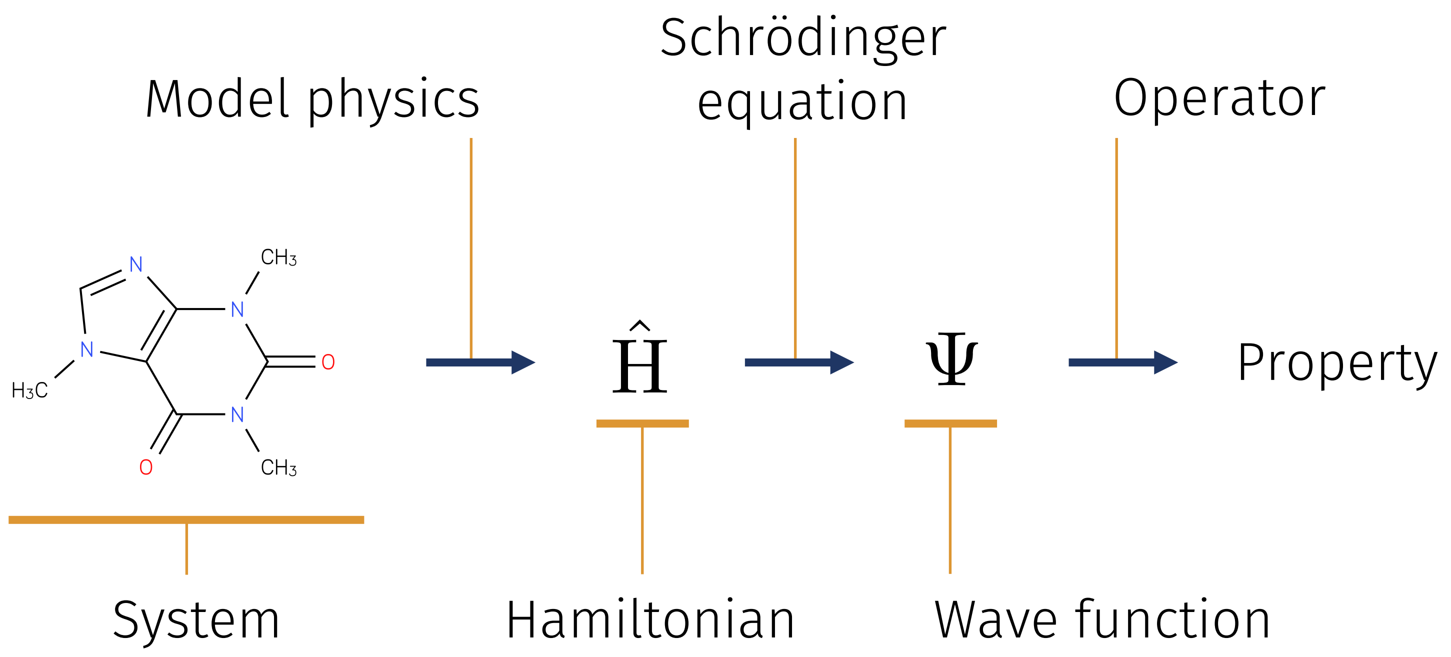
Machine Learning and Quantum Alchemy

Guido Falk von Rudorff, University of Kassel

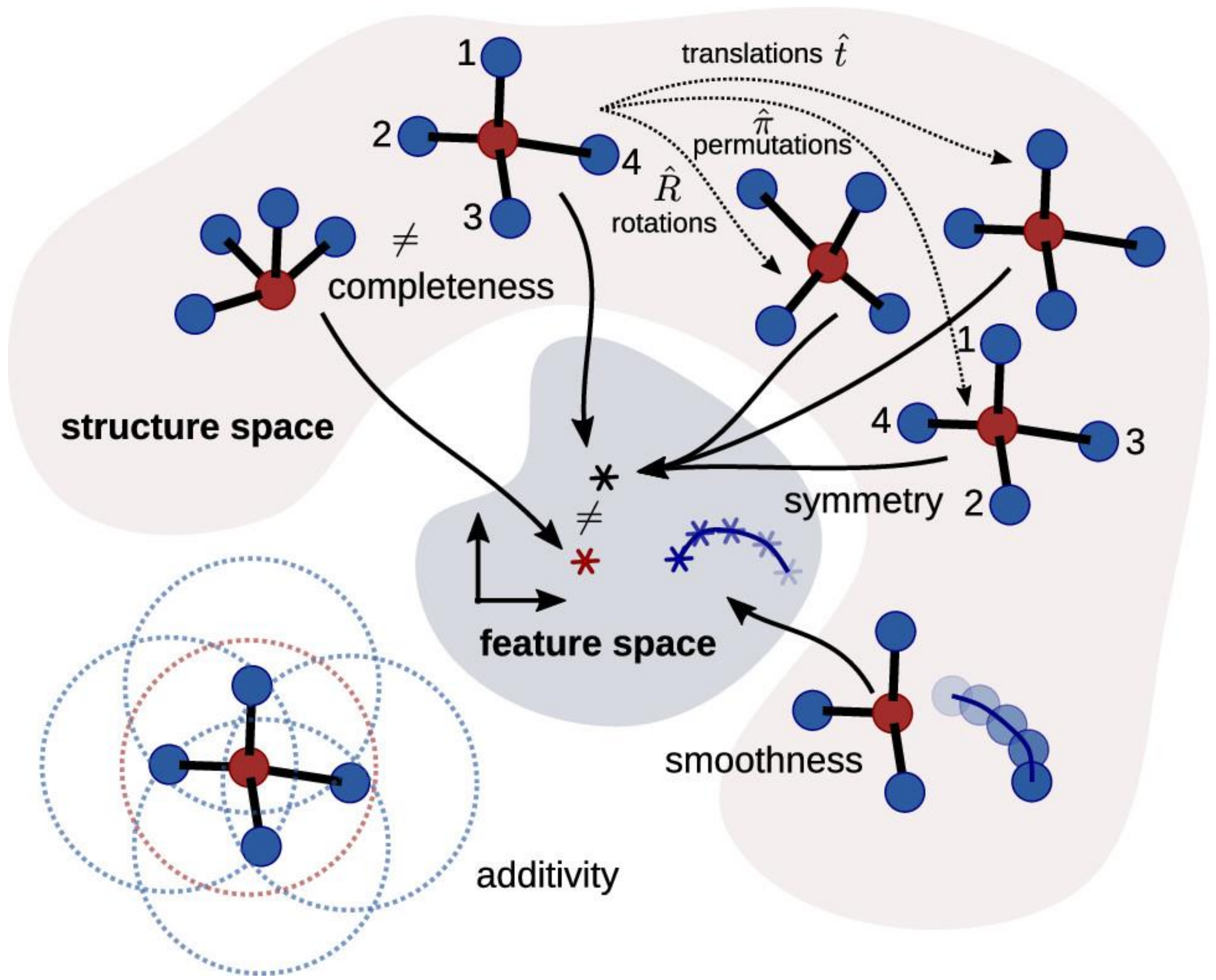
vonrudorff@uni-kassel.de nablachem.org/talks ferchault @ferchault

Which problem do we try to solve?

