From Molecules to Cells- Towards whole cell simulations

Dr. Jeffrey Skolnick

Director

Center for the Study of Systems Biology
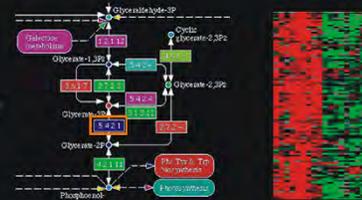**Georgia Institute of Technology**

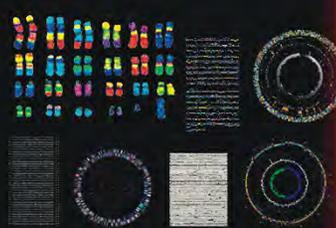# Overview of General Approach

Sequence to Function

EFICAz^2

AGSVVICDT	EC 4.2.1.1	TLAEPNG.A
GGTV.ICES		WKIDQQG.G
A.TI.ICES	YWRR.MGHL	TIAD.NAYG
A.SVVLCDS	YURR.IGRL	SLAE.QAYG
G.TLIICET	TUKYHMAN.	SIAE.NAFA
EC 1.2.1.3	FWRRKLGR.	EC 5.4.1.1
	YWKRHMGH.	

Enzyme Function Inference

Computational Metabolomics

C•Met

Cancer Biology

Sequence to Structure

PROSPECTOR_3.5

TASSER

FINDSITE^{LHM}

FINDSITE

Binding Site Prediction

multi-FINDSITE

Prediction of Protein-Protein Interactions

Druggable Target Prediction

m-TASSER

Structure to Function

LIGAND HOMOLOGY MODELING AS A NEW COMPUTATIONAL PLATFORM TO SUPPORT MODERN DRUG DEVELOPMENT

LIGAND HOMOLOGY MODELING: INTRO
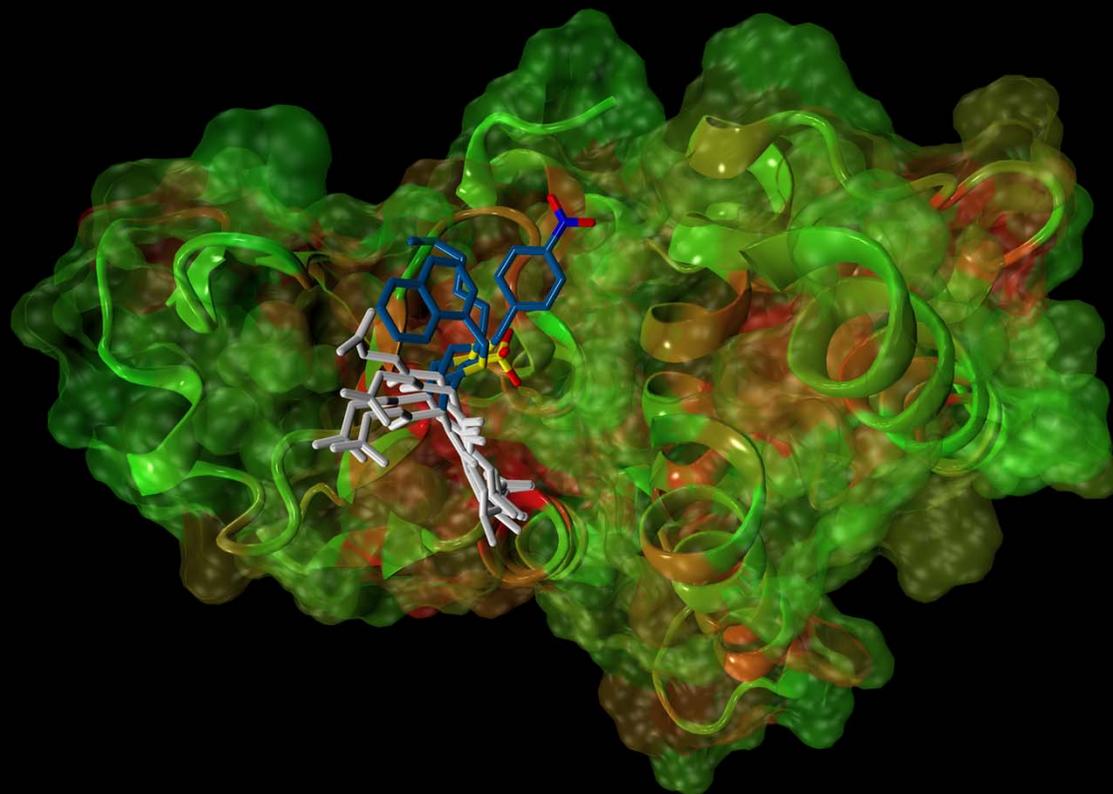
Structure-based virtual screening in drug discovery

Target structure ← X-Ray / NMR
apo / holo
- - - - - - - - - - - -
Protein models
close / remote homology

