Chemometrics

Evgenii B. Rudnyi and Jan G. Korvink
IMTEK
Albert Ludwig University
Freiburg, Germany

Preliminaries

Learning Goals

- Introduction
  - Drug Design
  - Engineering Polymers
- Chemoinformatics
  - Databases
  - Descriptors
- QSAR and QSPR
  - Empirical Modeling
  - Multilinear Regression
  - Principal Component Analysis
  - Partial Least Squares

References

- Leach, A.R., *Molecular modelling : principles and applications*.

On-line resources

- *Multiway calibration in 3D QSAR*, Thesis, www.ub.rug.nl/eldoc/dis/science/j.nilsson/
- Electronic Statistics Textbook, www.statsoft.com/textbook/stathome.html
- Chemometrics World (links), www.spectroscopynow.com/Spy/basehtml/SpyH/1,2466,2-0-0-0-0-home-0-0,00.html

Introduction

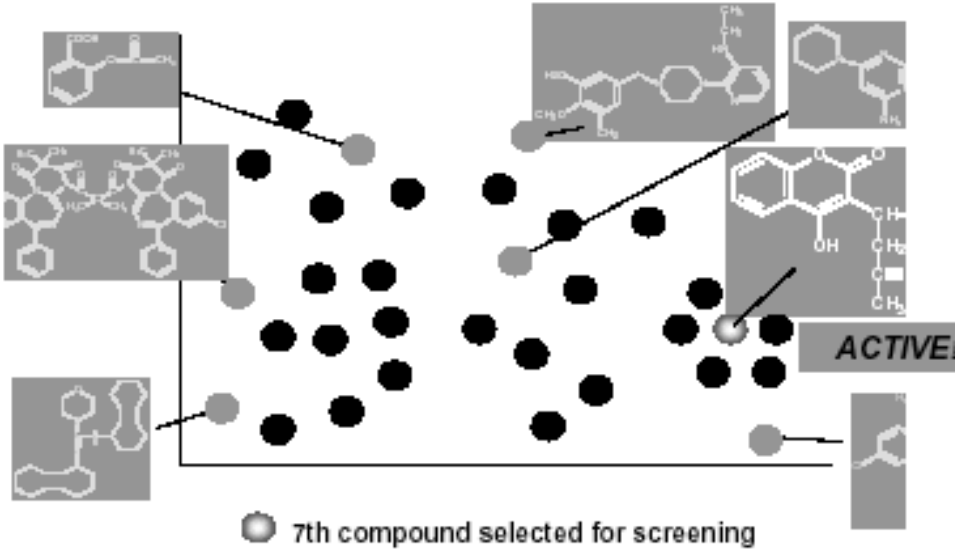
Drug Design

- Drug interact with proteins:
  - Docking problem.
  - Free energy (binding constant).
  - Needs pass through cell membranes.
- From the first principles the problem is untractable.
- How to choose next molecule?

Strategy

- To find a “lead series” and then “lead optimization”.

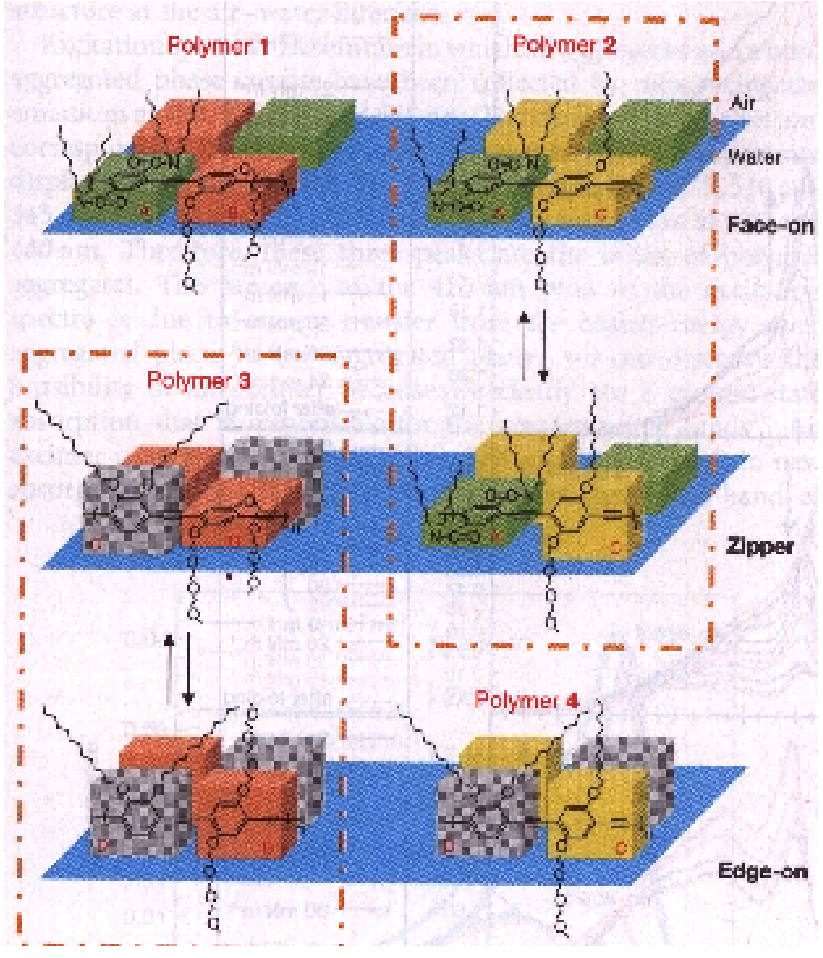
- Databases of chemical compounds.
  - American Chemical Society - a database of more than 18 million compounds.

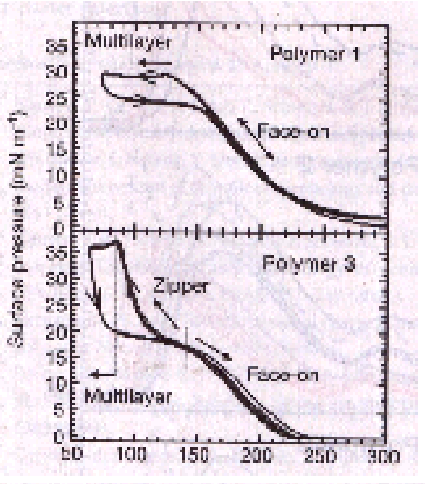


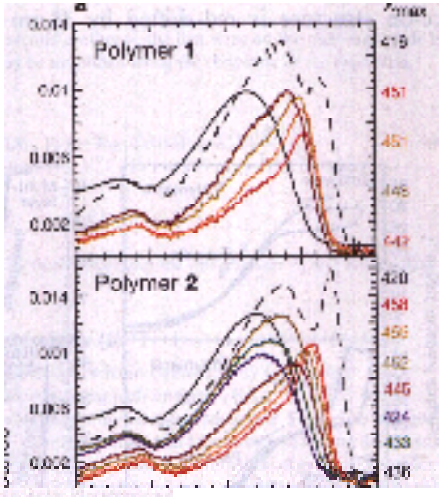
http://www.cc.nih.gov/ccc/principles/pdf/00-01/gorman-slides.pdf

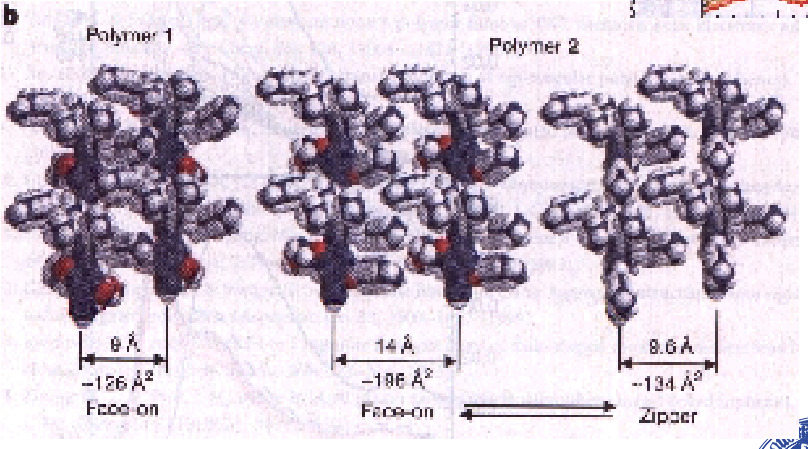
Introduction: Engineering Polymers

- J. Kim, T.M, Swager. Control of conformational and interpolymer effects in conjugated polymers, Nature, 2001, v. 411, p. 1030-1034