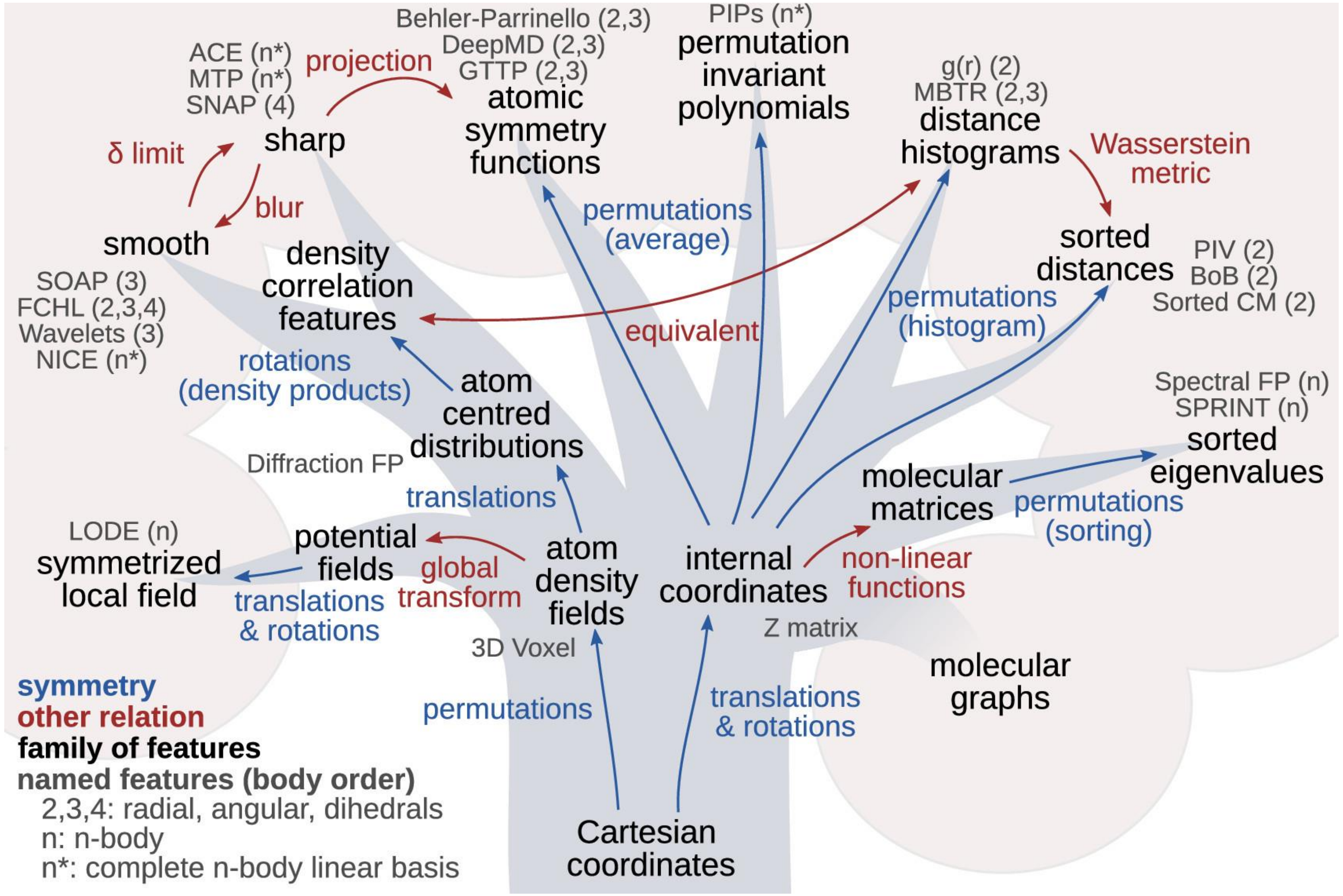


Representations

Requirements



Overview


Behler-Parrinello (2,3)
DeepMD (2,3)
GTTP (2,3)
atomic symmetry functions
PIPs (n*)
permutation invariant polynomials
ACE (n*)
MTP (n*)
SNAP (4)
projection
sharp
δ limit
blur
smooth
g(r) (2)
MBTR (2,3)
distance histograms
Wasserstein metric
sorted distances
PIV (2)
BoB (2)
Sorted CM (2)
permutations (average)
SOAP (3)
FCHL (2,3,4)
Wavelets (3)
NICE (n*)
density correlation features
equivalent
permutations (histogram)
rotations (density products)
atom centred distributions
Spectral FP (n)
SPRINT (n)
sorted eigenvalues
molecular matrices
permutations (sorting)
Diffraction FP
translations
LODE (n)
symmetrized local field
potential fields
global transform
atom density fields
internal coordinates
non-linear functions
translations & rotations
3D Voxel
Z matrix
molecular graphs
permutations
translations & rotations
Cartesian coordinates
symmetry
other relation
family of features
named features (body order)
2,3,4: radial, angular, dihedrals
n: n-body
n*: complete n-body linear basis

Categorial: One-Hot

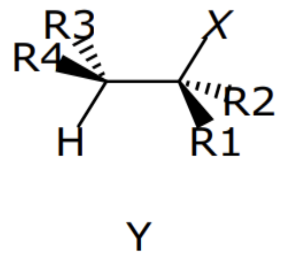
Categorial data vs regression

Solution: „binary“, „dummy“, „one-hot“ encoding

