Molecular shape descriptors

… parametrization of a shape

How to **describe a shape** (of a ligand … **cluster**)

with **a few real numbers**?

With a **vector**? (fixed dimension)

Such that

Similarity (sh1 , sh2) ~ $||d_{sh1} - d_{sh2}||$

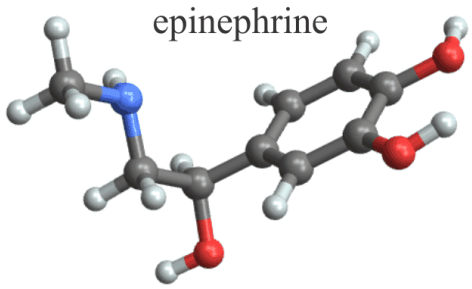
For

- Quick testing of similarity (**screening**),
- **Clustering** in space of shapes/properties,
- **Lossy compression** of databases …

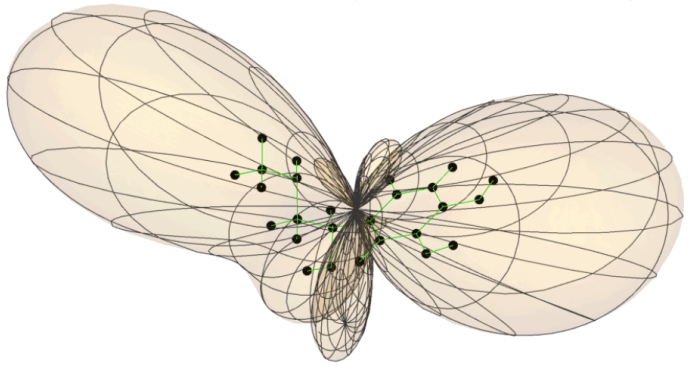
Einstein: “Everything should be made as simple as possible,

but not simpler.”

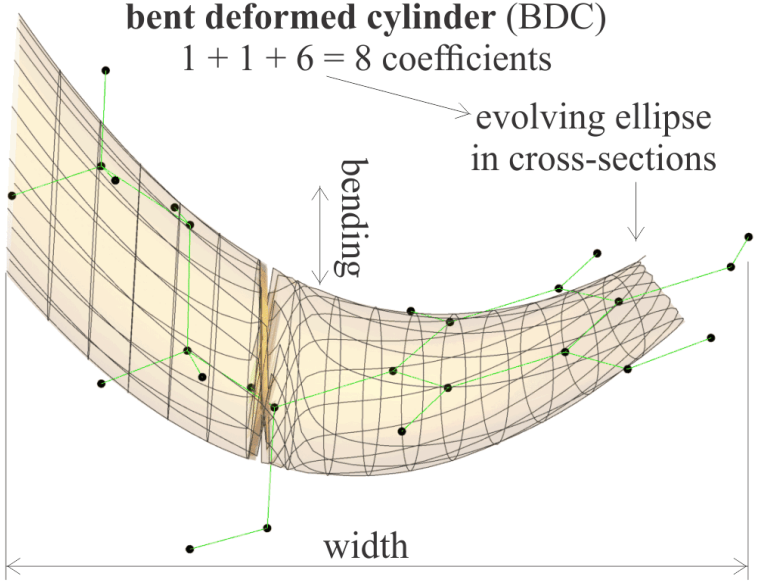
Jarek Duda, Kraków, TFML 16.II.2017



spherical harmonics $l = 0,1,2,3$, $m = -l \,..\, l$
16 coefficients (3 vanish + 3 small)



bent deformed cylinder (BDC)
1 + 1 + 6 = 8 coefficients

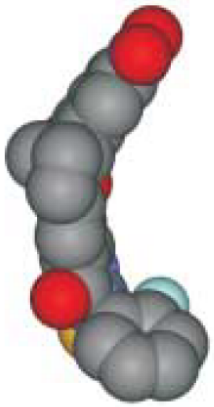


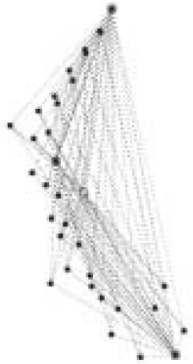
Ultrafast shape recognition (USR, 2007, Ballester, Richards)

12 numbers – 3 moments of distance from 4 points:

the centroid (ctd), the closest atom to ctd (cst),

the farthest atom to cst (fct) and the farthest atom to fct (ftf).





$$\left\{d_k^{\mathrm{ctd}}\right\}_{k=1}^{N}$$

$$\left\{d_k^{\mathrm{cst}}\right\}_{k=1}^{N}$$

$$\left\{d_k^{\mathrm{fct}}\right\}_{k=1}^{N}$$

$$\left\{d_k^{\mathrm{ftf}}\right\}_{k=1}^{N}$$

$$\vec{M}^q = \begin{pmatrix} \mu_1^{\mathrm{ctd}} \\ \mu_2^{\mathrm{ctd}} \\ \mu_3^{\mathrm{ctd}} \\ \mu_1^{\mathrm{cst}} \\ \mu_2^{\mathrm{cst}} \\ \mu_3^{\mathrm{cst}} \\ \mu_1^{\mathrm{fct}} \\ \mu_2^{\mathrm{fct}} \\ \mu_3^{\mathrm{fct}} \\ \mu_1^{\mathrm{ftf}} \\ \mu_2^{\mathrm{ftf}} \\ \mu_3^{\mathrm{ftf}} \end{pmatrix}$$

What should we expect from shape descriptors:

1) **representativeness** - close fingerprints should correspond to similar molecules, or in other words: distant molecules to distant fingerprints (**no false positives**),

2) **continuity** - small perturbation of molecule should not lead to a large change of its fingerprint (**no false negatives**),

3) **selectiveness** - for performance reasons we would like the vectors to be as short as possible, maximally exploit all the used coefficients, so they

should represent only the most significant features which might be meaningful for the interesting process like ligand bonding,

4) **independence** - analogously, the coefficient should not be correlated,

5) **decodability** - the fingerprint should allow to reconstruct the used approximation of shape – **can be treated as its lossy compression**,