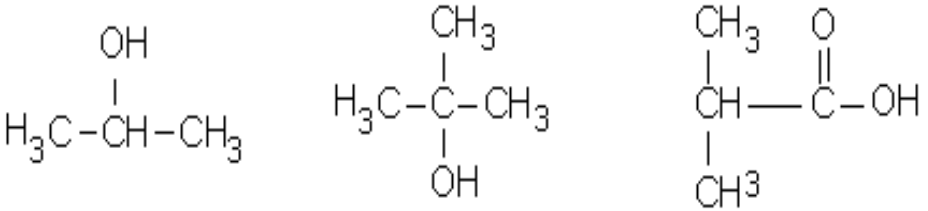




Chemoinformatics

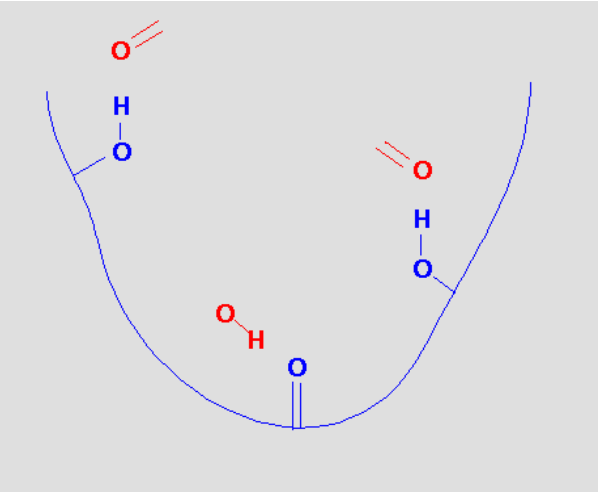
Databases

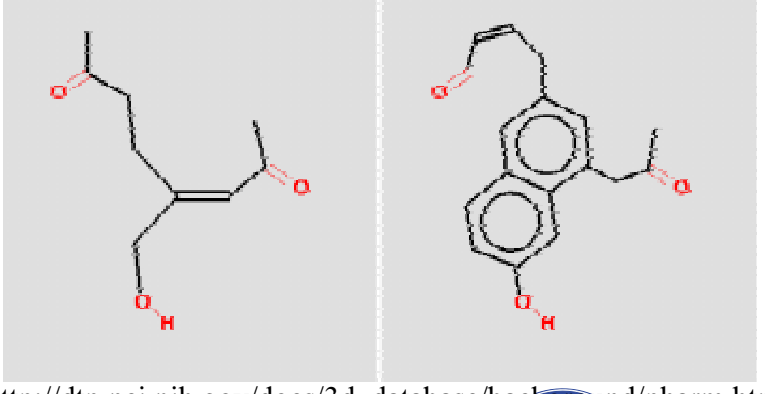
- 2D databases
- Search for molecules with given functional groups:
  - Graph representation
  - Simplified Molecular Input Line Entry System (SMILES)



http://esc.syrres.com/interkow/docsmile.htm

- Virtual molecules
- 3D databases: pharmacophores

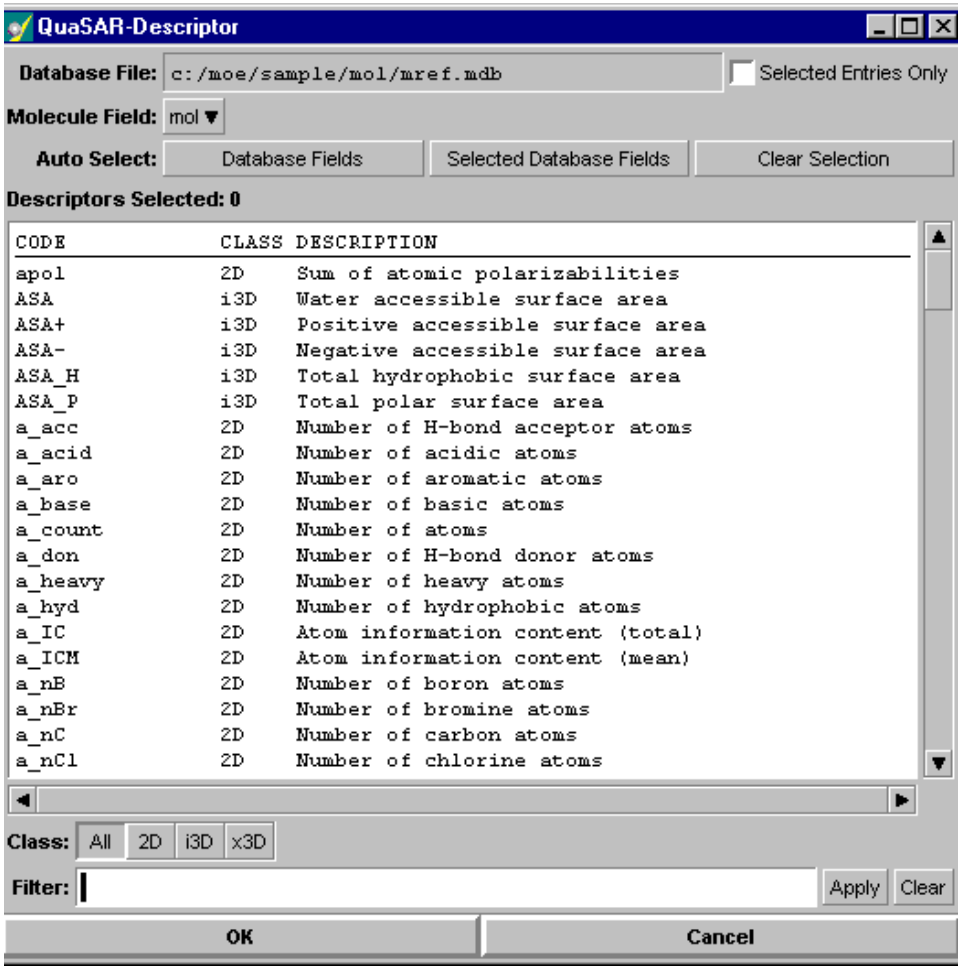




http://dtp.nci.nih.gov/docs/3d_database/background/pharm.html

Descriptors

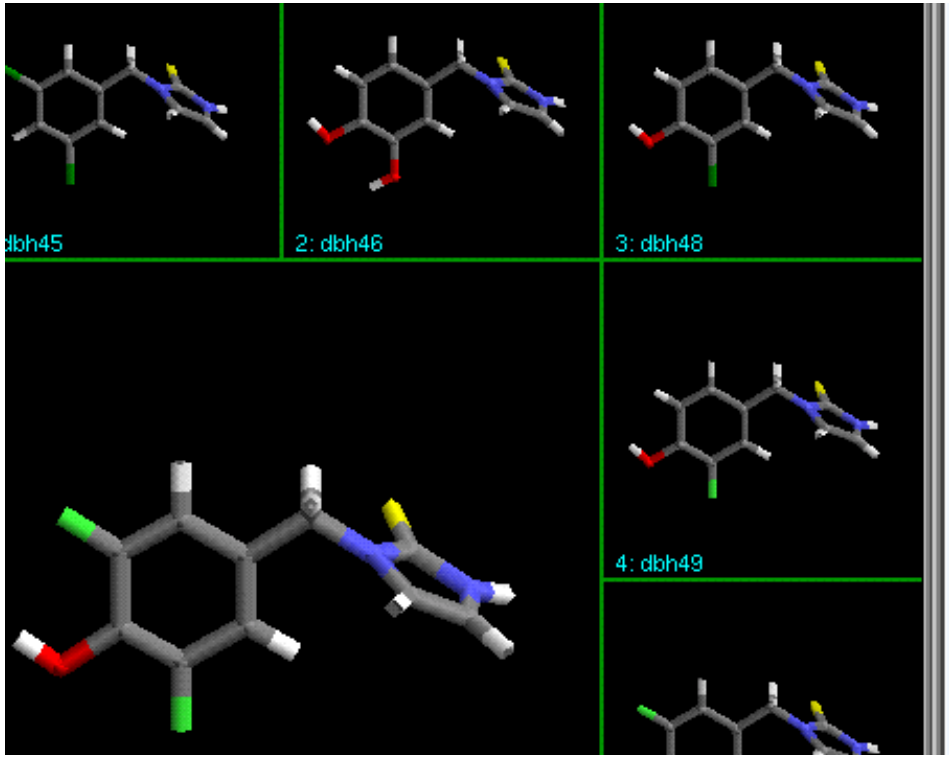
- Properties of the molecule that can be easily computed from its structure:
  - Molecular weight,
  - Molar refractory,
  - HOMO, LUMO (from QC),
  - and so on.
- Examples at www.chemcomp.com/feature/descr.htm
- Descriptors can be used to find similar molecules:
  - various similarity coefficients.