Encode *n* categories as vector of length *n-1* with one category (arbitrary) being the null vector.

Example:
- A: (0, 0)
- B: (1, 0)
- C: (0, 1)

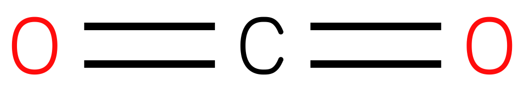
Chemistry example: (ABBA|CD) -> (0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0|0,1,0,0,1)



	A	B	C	D	E
R*k*	H	NO_2	CN	CH_3	NH_2
X	F	Cl	Br		
Y	H	F	Cl	Br	

Graph-based

Often: molecules = well-defined bonds



Adjacency matrix
- 1 if atoms i and j are bonded
- 0 otherwise

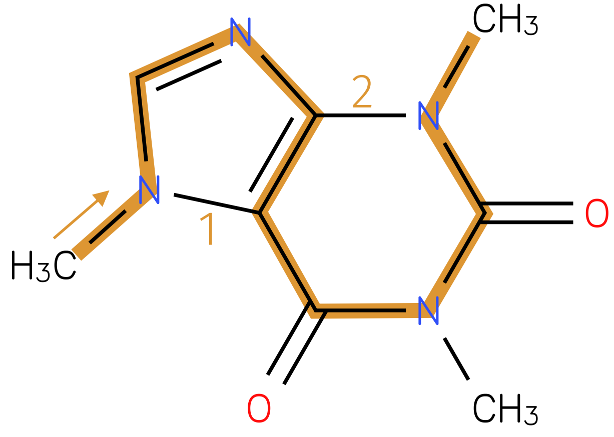
$$\begin{matrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{matrix}$$

Bond order matrix
- Bond order if atoms i and j are bonded
- 0 otherwise

$$\begin{matrix} 0 & 2 & 0 \\ 2 & 0 & 2 \\ 0 & 2 & 0 \end{matrix}$$

SMILES

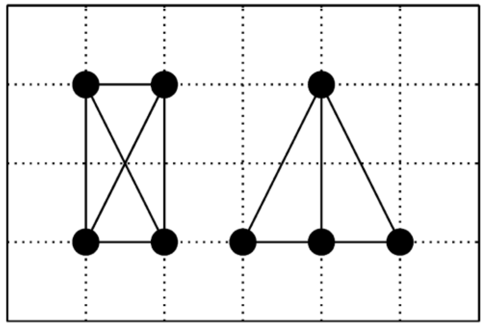
- Bonds: Nothing (1), = (2), # (3), \$ (4)	O=C=O
- Partial charges	[Na+]
- Fragments: .	[Na+].[Cl-]
- Rings: labels	C1CCCCC1
- Branches: parentheses	

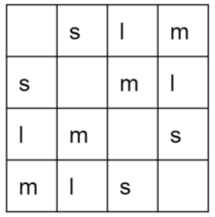


CN1C=NC2=C1C(=O)N(C)C(=O)N2C

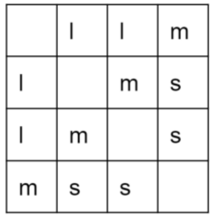
Distance-based

Many-body descriptions, e.g. all pairwise distances





	s	l	m
s		m	l
l	m		s
m	l	s	



	l	l	m
l		m	s
l	m		s
m	s	s	

Even three- and four-body interactions not unique [1].