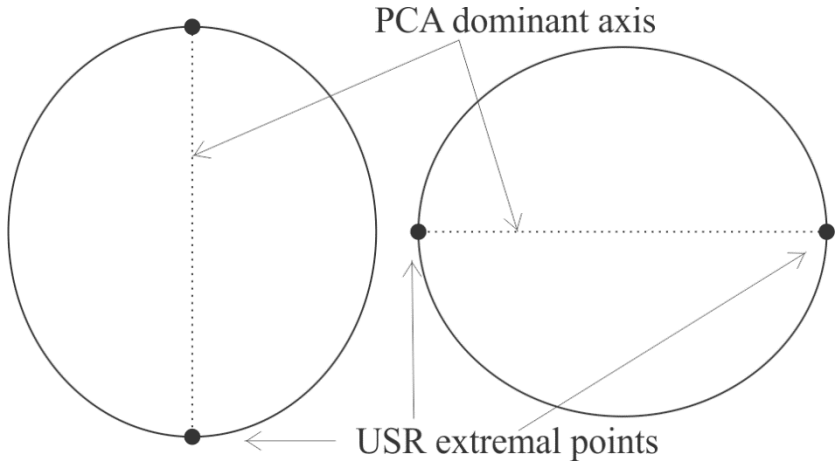
6) **faithfulness** - if decodable, the approximation **should agree** with essential qualitative and quantitative properties of the molecule, **should not introduce artifacts**.

USR fulfills none of them!

Continuity problem:

Small perturbation can switch the anchor of descriptor

(in rare symmetric cases)



General solution – **smoothen discontinuity**:

If close to a symmetric situation, find descriptors for both choices and use their average

Continuity at cost of reduced **decodability**

These requirements usually have disjoint applicability:

virtual screening **lossy compression**

general remark for **continuity**:

While **optimizing some quality measure**, we need to make sure that

small perturbation will not change the attractor

(real) **Spherical harmonics (SH)**

(Real: $e^{im\varphi} \rightarrow \sin(m\varphi), \cos(m\varphi)$

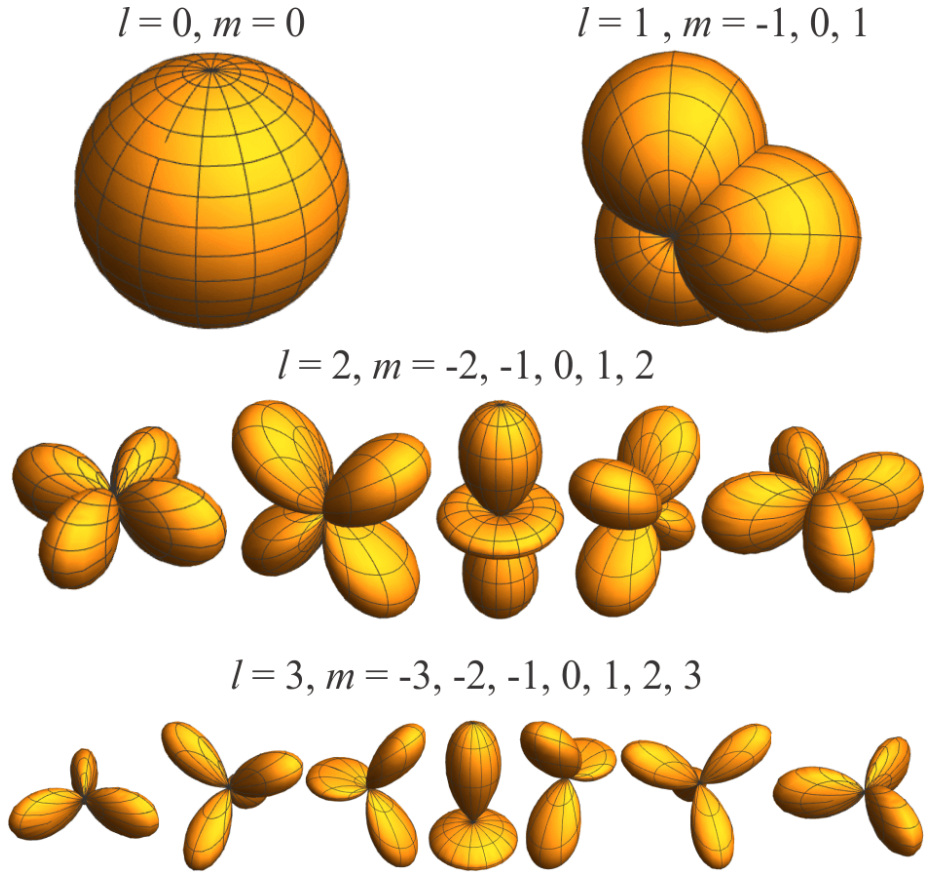
Assume **spherical envelope**: single $r(\theta, \varphi)$

$$r(\theta,\varphi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm} y_{lm}(\theta,\varphi)$$

Orthonormal for $(f,g) = \int f(\theta,\varphi) g(\theta,\varphi) d\Omega$

Rotation changes "inside l" (R – Wigner matrix):

$$a'_{lm} = \sum_{m'=-l}^{l} R^{l}_{mm'} a_{lm'}$$



We could find optimal rotation to minimize MSE:

$$MSE = \sum_{l} \sum_{m} (a_{lm} - b'_{lm})^2$$

To avoid searching - rotationally invariant fingerprints (RIFs): $A_l = \sqrt{\sum_m a_{lm}^2}$

This averaging discards lots of potentially valuable information!(selectiveness, decodability)

Let's **normalize** the rotation to use all **low order (more important) coefficients** …