QSAR and QSPR

Empirical Modeling

- Quantitative Structure - Activity Relationship (QSAR)
- Quantitative Structure - Property Relationship (QSPR)



RUN | GFA | Current Defaults Set: QSAR

http://www.accelrys.com/cerius2/images/qsar_1.gif

R1 C3 (Derived): ==3.14769 + 0.028503 * col "Pi-0^2" + 1.0661 * SP

	Y1 -log(IC50)	GFA Predicte	GFA Residual	STEPWISE Pr	STEF
1.	3.00	3.71	-0.71	2.84	
2.	3.15	3.75	-0.60	3.64	
3.	3.30	3.43	-0.13	2.32	
4.	3.45	3.95	-0.50	3.94	
5.	3.47	3.69	-0.22	3.68	
6.	3.47	3.61	-0.14	3.78	
7.	3.70	3.86	-0.16	3.74	
8.	3.76	3.67	0.09	3.61	
9.	3.81	3.80	0.01	3.32	

QSAR and QSPR

Multilinear Regression

- $\boldsymbol{y} = X \cdot \boldsymbol{b} + \boldsymbol{\varepsilon}$
  - $\boldsymbol{y}$ - dependent variable
  - $\boldsymbol{b}$ - unknown parameters
  - X - matrix of descriptors
    - rows - molecules
    - columns - properties
  - $\boldsymbol{\varepsilon}$ - error
- Can be generalized to several outputs $Y = XB + E$.
- Linear in parameters

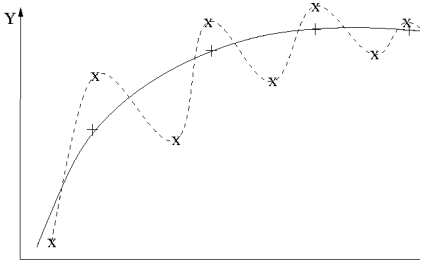
  $y = a + bx^2 + c\ln(x)$.
- Estimate $\boldsymbol{b}$ from a known data set and use it for new molecules.
- Ordinary least squares

  $\boldsymbol{b} = (X^T X)^{-1} X^T \cdot \boldsymbol{y}$

Problems:

- Colinearity in X,
- Number of columns may be greater than that of rows,
- Strong correlation between columns,
- There are errors in X,
- Overfitting - bad predictive power.

http://www.cs.duke.edu/~natsev/thesis/node10.html



Principal Component Analysis

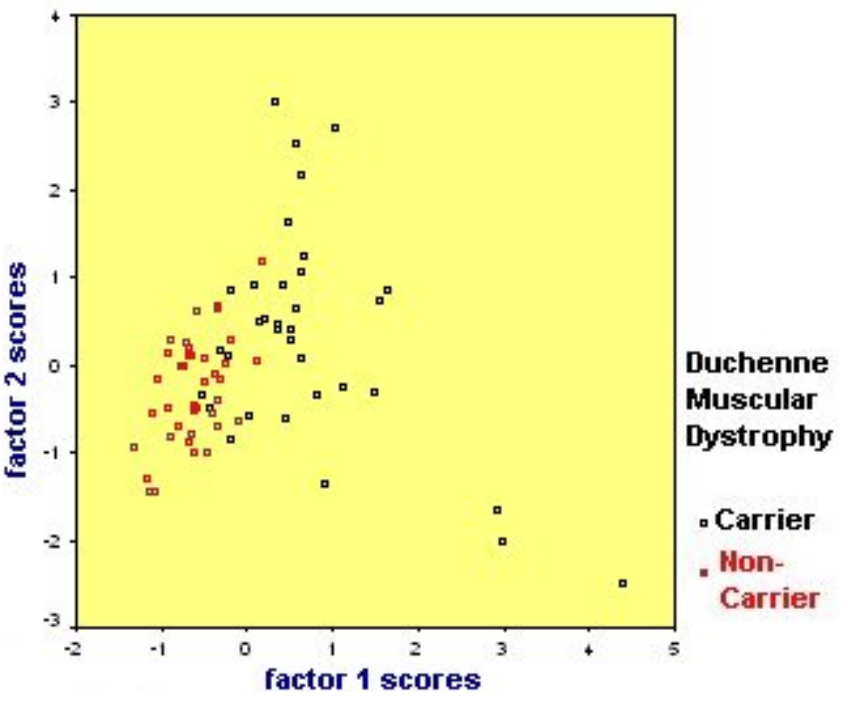
- Low-rank approximation of X (low-dimensional subspace)

$$X = YZ^T + E$$

- Can be computed by using first dominant SVD-vectors

$$\hat{X} = \hat{U} \cdot \hat{\Sigma} \cdot \hat{V}^T = \sum_{i=1}^{k} \sigma_i (\boldsymbol{u}_i \boldsymbol{v}_i^T)$$

- Can be used for classification.
- Principal Component Regression - uses first factors.



http://obelia.jde.aca.mmu.ac.uk/multivar/pca_eg2b.htm

QSAR and QSPR

Partial Least Squares

- Similar to PCR but uses $\boldsymbol{y}$ to find factors in X.
- Regression in latent variables $\boldsymbol{y} = \boldsymbol{t} \cdot \boldsymbol{b} + \boldsymbol{\varepsilon}$.
- Latent variables are determined by descriptors $\boldsymbol{t} = C\boldsymbol{x}$.
- There are just few latent variables.

Model Validation

- How to validate regression model?
  - If errors possess the normal distribution, there are strait-forward formulas.
  - To divide the molecules to training and validation sets.
- Nonparametric statistics
  - Jack-knife
    - Take one sample out, make the calculations, then repeat for all samples.
  - Bootstrap
    - Take some samples out but replace them by others.
  - Cross-validation

Applications

Evgenii B. Rudnyi and Jan G. Korvink
IMTEK
Albert Ludwig University
Freiburg, Germany

Learning Goals

- Polymers
  - Tethered polymers
  - Branched polymers
  - Polymeric networks and gels
  - Coarse-grained models of polymers
- Inorganic Materials

- Steered Molecular Dynamics

Tethered Polymers

- Gary S. Grest, Normal and Shear Forces Between Polymer Brushes, Advances in Polymer Science, 1999, Vol.138, p. 149-183.