Chemistry: **Coulomb Matrix [2]**

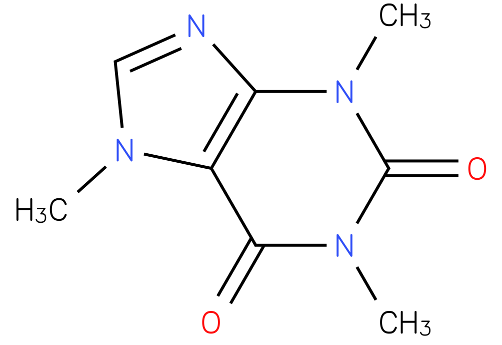
- Diagonal: $0.5Z_i^{2.4}$
- Off-diagonal: $Z_iZ_j/\|\mathbf{R}_i - \mathbf{R}_j\|$
- Problem: sorting, uniqueness

[1] Pozdnyakov et al., *Phys Rev Lett* 2020. [2] M. Rupp et al., *Phys Rev Lett* 2012.

Fingerprints

Collect relevant features of a molecule into a vector.

- Fixed list of functional groups (e.g. Joback method)
- Hash of generated circular atom environments (e.g. ECFP)
- Kinds of generated local environments (e.g. Morgan)
    - Element, # heavy neighbors, # protons, charge, part of ring…
- Fixed checklist of features (e.g. MACCS keys)
    - Subgraphs



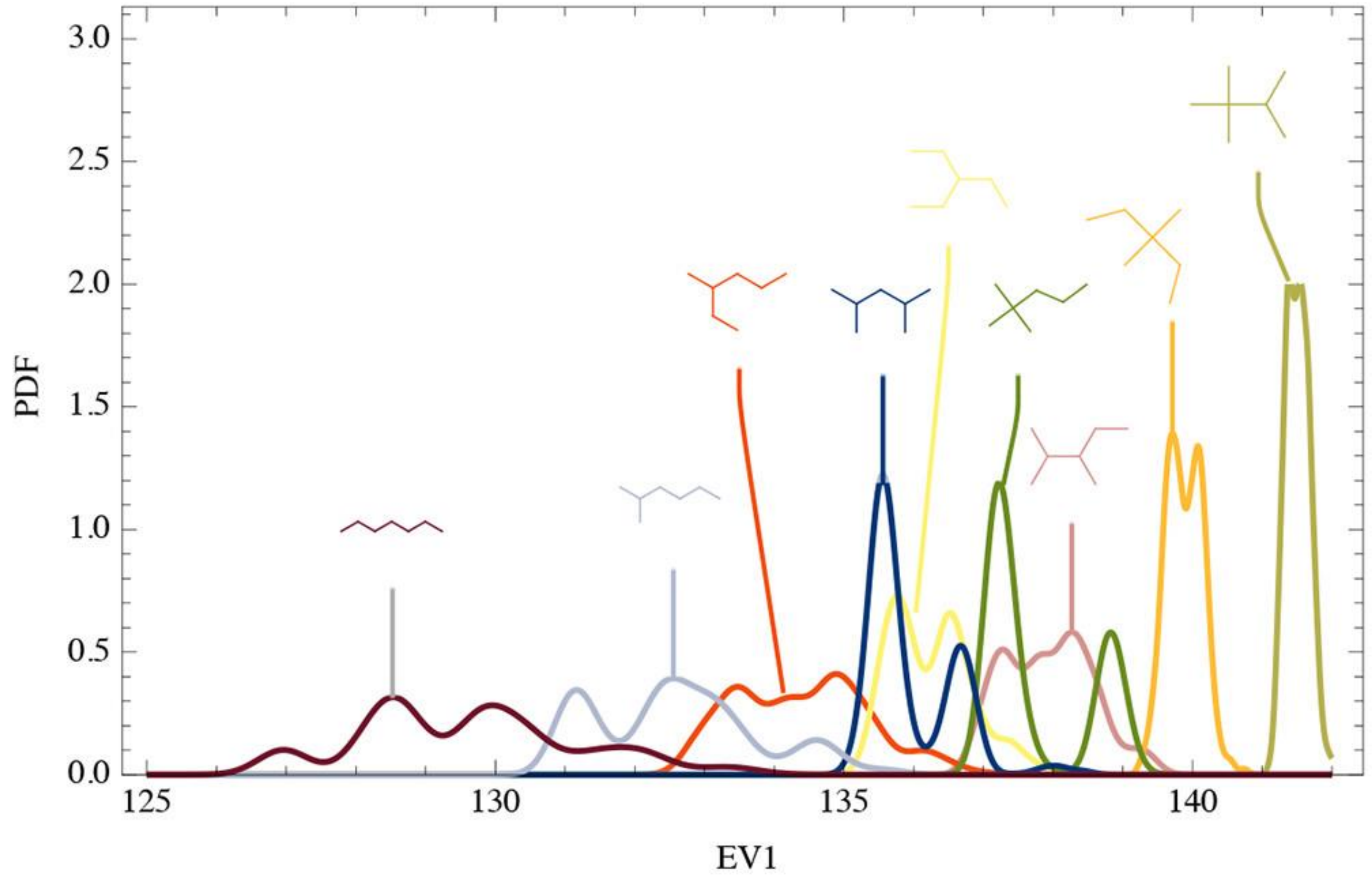
Morgan (feature: count)
10565946: 2, 348155210: 1, 476388586: 1, 540046244: 1, 553412256: 1, 864942730: 2, 909857231: 1, 1100037548: 1, 1333761024: 1, 1512818157: 1, 1981181107: 1, 2030573601: 1, 2041434490: 1, 2092489639: 3, 2246728737: 3, 2370996728: 1, 2877515035: 1, 2971716993: 1, 2975126068: 2, 3140581776: 1, 3217380708: 4, 3218693969: 1, 3462333187: 1, 3657471097: 3, 3796970912: 1