PCA-SH: normalize rotation using principal component analysis

To be able to use all low order a_{lm}

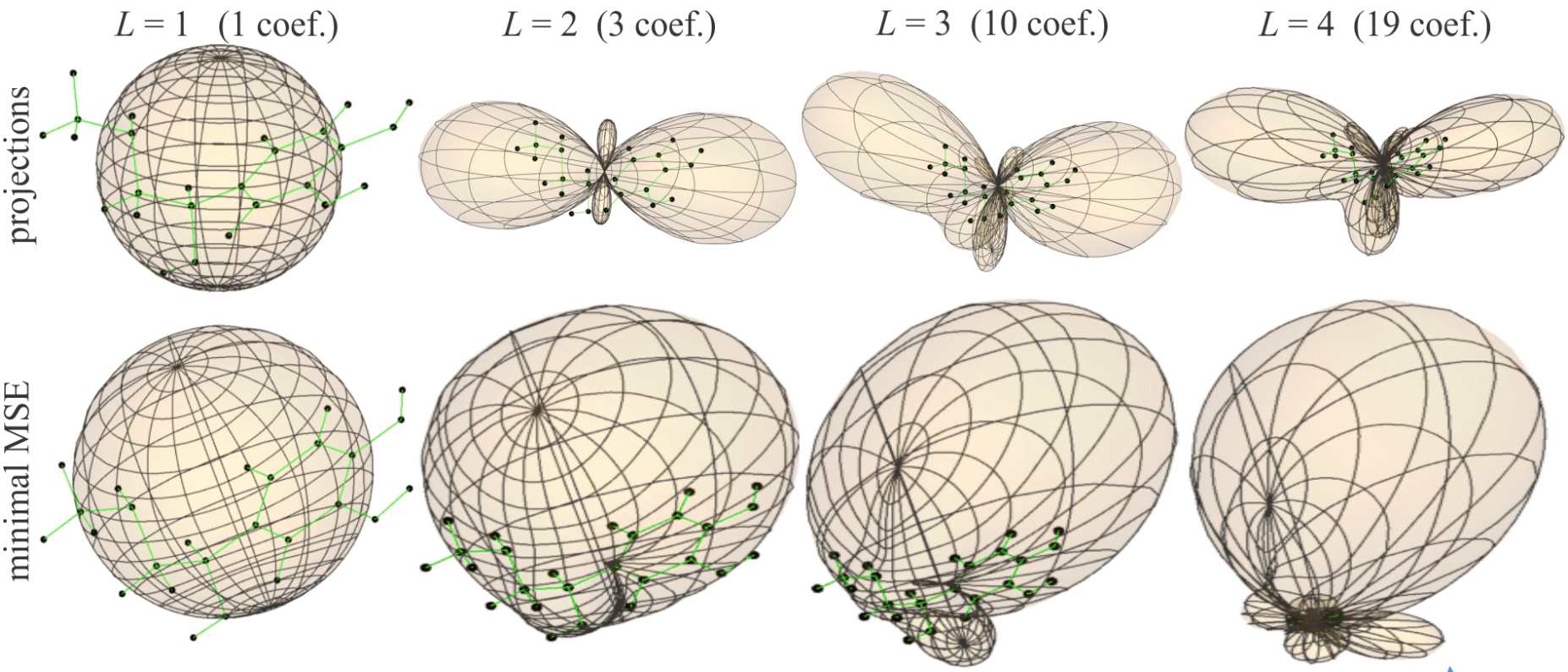
1) move the center of coordinates to the **centroid**,
2) calculate covariance matrix and its sorted **eigenvectors** $\boldsymbol{e_k}$,
3) change **signs** of e_1 and e_2 if needed to ensure $-min_k < max_k$,
4) change sign of e_3 if needed to ensure **preserved orientation**
5) **transform points** to the new base

Thanks to moving to centroid: all 3 coefficients for $l = 1$ nearly vanish

Thanks to PCA rotation: 3 of 5 coefficients for $l = 2$ are small

There remains: 1 for $l = 0$ (average radius), 2 for $l = 2$ (elongation),

7 for $l = 3$, 9 for $l = 4$ ….



For performance reasons, instead of $(f, g) = \int f(\theta, \varphi) g(\theta, \varphi) d\Omega$

We use $[f, g] \coloneqq \sum_i f(\theta_i, \varphi_i)\, g(\theta_i, \varphi_i)$ over all (**point**) atoms of molecule

Our base is no longer orthogonal for the $[\cdot,\cdot]$ scalar product …

However, minimizing MSE instead of direct projections, leads to **overfitting**.

A difficult question: what surface do we really want to represent?

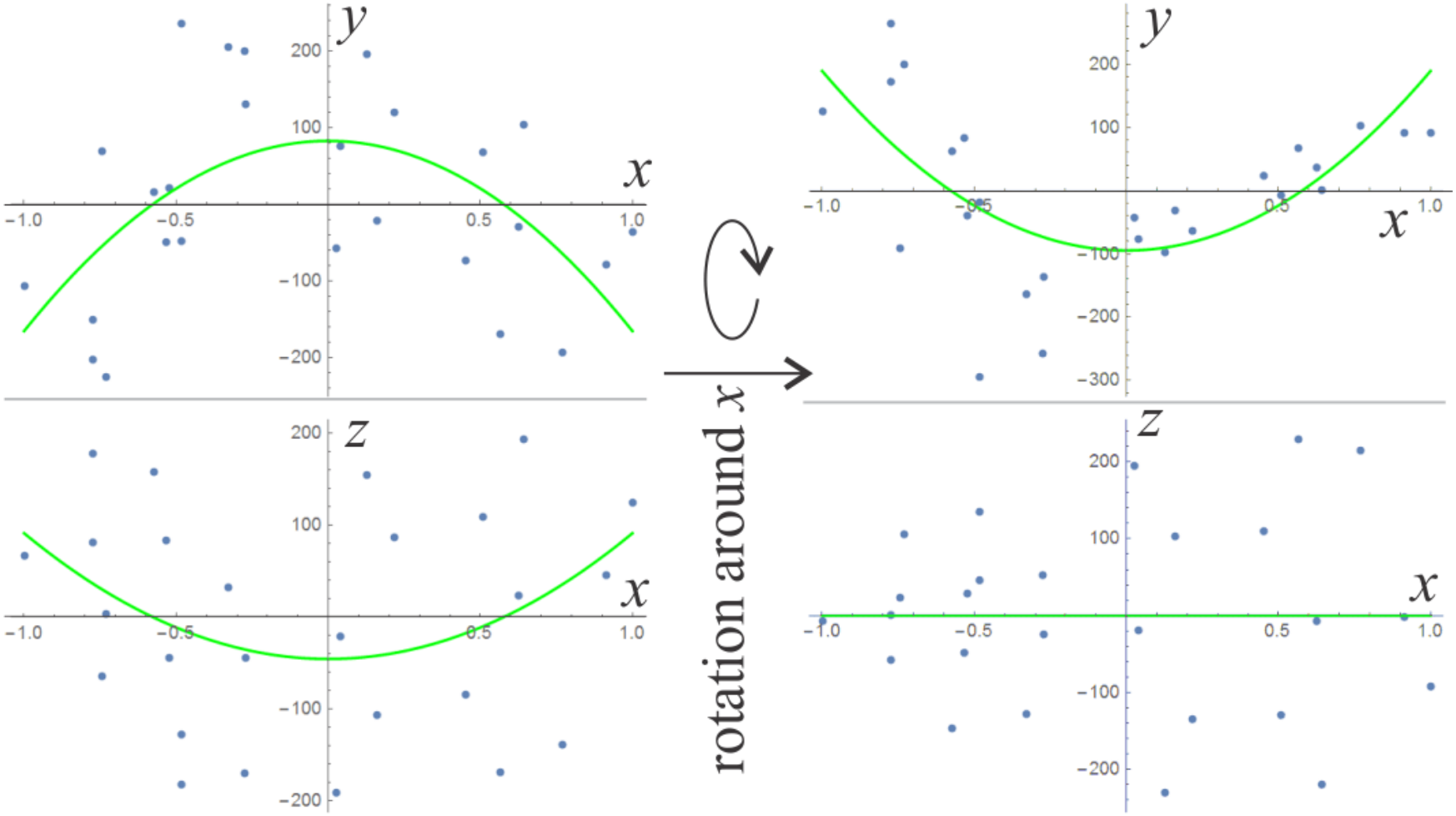
Minimizing MSE – going through all the atoms – it is not what we want …