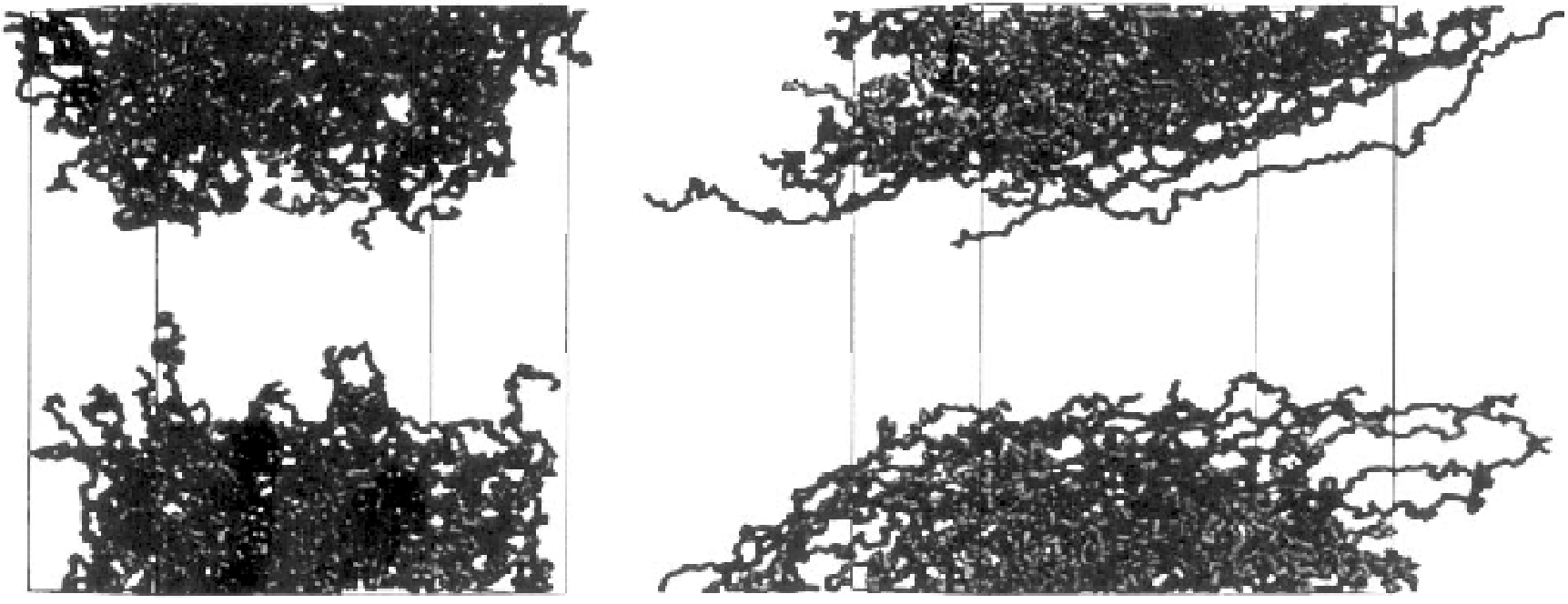
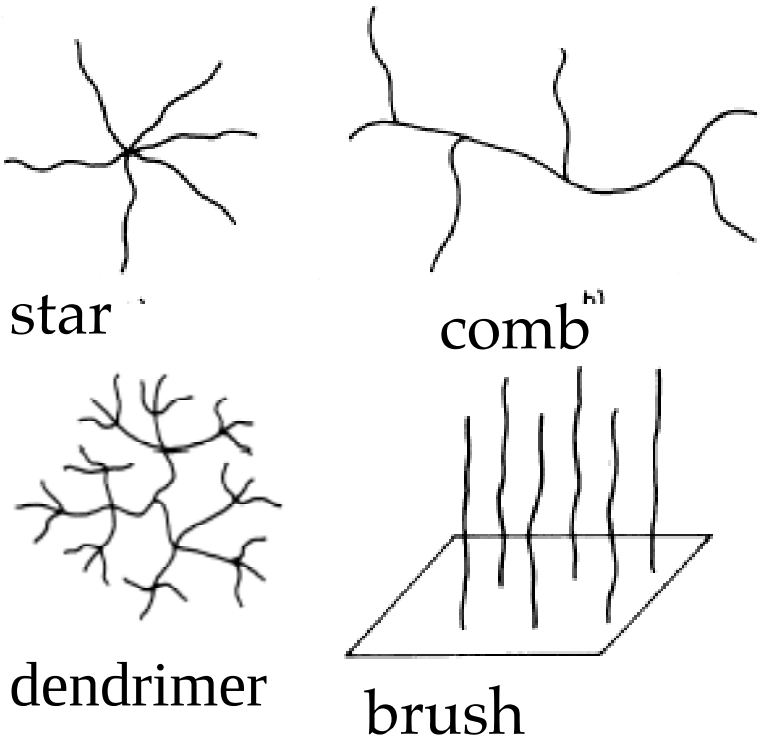


Fig. 5. Typical configurations of a brush of chain length N=100 immersed in a melt of mobile dimers for ρ_a=0.03σ^{-2}. The solvent is not shown for clarity. The shear velocity is v_w=0 (**a**) and 0.2σ/τ (**b**). The dimensions of the cell are A=40.8$^2\sigma^2$ and D=64.9σ. Result from molec-

Branched Polymers

- Juan J. Freire, Conformational Properties of Branched Polymers: Theory and Simulations, Advances in Polymer Science, 1999, Vol.138, p. 35-112



