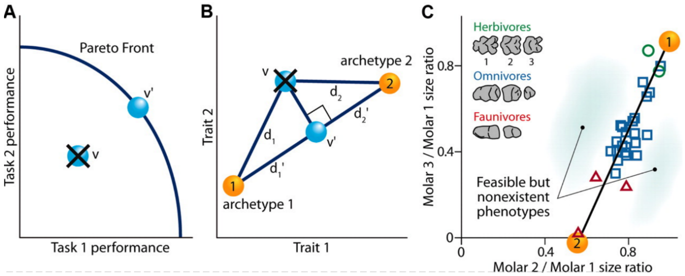
Sciencexpress

published in 2012

Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space

O. Shoval,[1] H. Sheftel,[1] G. Shinar,[1] Y. Hart,[1] O. Ramote,[1] A. Mayo,[1] E. Dekel,[1] K. Kavanagh,[2] U. Alon[1*]

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel.



Archetypes and Evolutionary Trade-offs

- In a biological system, a **phenotype** is defined by a vector of **traits** (quantitative measurements)
- Space of phenotypes: **morphospace**
- **Natural selection:** optimize fitness function $\rightsquigarrow$ point in morphospace.
- **But** organisms need to perform **multiple tasks** that all contribute to their fitness $\rightsquigarrow$ **multi-objective** optimization problem.
- **Pareto front**: best trade-offs between different requirements.
- Point on Pareto front depends on **relative contribution of tasks** to fitness.

Computational Archetype Selection

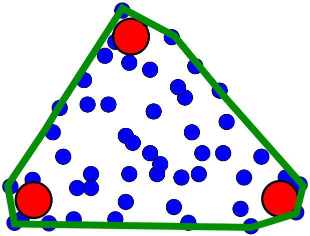
Cutler &Breiman, *Archetypal Analysis*, Technometrics 1994.

- n observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^p$, as rows of data matrix $X \in \mathbb{R}^{n \times p}$
- Aim: find K archetypes $\Rightarrow Z \in \mathbb{R}^{K \times p}$; $K \ll n$ fixed.
- **Observations are convex mixtures of archetypes:**

$$\boldsymbol{x}_i = Z^t \boldsymbol{a}_i + \epsilon_i, \quad a_{ij} \geq 0 \text{ and } \sum_{j=1}^{K} a_{ij} = 1.$$

- **Archetypes are convex mixtures of observations:**

$$\boldsymbol{z}_i = \sum_{j=1}^{n} b_{ij} \boldsymbol{x}_j, \quad \text{where } b_{ij} \geq 0 \text{ and } \sum_{j=1}^{n} b_{ij} = 1$$

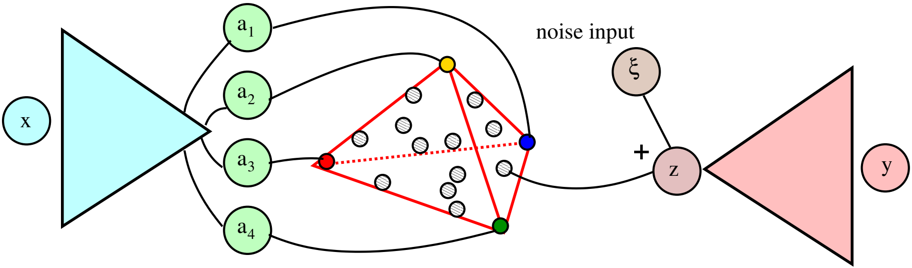


Archetypes approximate convex hull

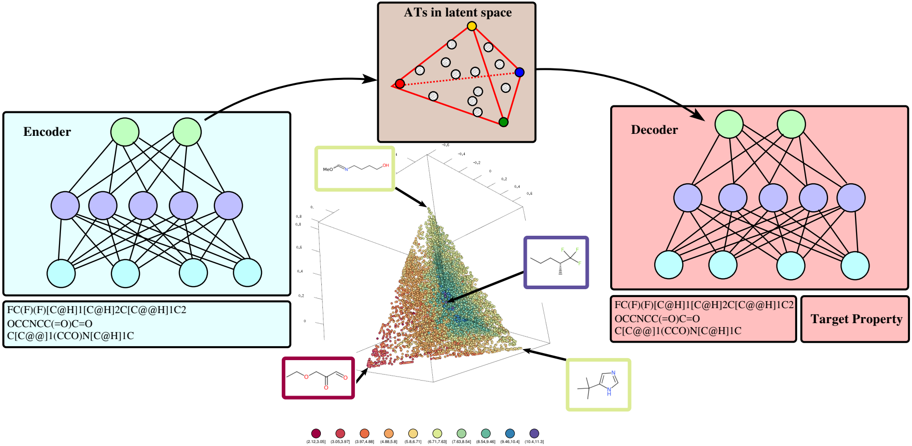
- Constrained optimization problem.