Spherical harmonics: require there is one point in each direction – not true for ‘U’

Good for globular (sphere-like) molecules, not **ligands** which are **elongated** and **bent**.

Let’s **normalize to [-1,1]** for the main axis (x), fit $3x^2 - 1$ Legendre polynomial,

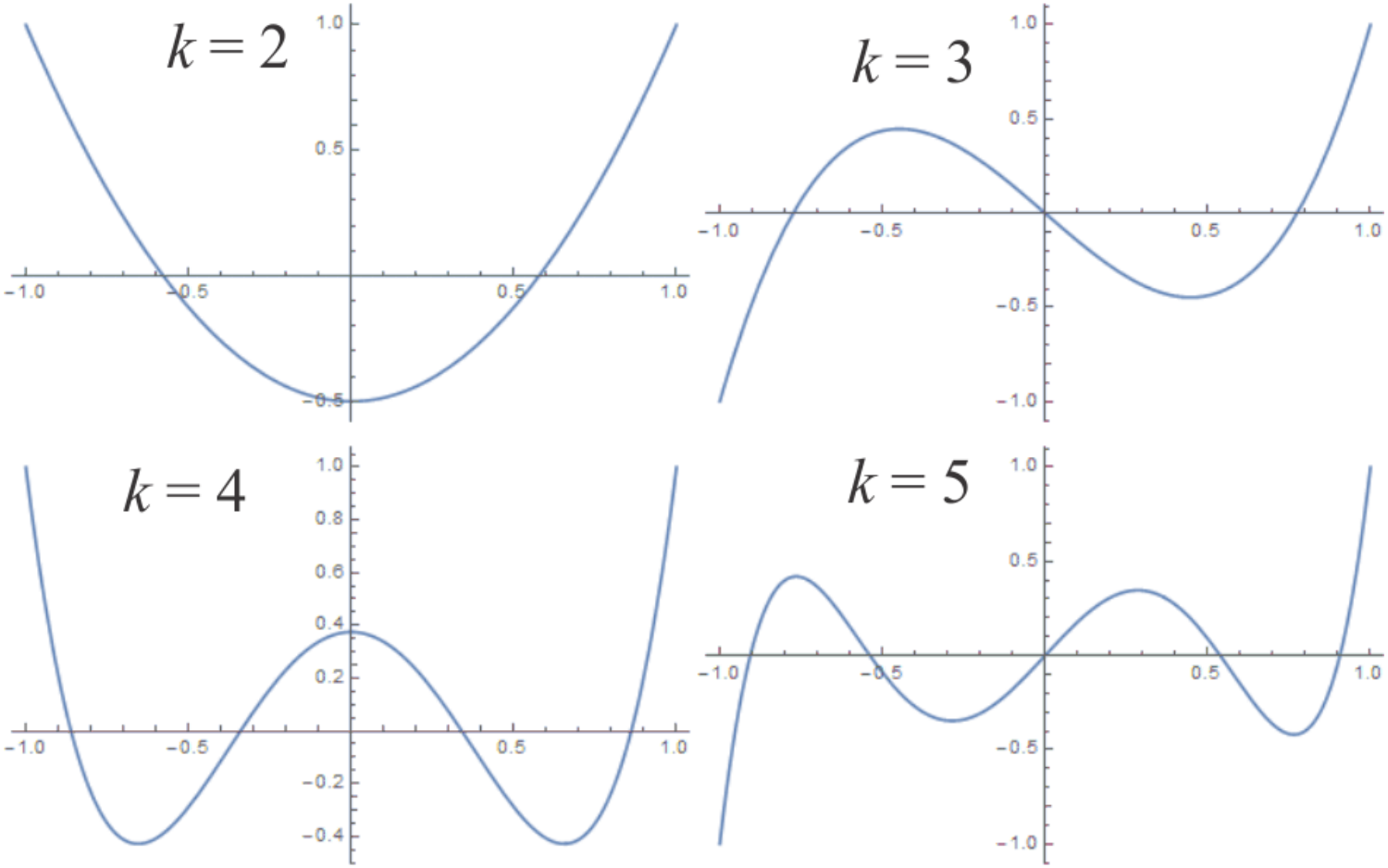
Then **rotate to ‘U’ shape in xy and flat in xz** – to discard the xz coefficeint