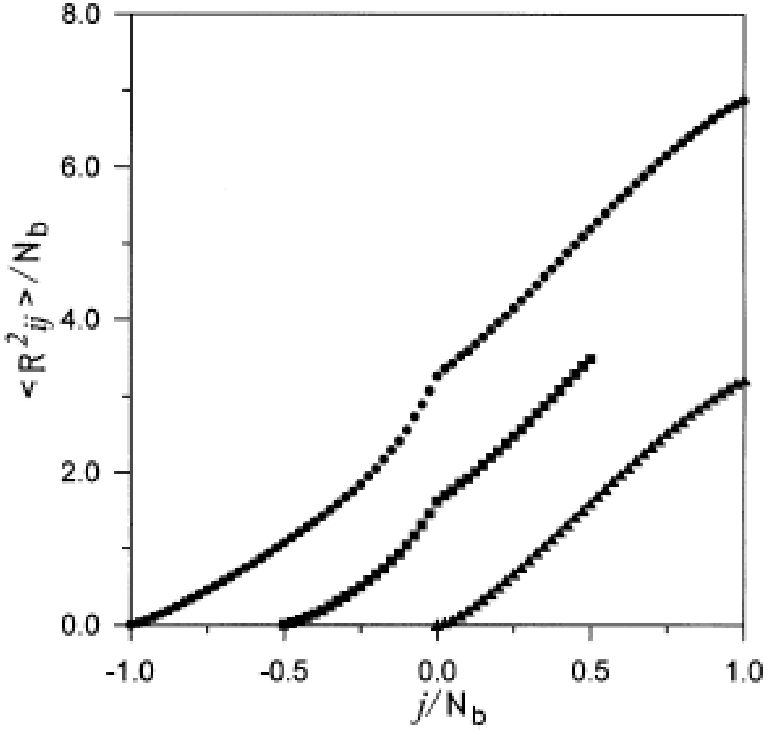
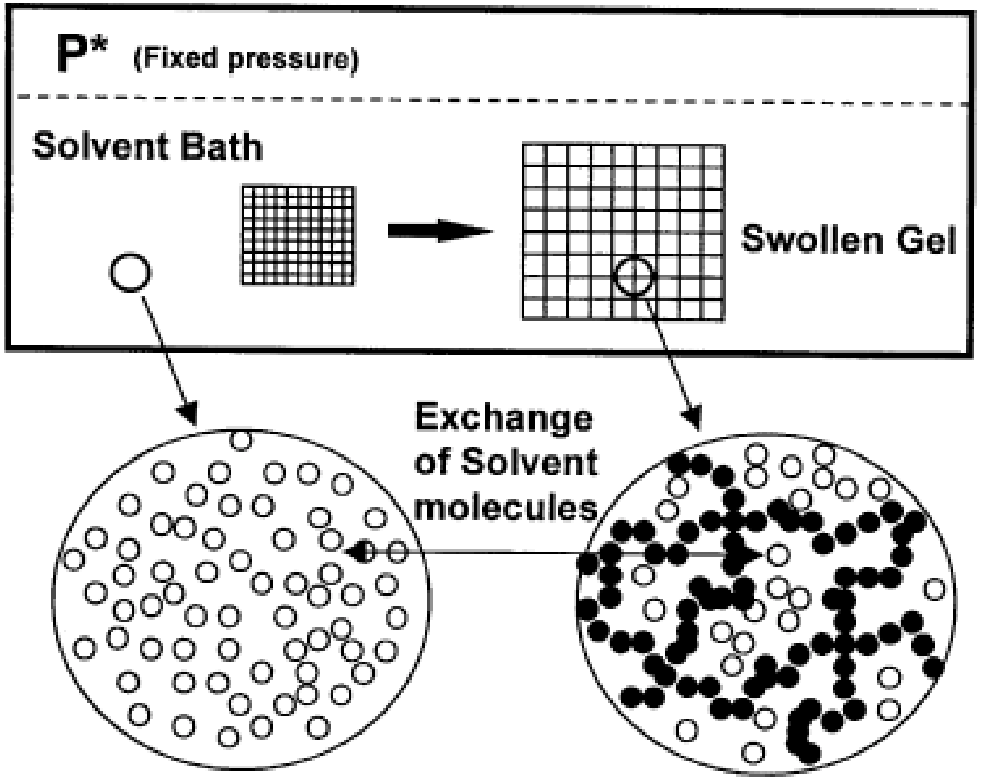
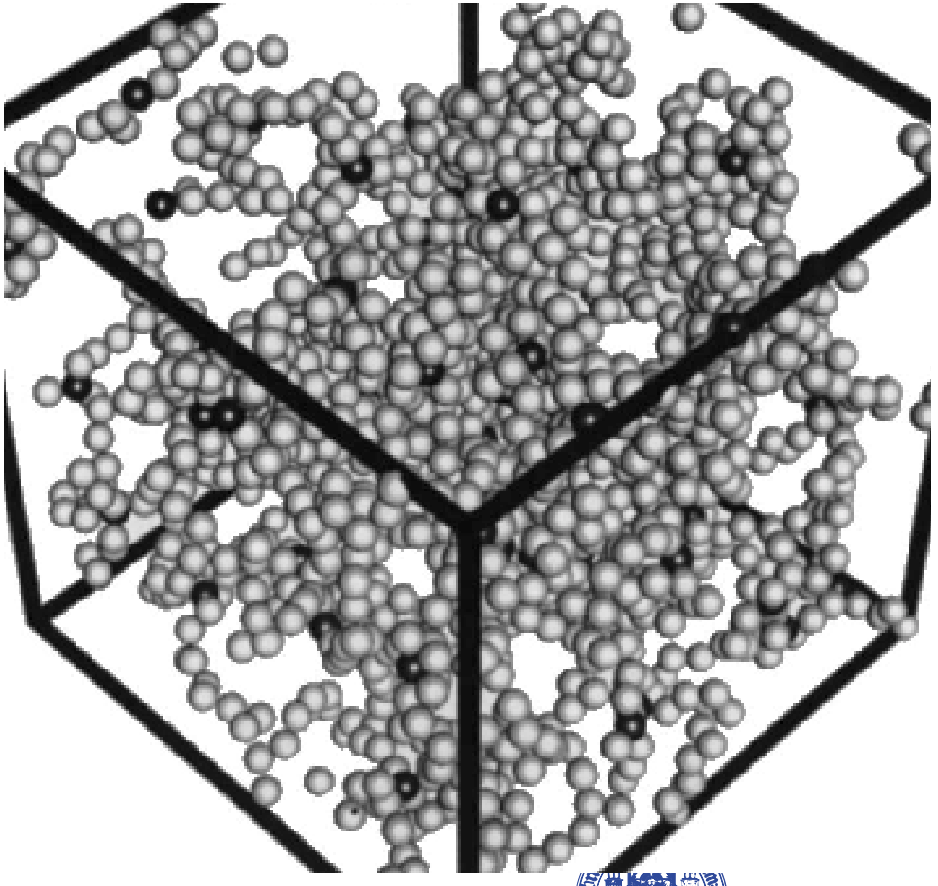


Fig. 14. Normalized averaged intramolecular distances plotted as a function of the position of bead j for a star with 12 arms with a total of 472 bonds. The beads are labeled as negative from $-N_b$ to 0 (the central atom) on the first arm and as positive up to N_b on the second arm.

Polymeric Networks and Gels

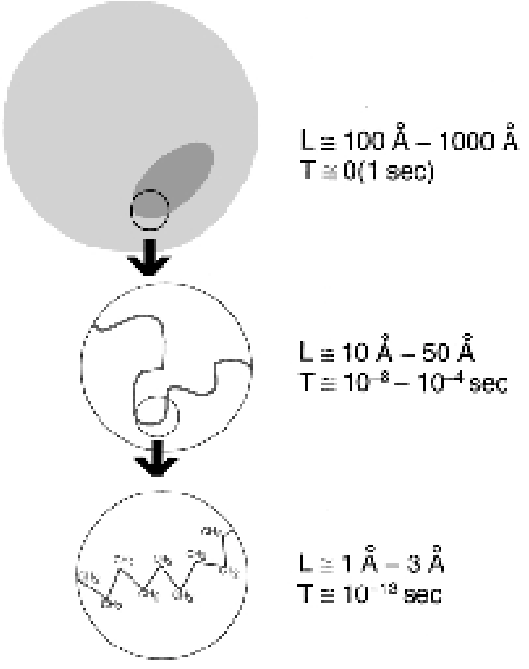
- Fernando A. Escobedo, Juan J. de Pablo, Molecular simulation of polymeric networks and gels: phase behavior and swelling, Physics Reports 318 (1999) 85-112

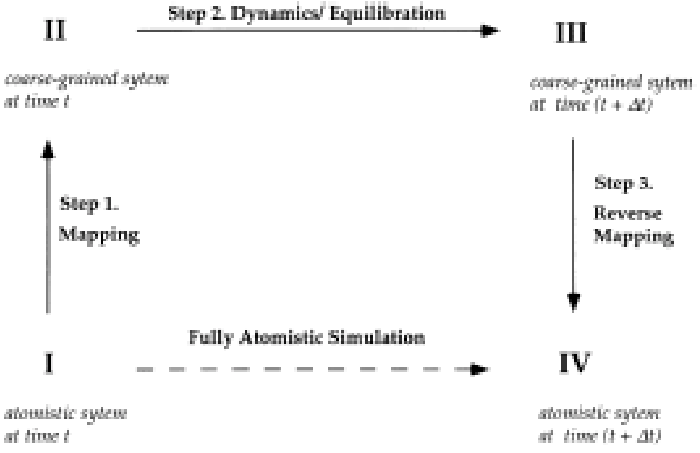


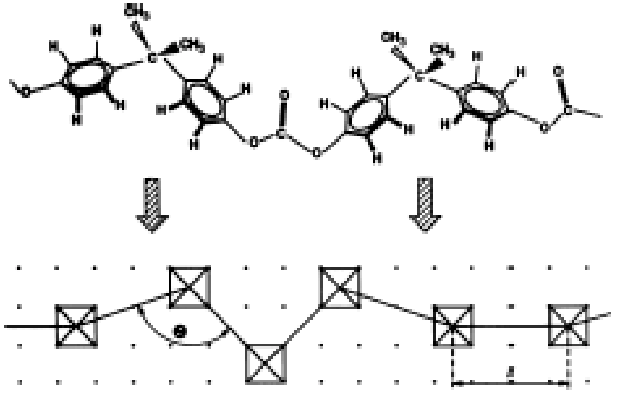


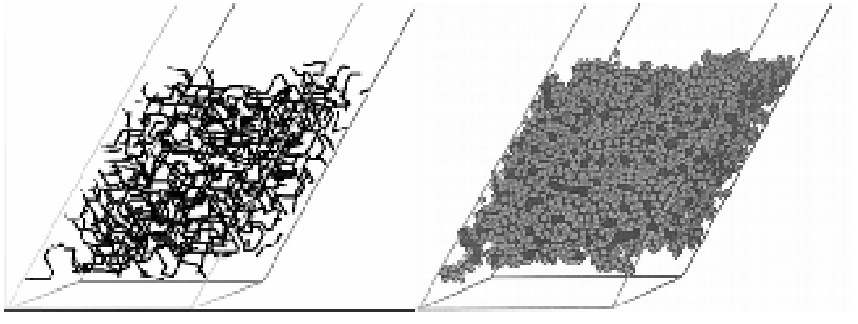
Coarse-grained Models

- J. Bashnagel et al, Bridging the gap between atomistic and coarse-grained models of polymers: status and perspective, Advances in Polymer Science, 2000, v. 152, p. 42-156