Deep Archetypes (Keller et al. 2018)

Problem: it is difficult to **find a representation where convex mixing works**...
Solution: fix the ATs at vertices of a simplex located in the **latent space of an IB** and use deep nets to **learn such a representation.**

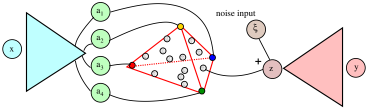


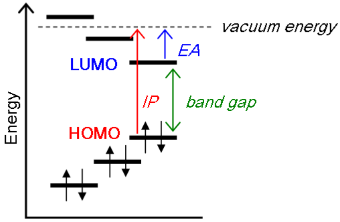
Deep Chemical Archetypes



Deep Chemical Archetypes

- **Input** x**:** SMILES strings and **3D molecule descriptors depending on atom positions** $\rightsquigarrow$ **conformational information.**
- **Target property:** energy difference between highest occupied molecular orbital and lowest unoccupied molecular orbital, **HOMO-LUMO gap.**
- **Deep SMILES** decoder, producing syntactically correct SMILES (O'Boyle & Dalke, 2018).





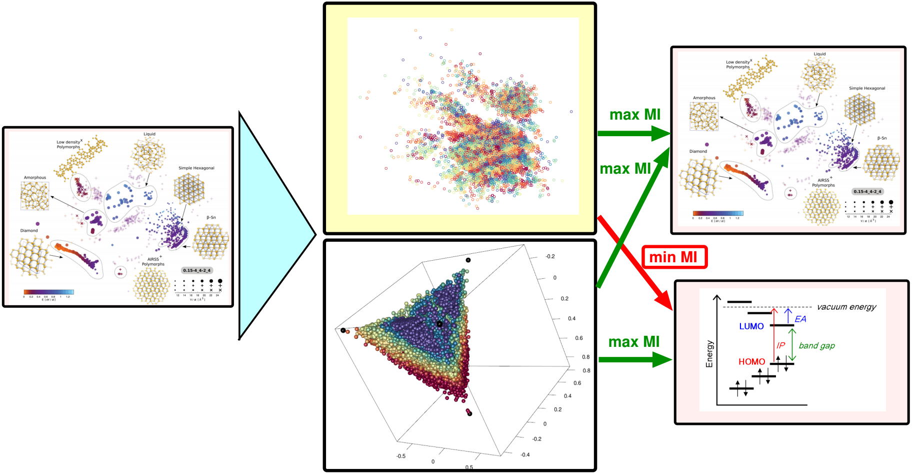
Wikipedia

Deep Chemical Archetypes: Encoding Invariances

Problem: many molecules have (roughly) the same homo-lumo gap!

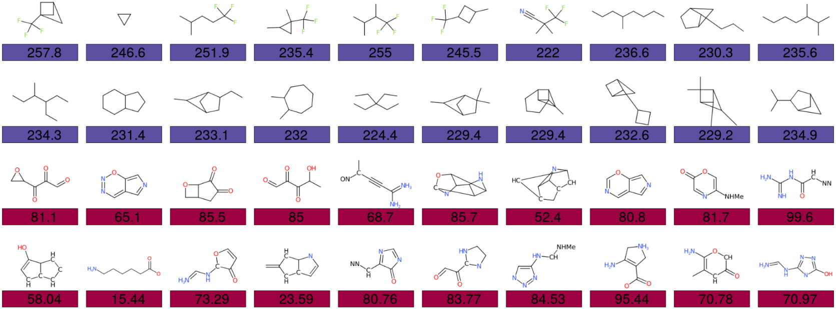
$\rightsquigarrow$ need more latent dimensions to capture structural variations

$\rightsquigarrow$ "orthogonal" space $\rightsquigarrow$ sample molecules with a given property!



Deep Chemical Archetypes: Results on QM9

- 13k organic molecules made up of H, C, N, O and F, with up to nine heavy (non-hydrogen) atoms. Properties calculated by DFT.



- Some potentially interesting molecules found via sampling the "orthogonal" space (more on that will appear elsewhere...)

Summary

- **Chemical space** $\rightsquigarrow$ challenging ML problems.
- Approach: visualize structure **conditioned on target properties**.
- **Deep information bottleneck** models are powerful tools for this purpose!
- **Generative model** allows us to **sample molecules** with desired properties.
- But still **many open questions:** large parts of the chemical space seem to be empty, transfer of models not trivial at all.

Thank you for your attention!

MARVEL
NATIONAL CENTRE OF COMPETENCE IN RESEARCH

FNSNF
SWISS NATIONAL SCIENCE FOUNDATION

75 NRP
Big Data
National Research Programme