MACCS keys:
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0,
0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0]

Spectra

Condense a matrix into its eigenvalues

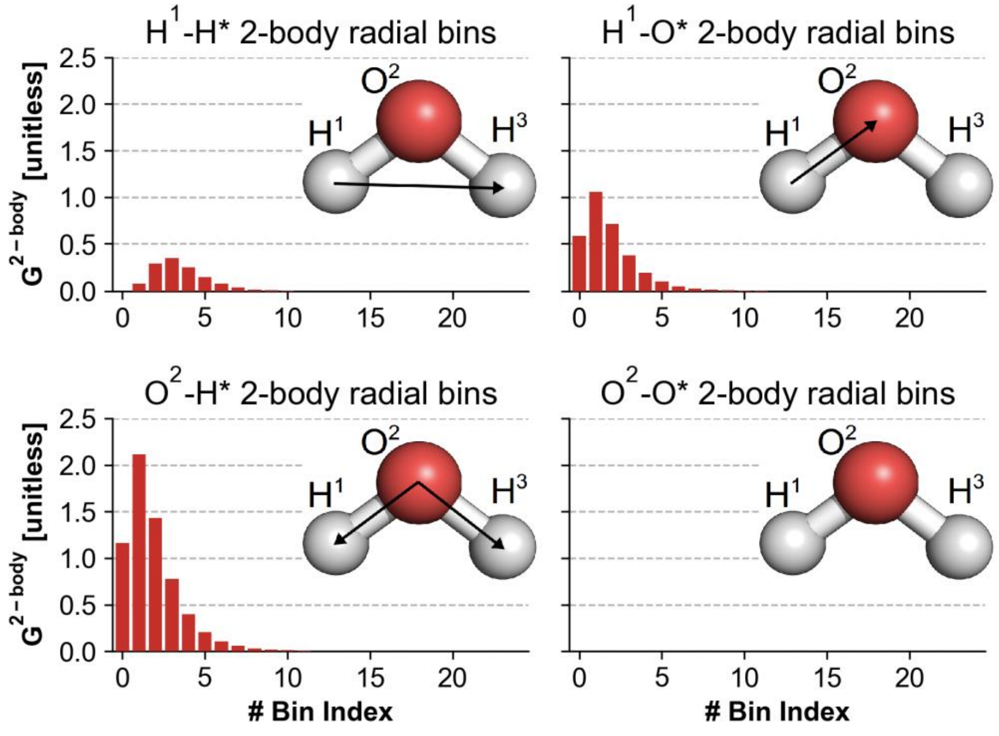
- No sorting issue, as permutationally invariant
- Smaller: N instead of N^2
- Lossy