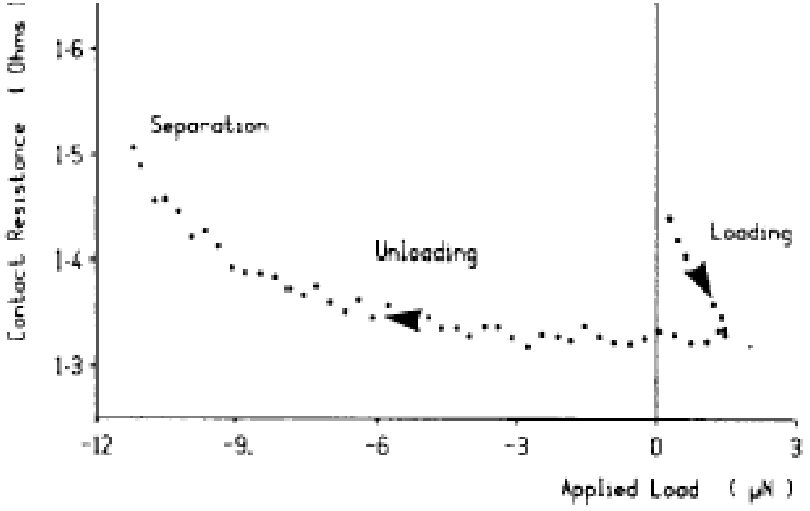


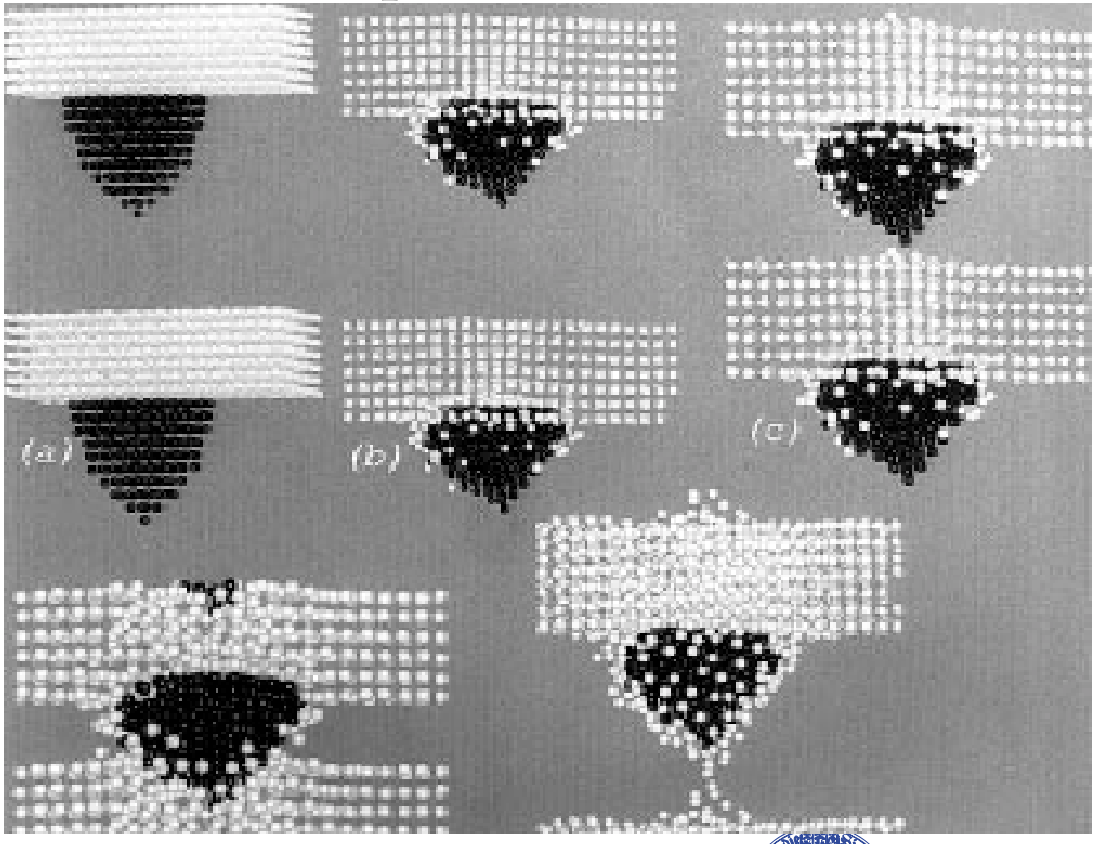
Inorganic Materials

- Nanocontacts
- Nano-scale films
- Crack propagation
- H. Raffi-Tabar, MODELLING THE NANO-SCALE PHENOMENA IN CONDENSED MATTER PHYSICS VIA COMPUTER-BASED NUMERICAL SIMULATIONS, Physics Reports 325 (2000) 239}310
- Demos from http://hal6000.thp.Uni-Duisburg.DE/~kai/index_1.html

Nanocontacts

- Material surface is quite rough on atomic and molecular scales and they are decorated with a variety of irregularities in all directions.
- Point-contact adhesion and indentation experiments