Binding site detection

Compound docking ← Screening library

Compound ranking → Ranked library
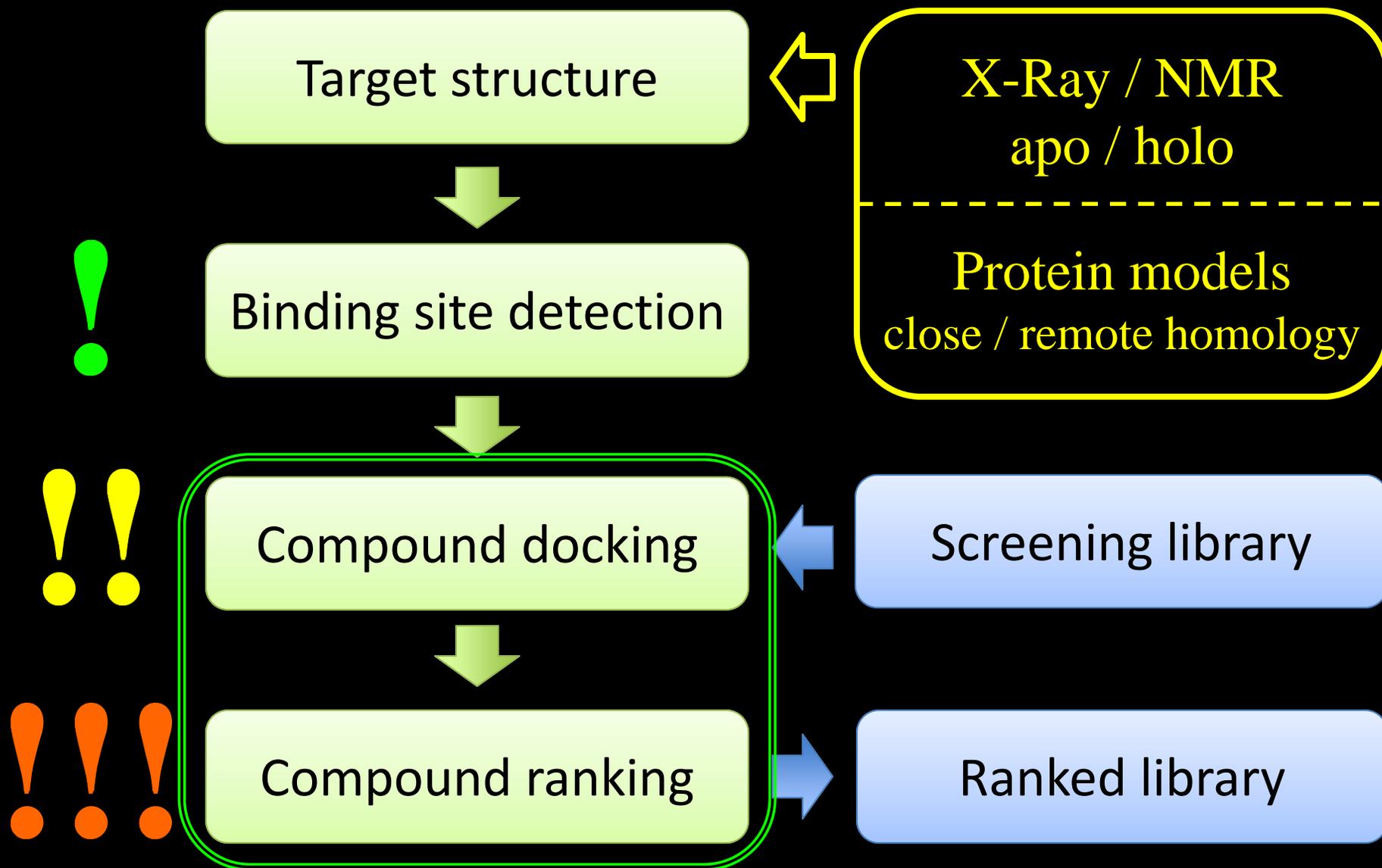
Requirements for ligand docking/scoring approaches:

① Accurate ligand binding pose prediction
② Reliable compound ranking in virtual screening
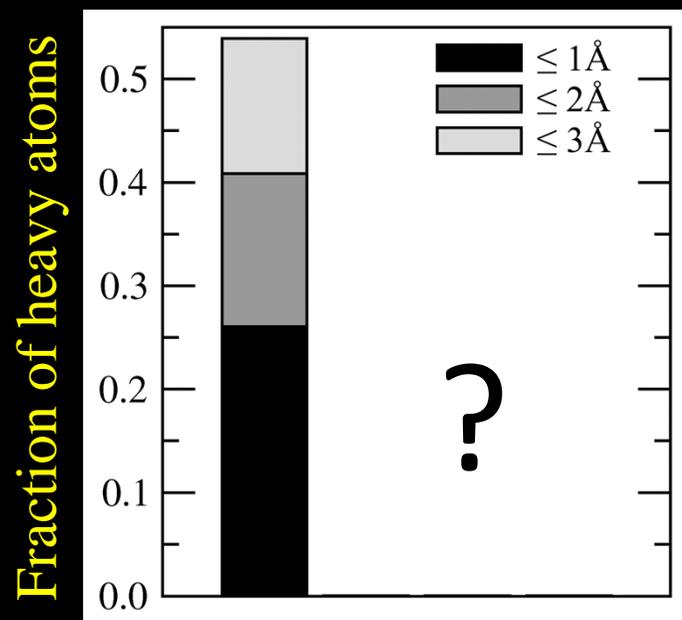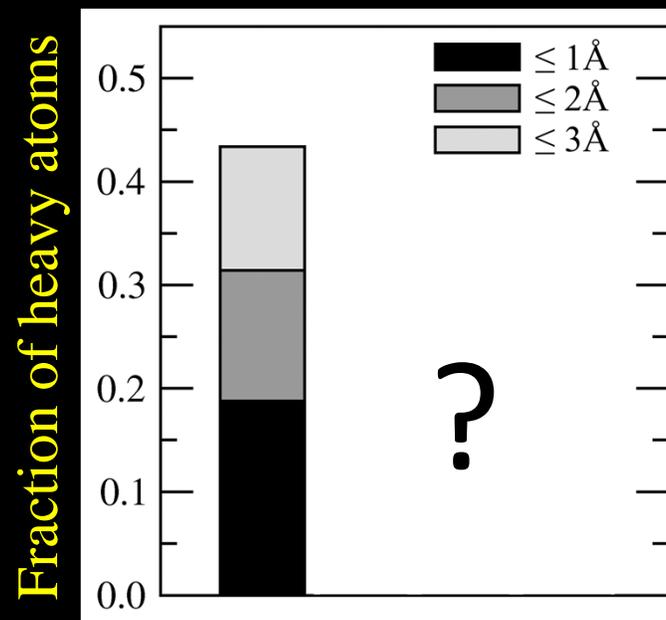
LIGAND HOMOLOGY MODELING: OVERVIEW

Template-based protein structure prediction

Template-based protein-ligand modeling

PDB

(meta) Threading
*PROSPECTOR, SP3, Sparks*

Structure templates

Ligand templates

Initial protein model

Binding site prediction
FINDSITE

Initial binding pose
FINDSITE^{LHM}

Specific potentials & restraints

Low-resolution refinement
TASSER

Low-resolution refinement
Q-Dock^{LHM}

Final protein model

Final binding pose
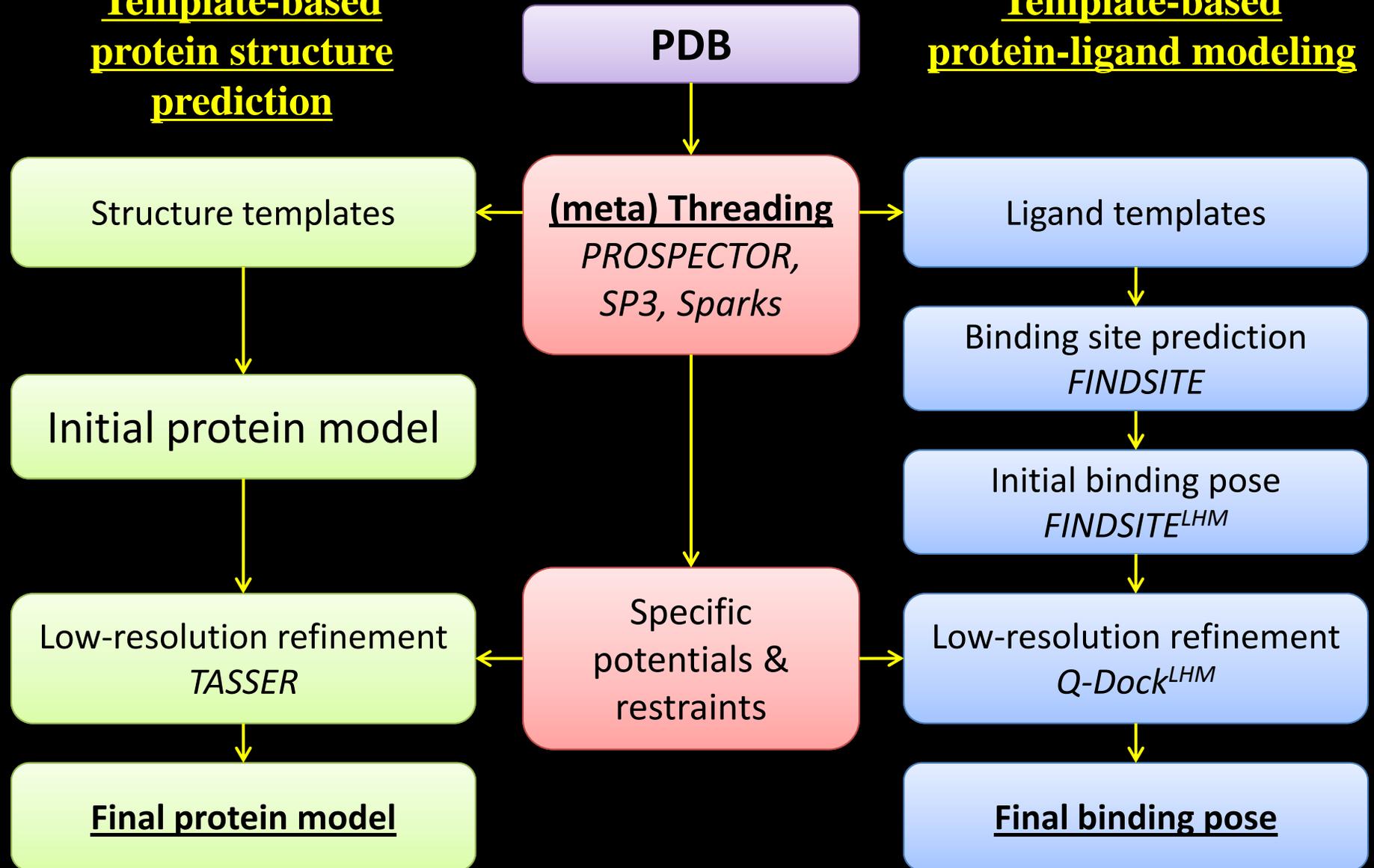
LIGAND HOMOLOGY MODELING: TECHNOLOGY & APPLICATION