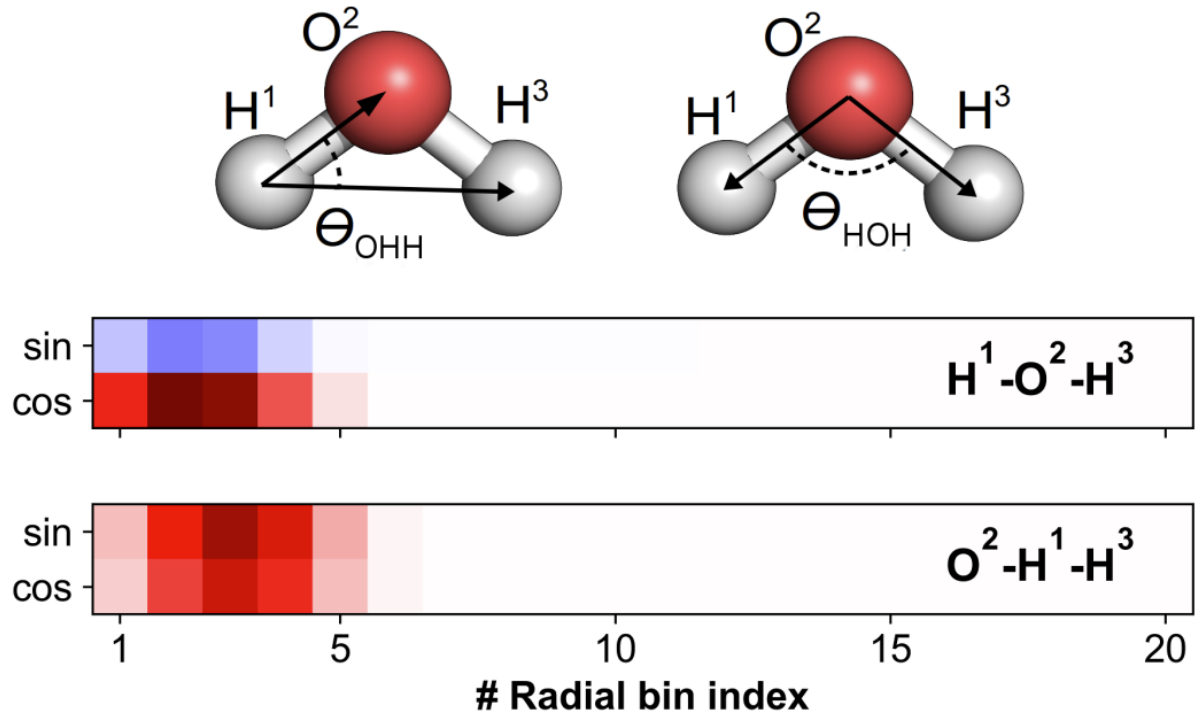


J. Schrier, *J Chem Inf Model*, 2020.

Densities

Smear positions into densities (FCHL / SOAP)





Christensen et al, *J Chem Phys*, 2020.

Kernel-Ridge-Regression

Core idea

Procedure

- Get *i* data points with scalar property (label) $\{q_i\}$
  - E.g. atomisation energy
- Calculate all representations $\{\mathbf{M}_i\}$
  - typically ~1k
- Find distance and kernel matrices $\mathbf{D}, \mathbf{K}$
  - Symmetric
- Train model for predictions $\{\tilde{q}_i\}$

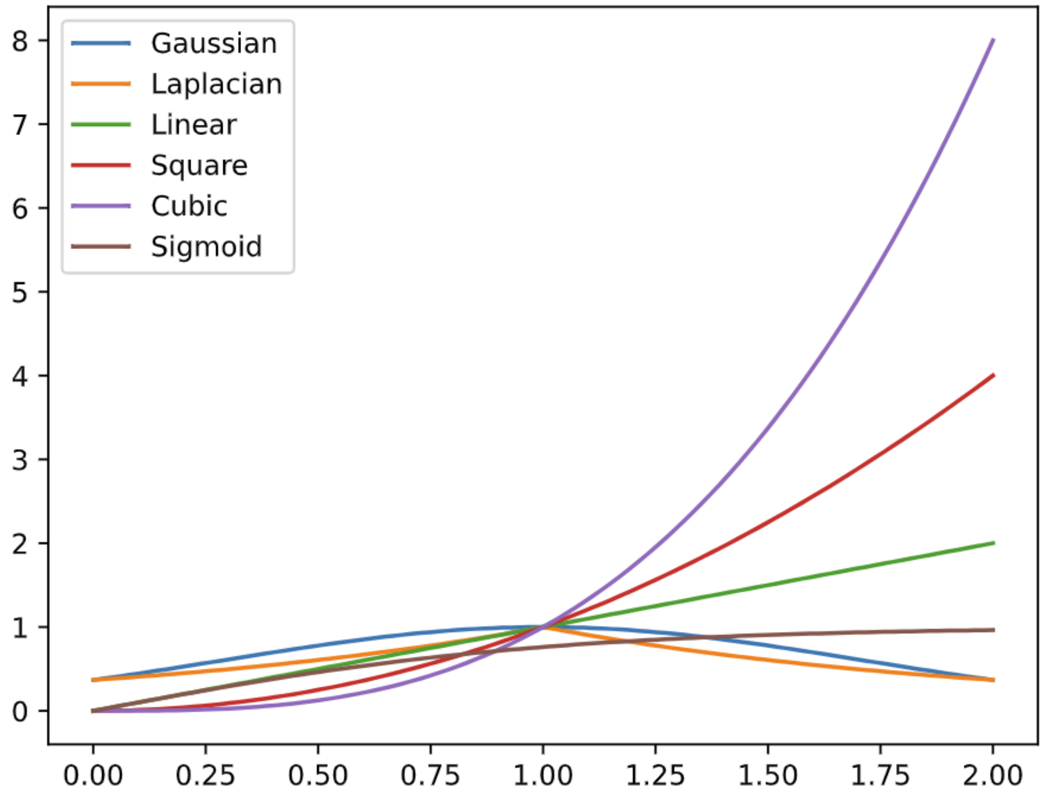
$$\arg\min_{\alpha} \sum_i (q_i - \tilde{q}_i)^2 + \lambda \sum_{ij} \alpha_i \alpha_j k_{ij}$$