Nano-scale Films

- Adsorption and intercalation of Ag atoms on graphite

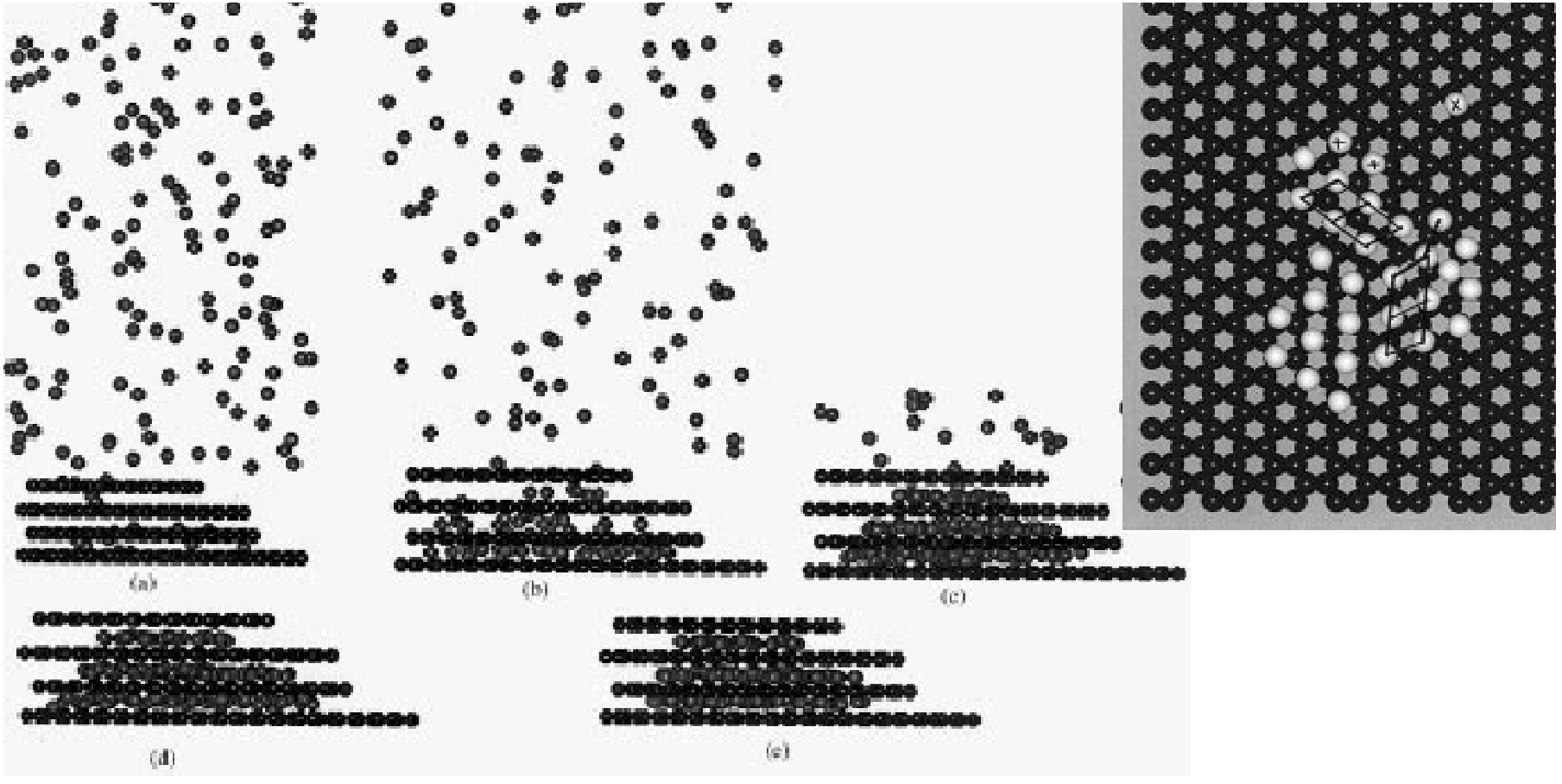
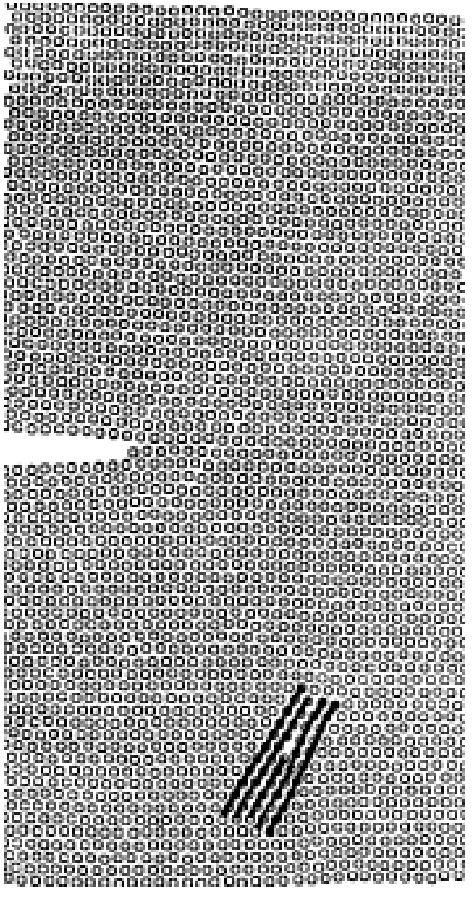
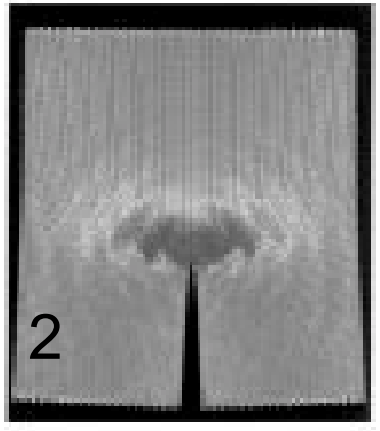
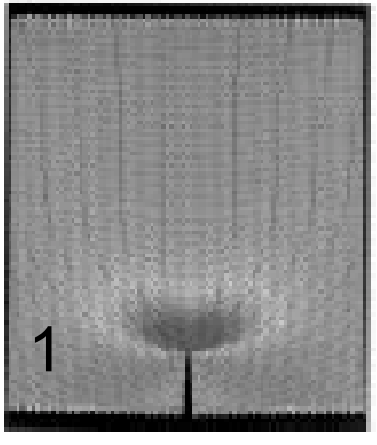


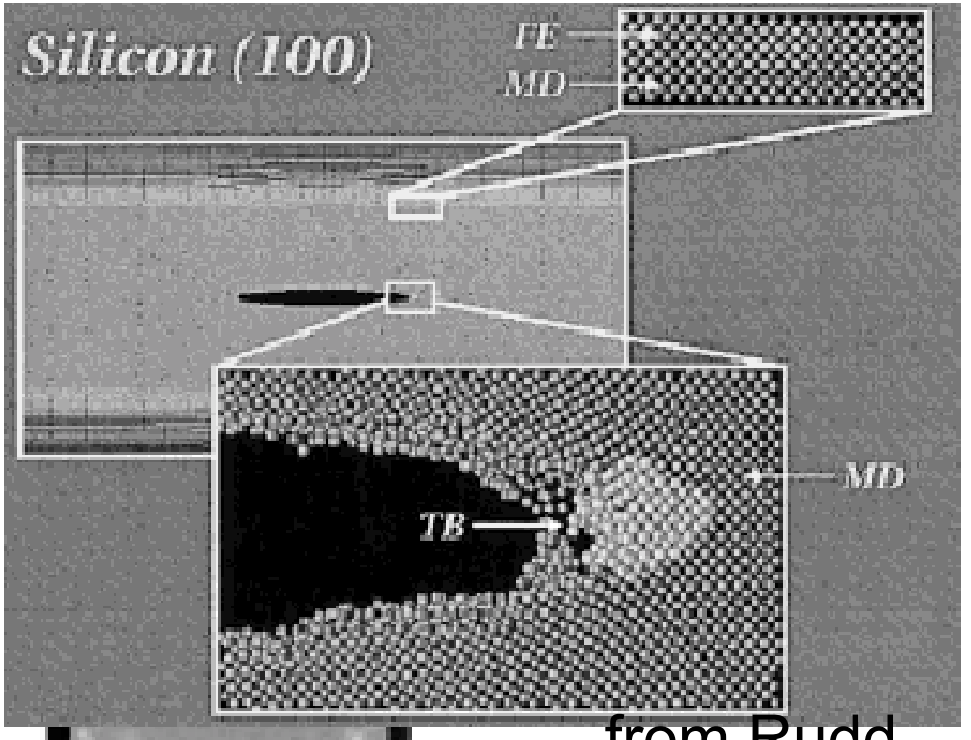
Fig. 26. Three-dimensional geometries of the simulated Ag/graphite system in the ⟨1 0 0⟩ direction after: (a) 7.5 ps; (b) 30 ps; (c) 75 ps; (d) 106 ps and (e) 180 ps, corresponding to Ag atom temperatures of respectively $T = 2355$; 2100; 1590; 1240; 400 K.