FINDSITE

Ligand binding site prediction
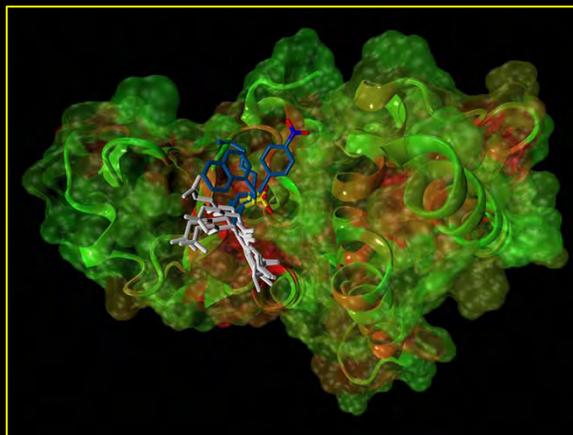
FINDSITE^{LHM}

Similarity-based ligand docking

Q-Dock^{LHM}
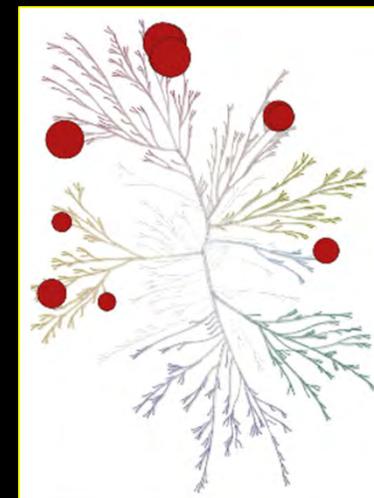
Low-resolution ligand docking/refinement
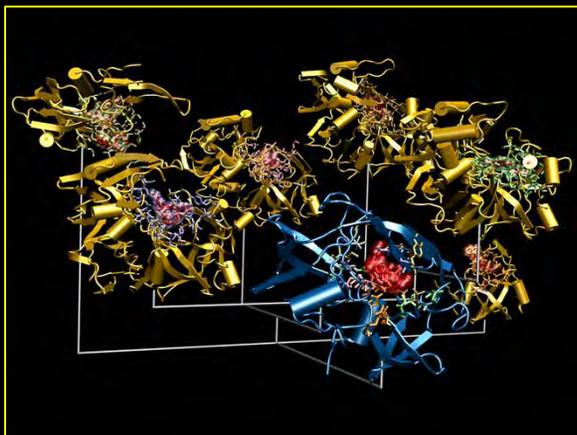
KINOME^{LHM}