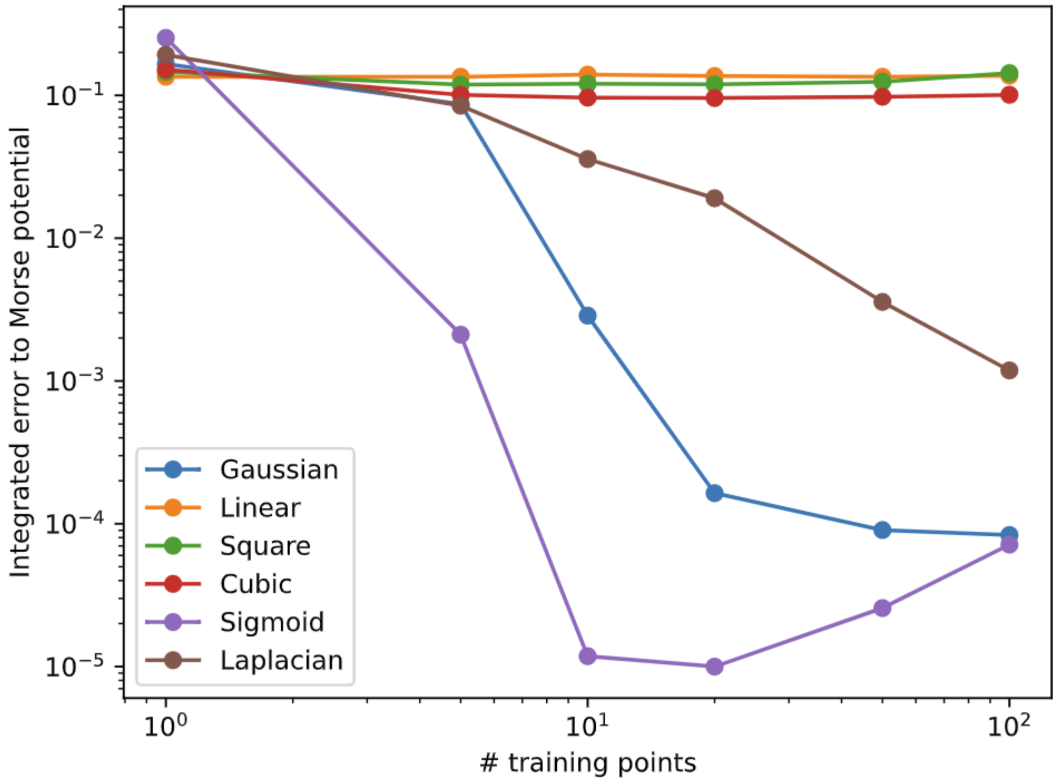
$$\Rightarrow \alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \text{y}$$

$$\tilde{q}(\mathbf{M}) = \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i)$$