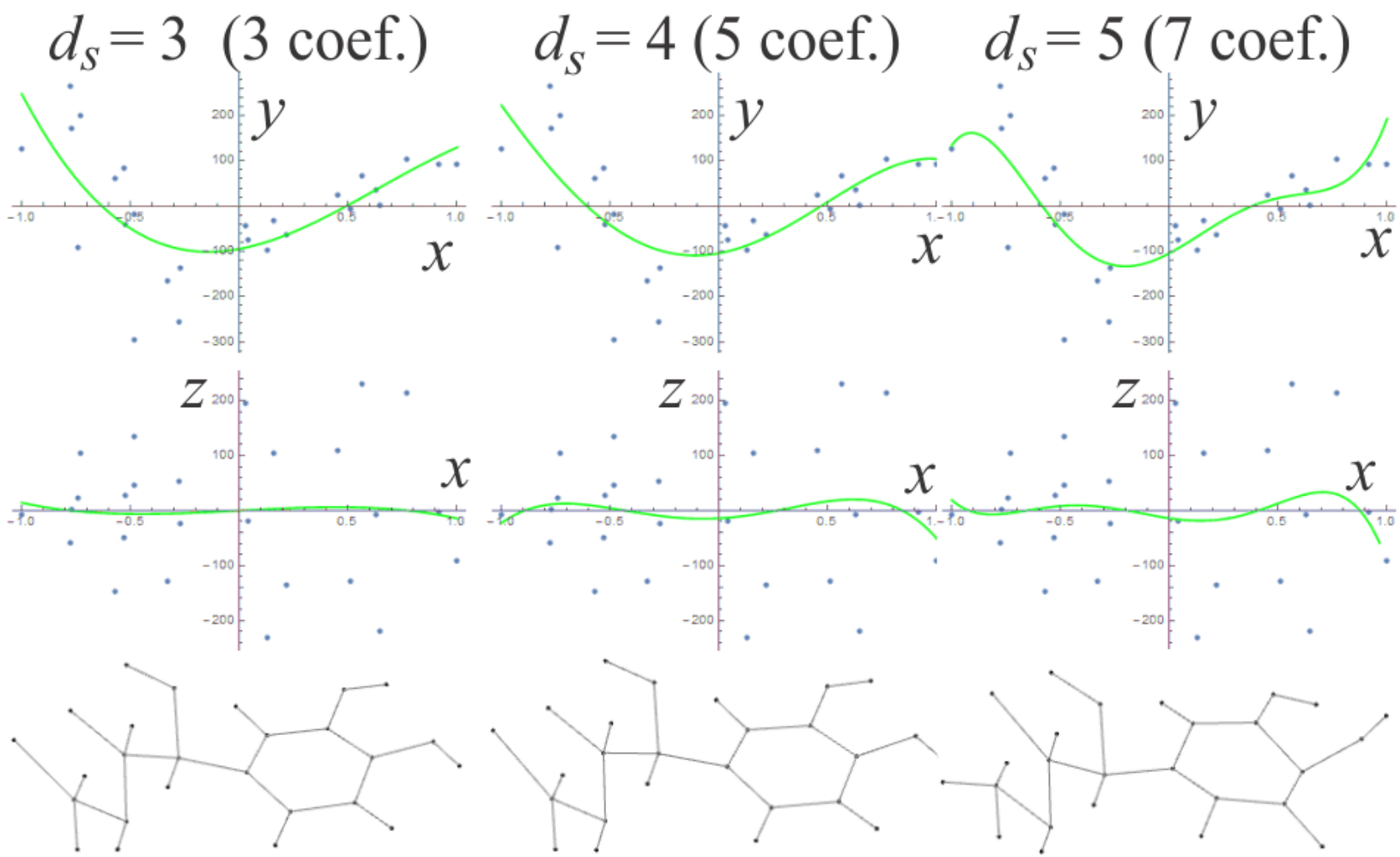


Legendre polynomial – orthogonal for $(f,g) = \int_{-1}^{1} f(x)g(x)dx$

Thanks to normalization, $k = 0,1$ terms vanish

for **‘**U**’**-like molecules $k = 2$ is sufficient, for ‘∼’-like we need $k = 3$ …



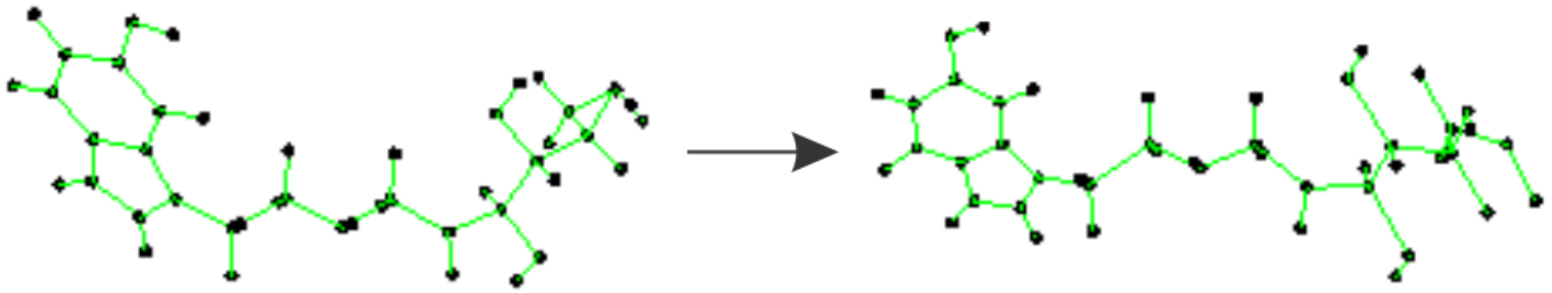


$k = 2$ 'spine' of molecule required 1 coefficient,

We can use higher order (2 coefficients/order) to improve agreement (and '∼' …)