Virtual screening of the human kinome

X-React^{KIN}

In silico drug profiling
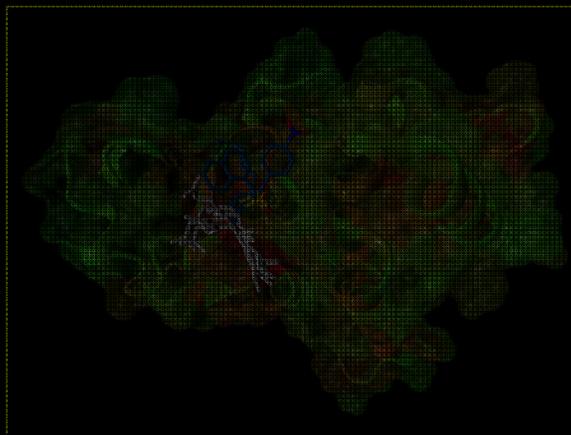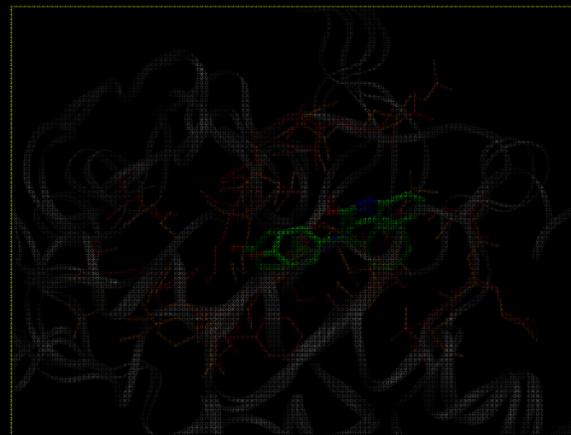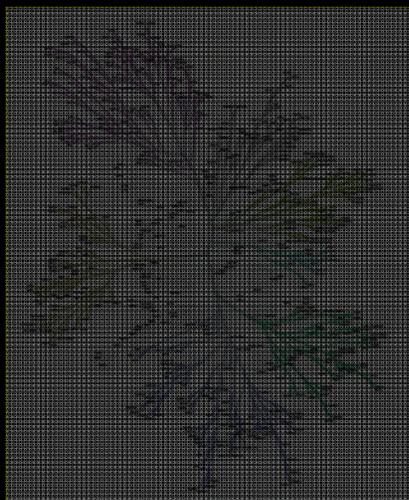
FINDSITE

Ligand binding site prediction

FINDSITE[LHM]

Similarity-based ligand docking
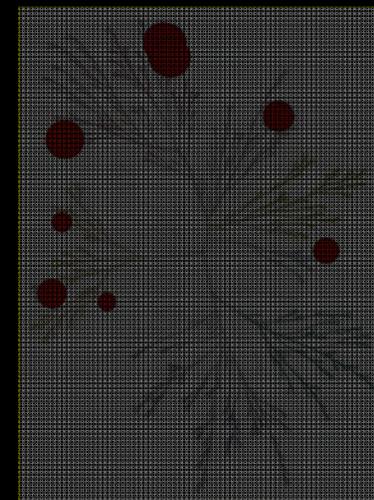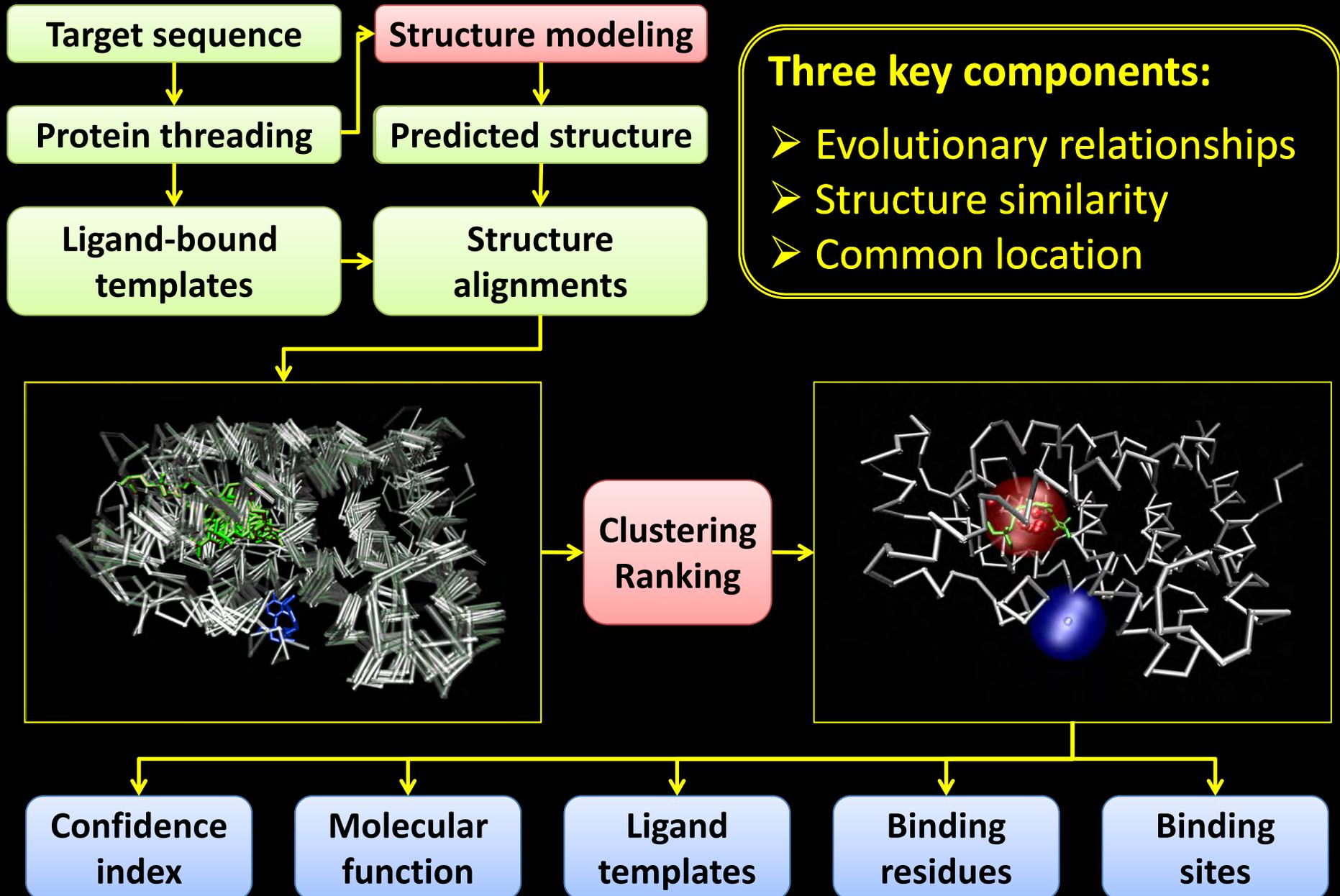
Q-Dock[LHM]