Gaussian kernel: $k(x, y) = \exp(-\gamma \|x - y\|^2)$

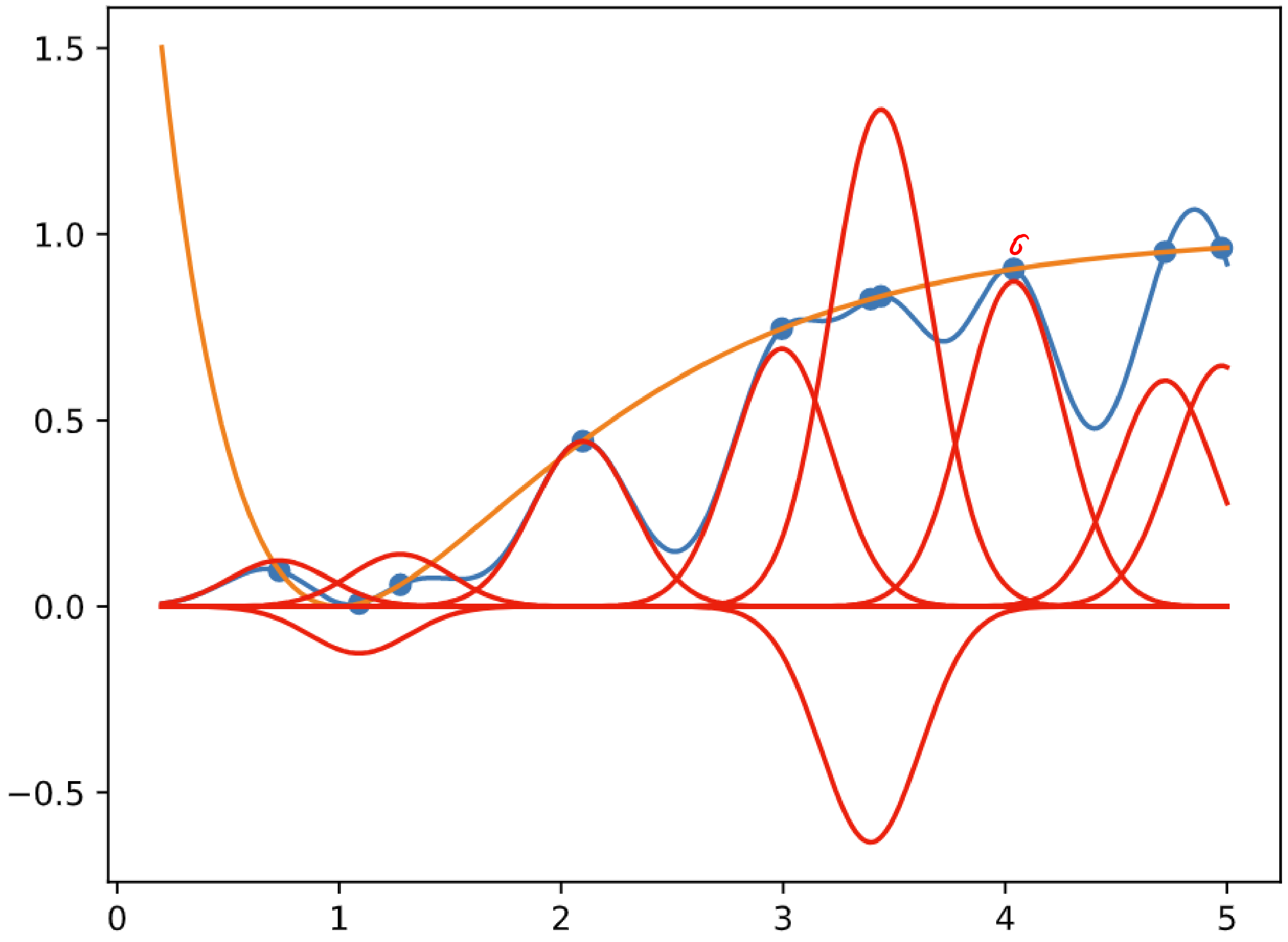
Kernel functions / Mercer‘s condition

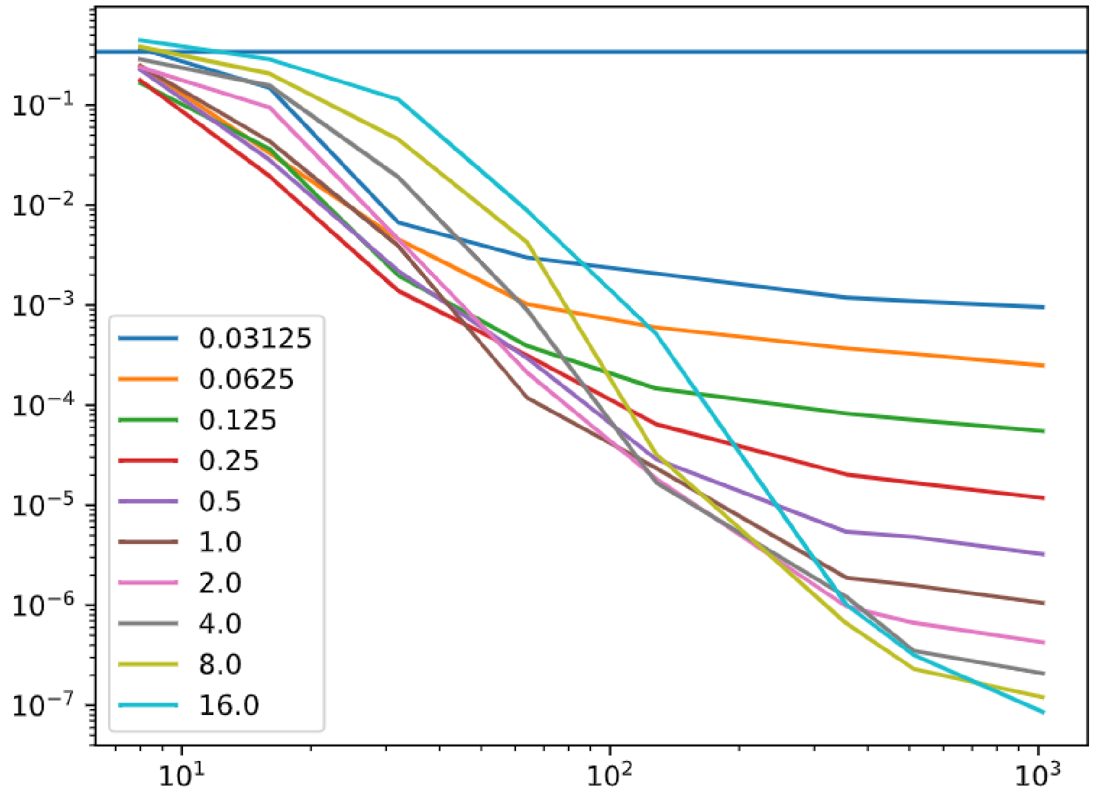




$$\iint g(x)K(x,y)g(y)\,dx\,dy \geq 0$$

Example model





When to use

Pros
- One-Step learning
- Explainable model
- Easy ablatation studies (=„what if certain data was not known“)
- Efficient cross-validation

Cons	How to fix
- Need to have features	Make them overcomplete
- Memory requirements can be challenging	Consider nearest neighbors only
- Plenty of hyperparameters	More compute
- Instable for large data sets	Regularize

Summary Kernel Ridge Regression

- Explainable method
- Requires kernel function
- Can be learned without optimization
- Requires hyperparameter scans
- Highly susceptible to kernel function choice
- Regularization required especially for noisy data
- Easy to implement, hard to implement efficiently