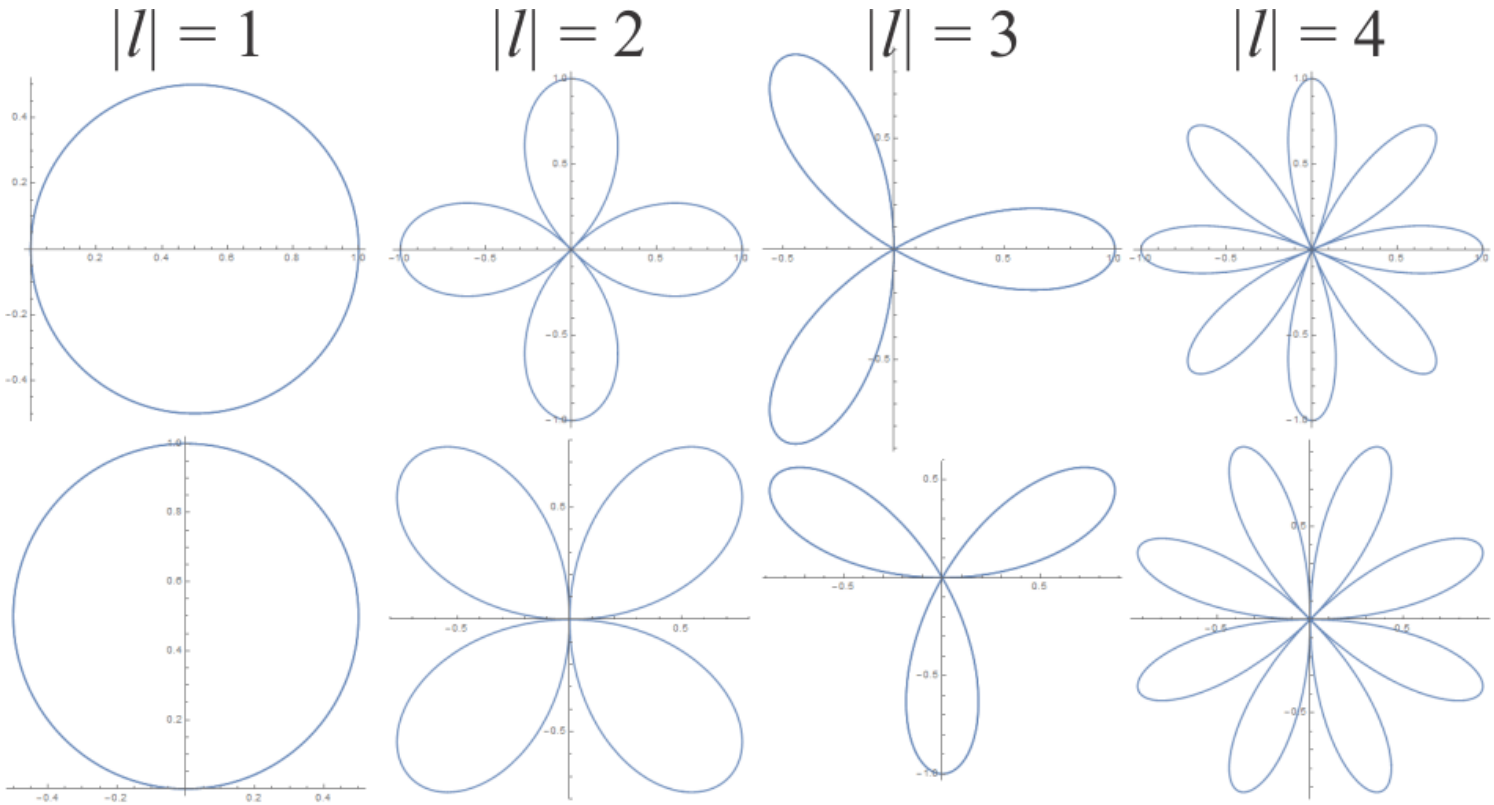
Finally: **unbend** the molecule – subtract the fitted polynomials ('spine')

Unbent molecule – now along the main axis x, we need to describe **evolution**



Cylindrical harmonics: $\sin l\varphi$ or $\cos l\varphi$ - orthogonal for $\int_0^{2\pi} f(\varphi)g(\varphi)\,d\varphi$

For **evolution**, we can use $P_k(x)c_l(\varphi)$ base



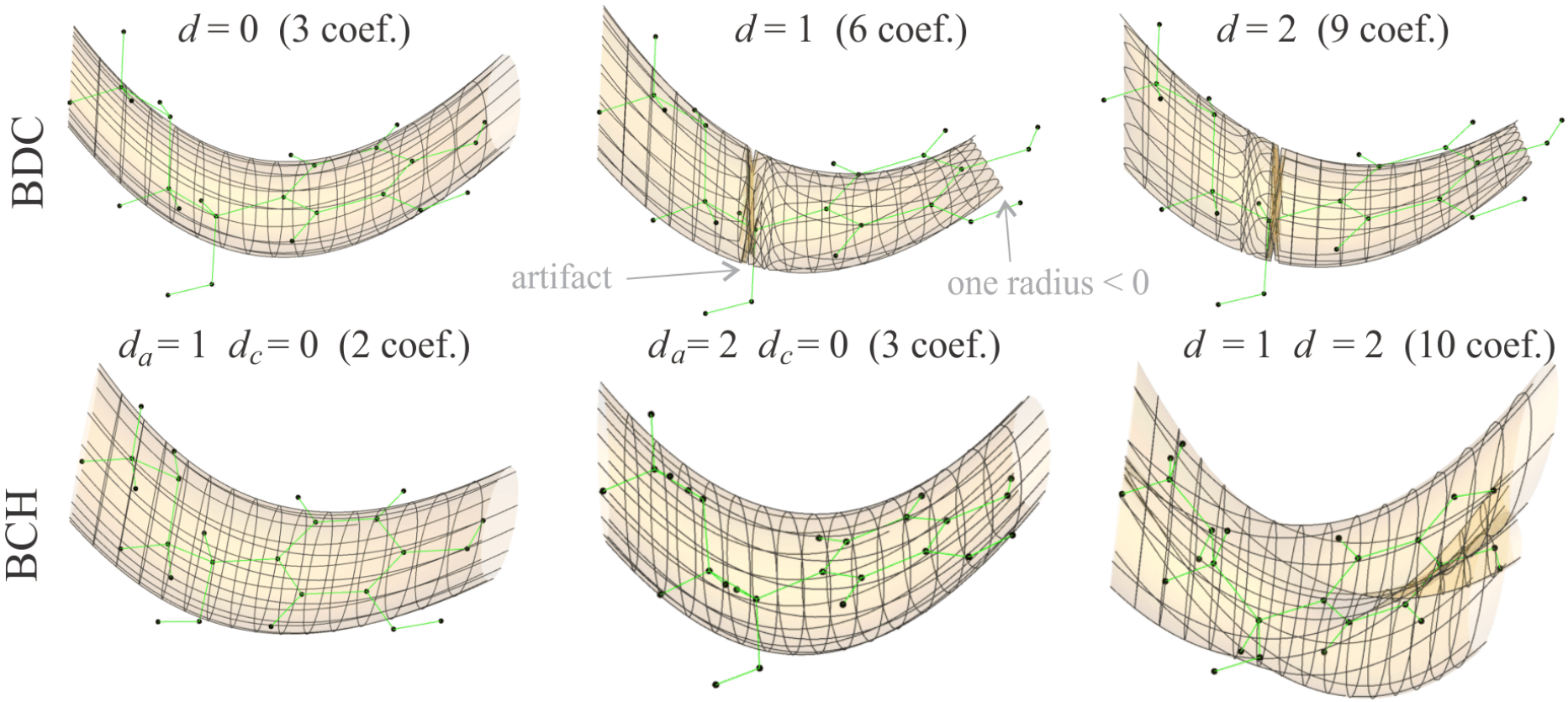
Bent cylindrical harmonics (BCH) – might be useful if molecule forks