Low-resolution ligand docking/refinement

KINOME[LHM]

Virtual screening of the human kinome

X-React[KIN]

*In silico* drug profiling

Ligand Homology Modeling: FINDSITE

Target sequence → **Protein threading** → **Ligand-bound templates** → **Structure alignments**

Structure modeling → **Predicted structure** → **Structure alignments**

Clustering Ranking

Three key components:
- Evolutionary relationships
- Structure similarity
- Common location

Confidence index · **Molecular function** · **Ligand templates** · **Binding residues** · **Binding sites**

Benchmarks carried out for a set of 901 proteins:

• Crystal structures

FINDSITE is compared to **LIGSITE**^{CSC}

Crystal structures
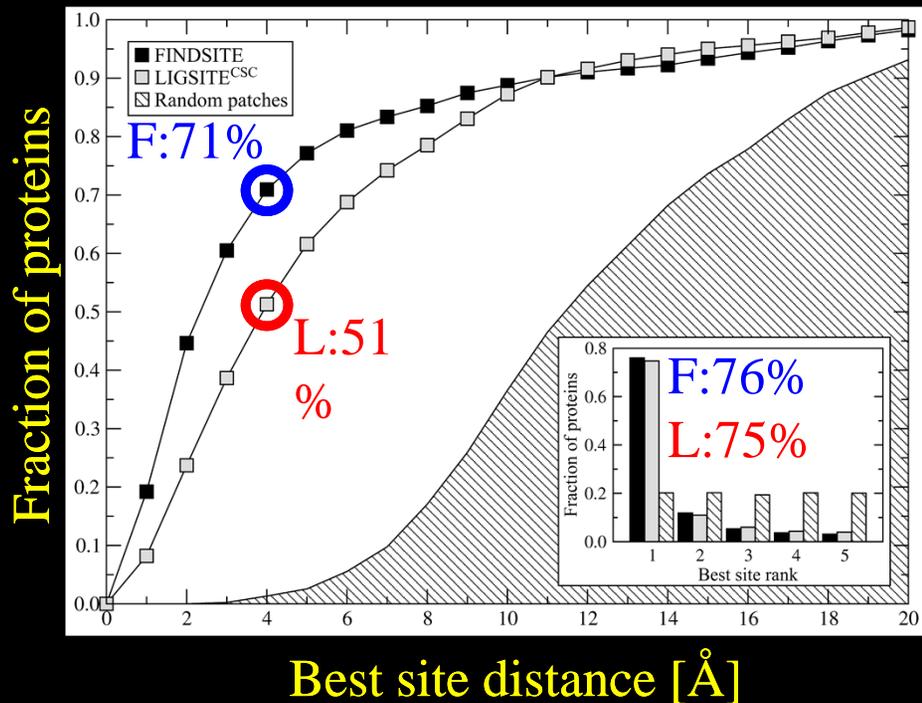
Best site distance [Å]

LIGAND HOMOLOGY MODELING: FINDSITE^LHM

Why does FINDSITE WORK?