Crack Propagation

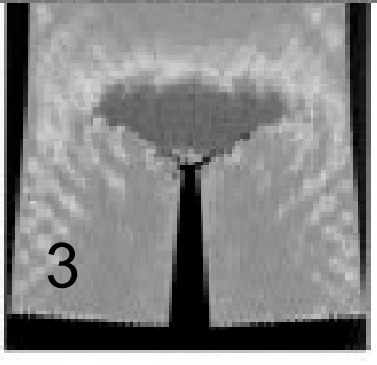








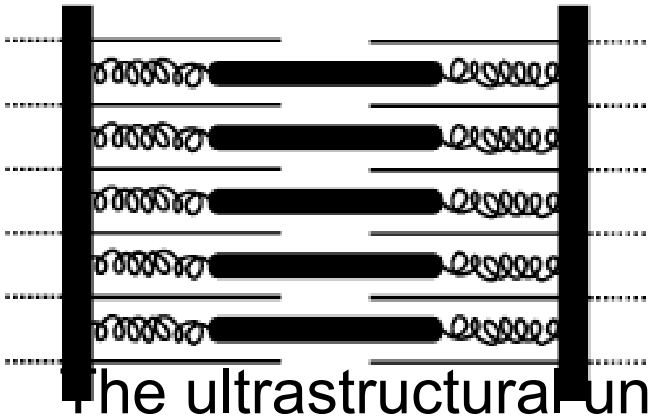
from Rudd



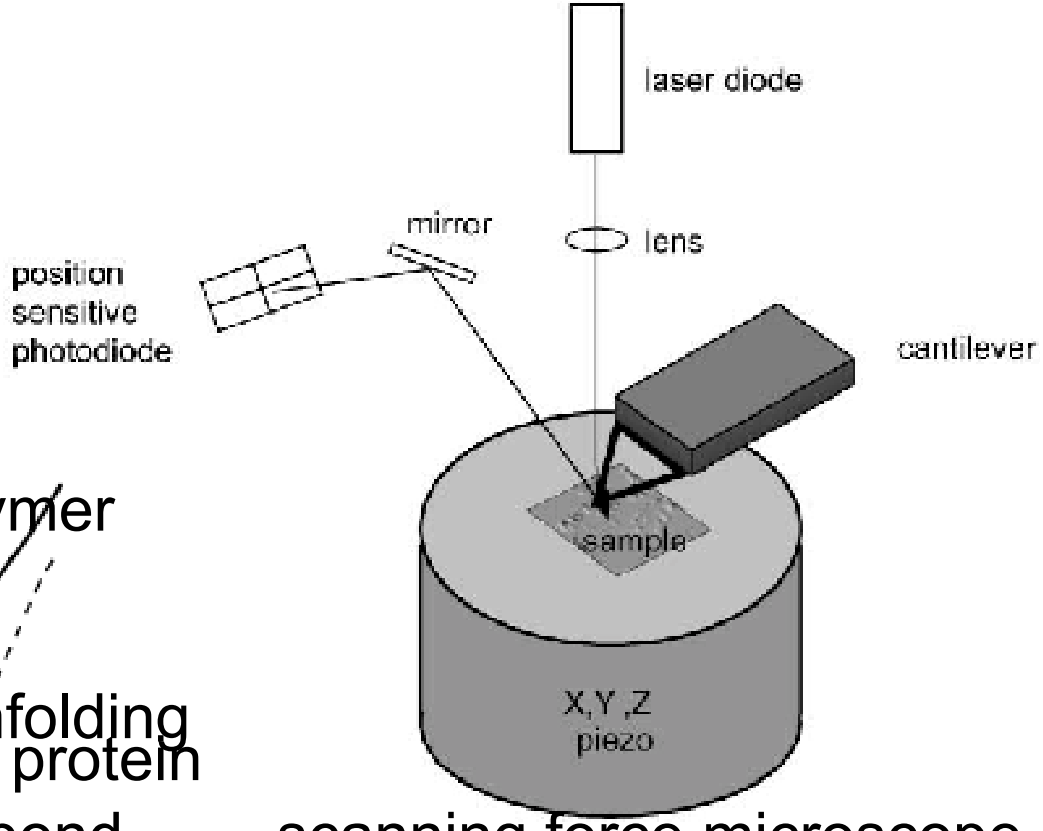
Steered Molecular Dynamics

- Experiments
  - R.Merkel, FORCE SPECTROSCOPY ON SINGLE PASSIVE BIOMOLECULES AND SINGLE BIOMOLECULAR BONDS, Physics Reports 346 (2001)343 -385
- Simulation
  - S. Izrailev et al, Steered Molecular Dynamics, http://www.ks.uiuc.edu/Research/
- Demo from http://www.ks.uiuc.edu/Research/

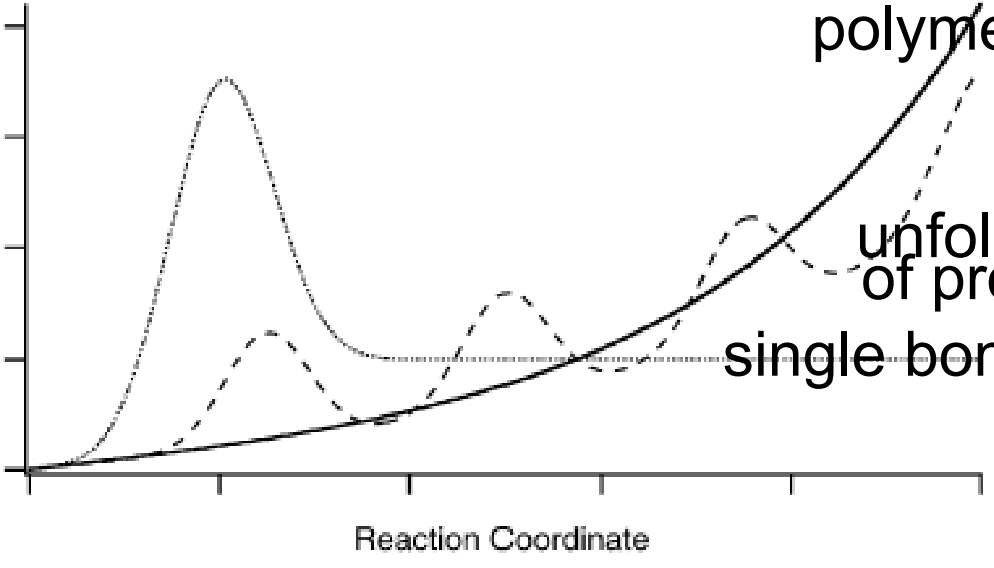
Experiments



The ultrastructural unit of skeletal muscle



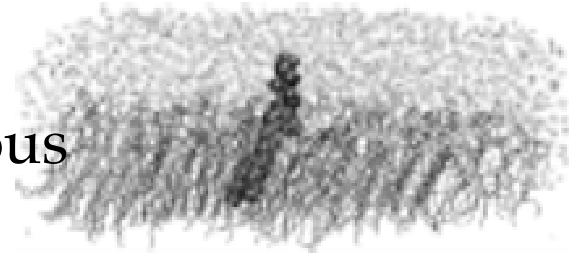
scanning force microscope



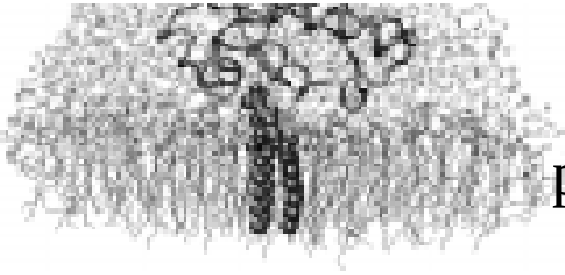
Simulation

- Extraction of a ligand from the binding pocket of a protein

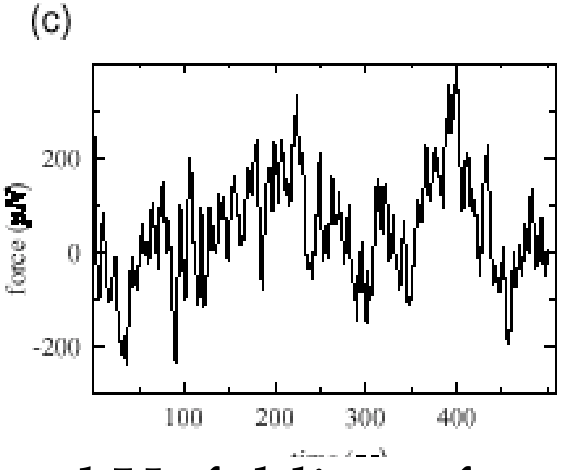
extraction to the aqueous phase

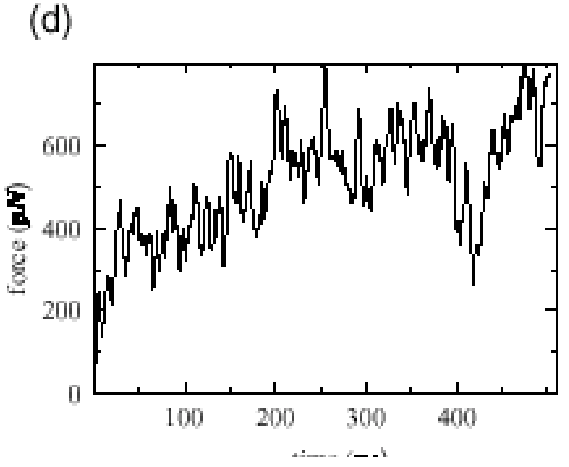


- Extraction of Lipids from Membranes



extraction into protein phospholipase





- Force-Induced Unfolding of Titin

Summary

- Polymers
- Inorganic materials
- Steered molecular dynamics