Bent deformed cylinder (BDC) – use **evolving ellipse** instead

Fitting – e.g. polynomials (of x) for y^2, z^2, yz covariance matrix as a function of x

Ellipse (cross-section) from its eigenvalues and eigenvectors

6 parameters for basic **evolving ellipse**: $cov(x) = \begin{pmatrix} a+bx & c+dx \\ c+dx & e+fx \end{pmatrix}$



we have e.g. 8

parameters for **shape** –

can be used to **complement**

other descriptors

we can add more to describe

different properties - e.g.

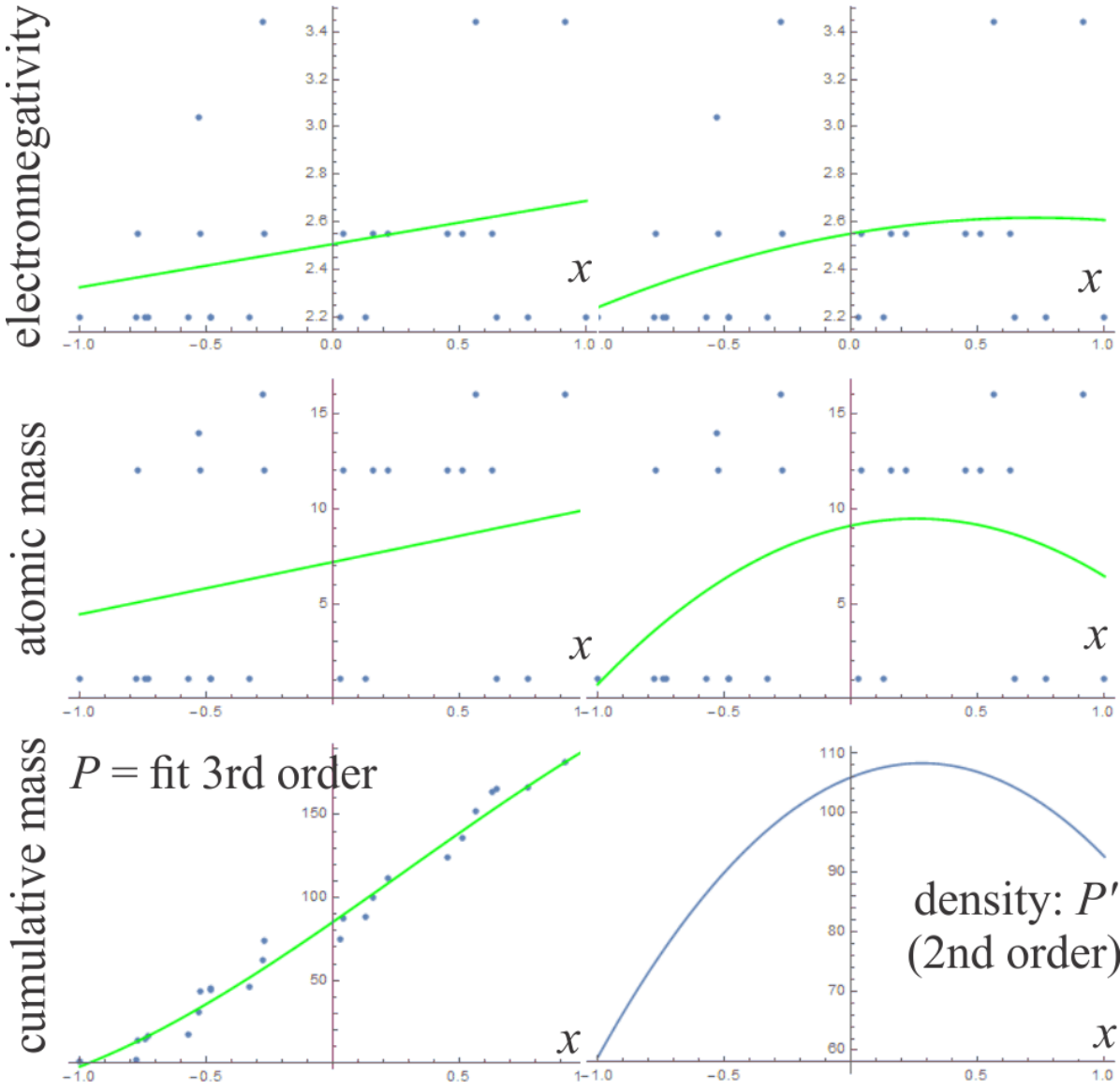
coefficients of polynomials along x:

- **Electronegativity**
- **Mass distribution**
- **flexibility**:

e.g take ensemble of conformations,

use **average** parameters

and their **variance** in ensemble

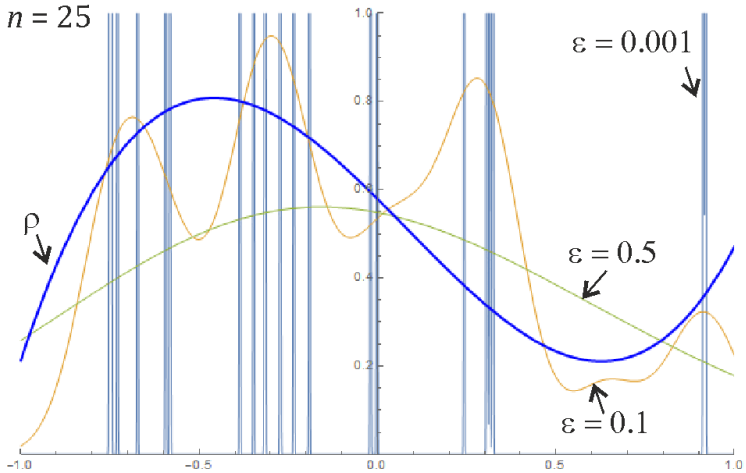


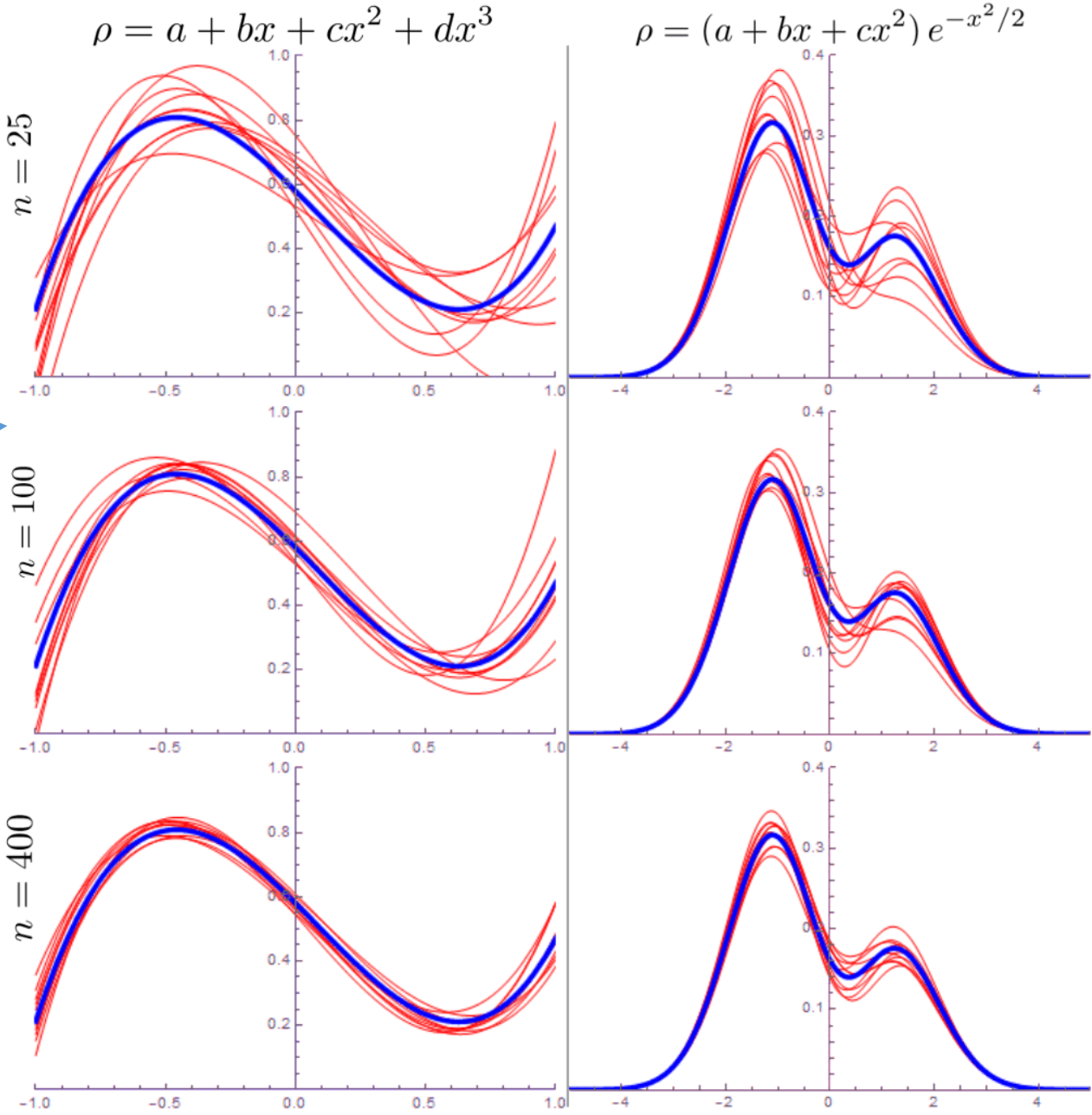
Parametric density estimation:

https://arxiv.org/abs/1702.02144

$$[f] = \frac{1}{n} \sum_{x \in \text{sample}} f(x)$$

$$\rho = \frac{5}{8}((1 - 3x^2) +$$
$$+[x](15x - 21x^3) +$$
$$+[x^2](9x^2 - 3) +$$
$$+[x^3](35x^3 - 21x))$$





General basic framework to describe molecules … shapes (e.g. **clusters**),

can be generalized to higher dimensions e.g. for various classification problems

Optimized for the interesting -

ligands – usually **elongated, bent** and **flattened** to fit into a protein:

- **Normalize** position and **rotation**,
- describe **bending** (spine)
- then **evolution of cross-section** along main axis (polynomial coefficients)
- then evolution of other properties along this axis

Further perspectives:

- **optimize for interesting types** of molecules, evaluate, compare …
- which interesting properties should we add to description? How?
- how to include **changes of conformations**?
- how to describe evolution of cross-section, fit parameters (continuity…)

e.g. for **disconnected cross-sections**? Algebraic manifolds? Fitting them?

Preprint: http://arxiv.org/pdf/1509.09211

Mathematica impl.: https://dl.dropboxusercontent.com/u/12405967/shape.nb