FINDSITE
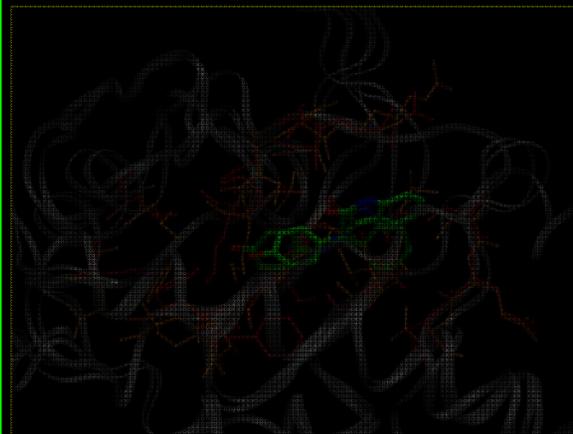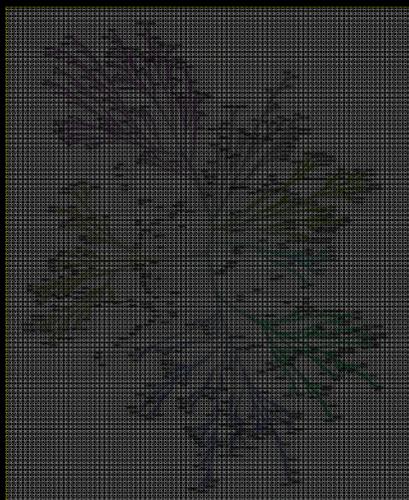

Ligand binding site prediction

FINDSITE[LHM]
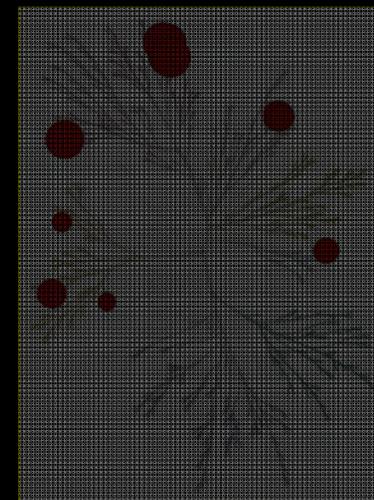

Similarity-based ligand docking

Q-Dock[LHM]

Low-resolution ligand docking/refinement
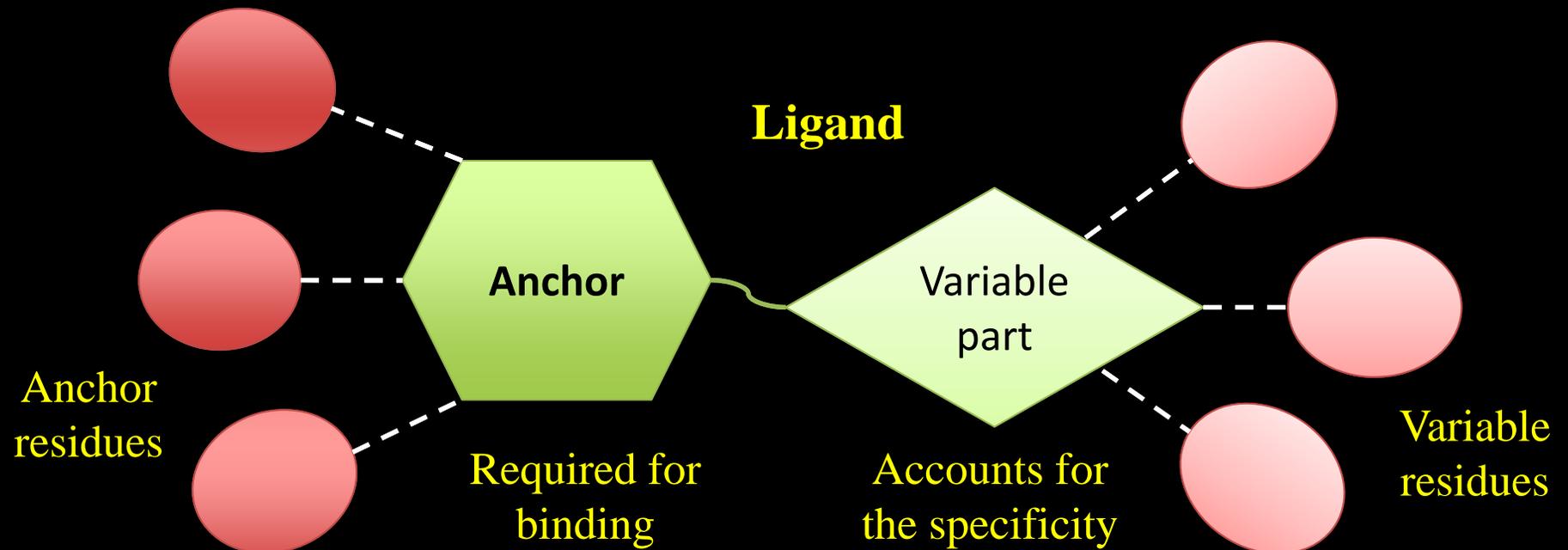
## KINOME[LHM]

Virtual screening of the human kinome

X-React[KIN]

In silico drug profiling

Anchor residues

Anchor

Ligand

Required for binding

Variable part

Accounts for the specificity

Variable residues

- Can an anchor be identified in ligands bound to evolutionarily related proteins?

- What is the sequence/structure conservation of the anchor and variable regions?

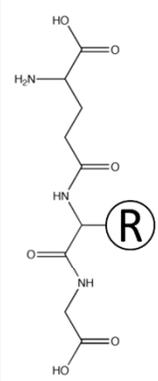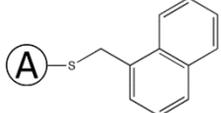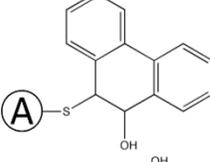- Can the consensus binding mode be used for ligand docking into the predicted pockets?
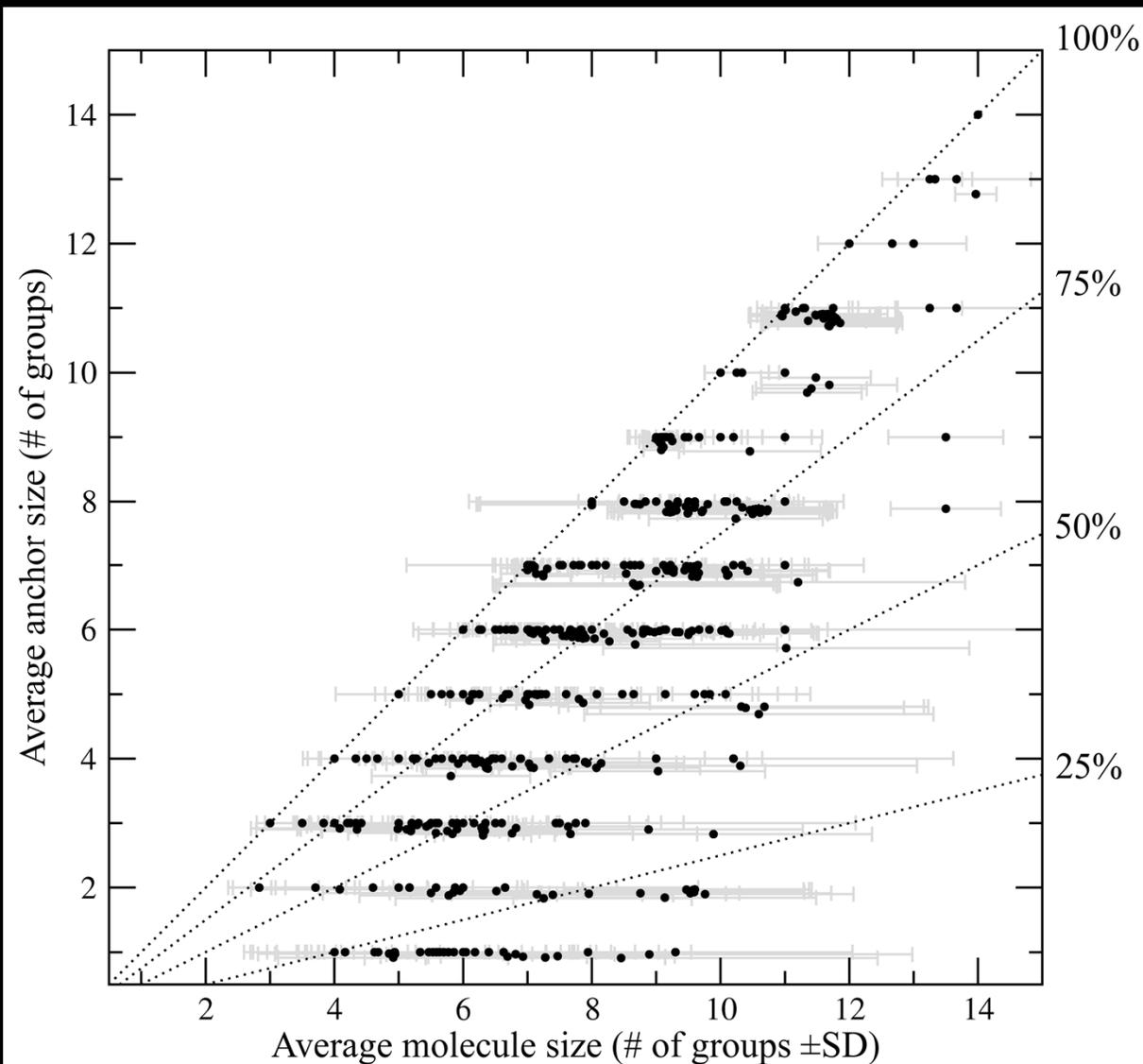
Average molecule size vs. average anchor size

Inset: average pairwise RMSD of the anchor groups

Glutathione S-transferase from *E. Coli* complexed with glutathionesulfonic acid (PDB-ID: 1a0f)

Variable parts (extracted from remote templates)

LIGAND HOMOLOGY MODELING: FINDSITE^{LHM}

Glutathione S-transferase from *E. Coli* complexed with glutathionesulfonic acid (PDB-ID: 1a0f)

Conserved substructure – white, **Variable region** – black
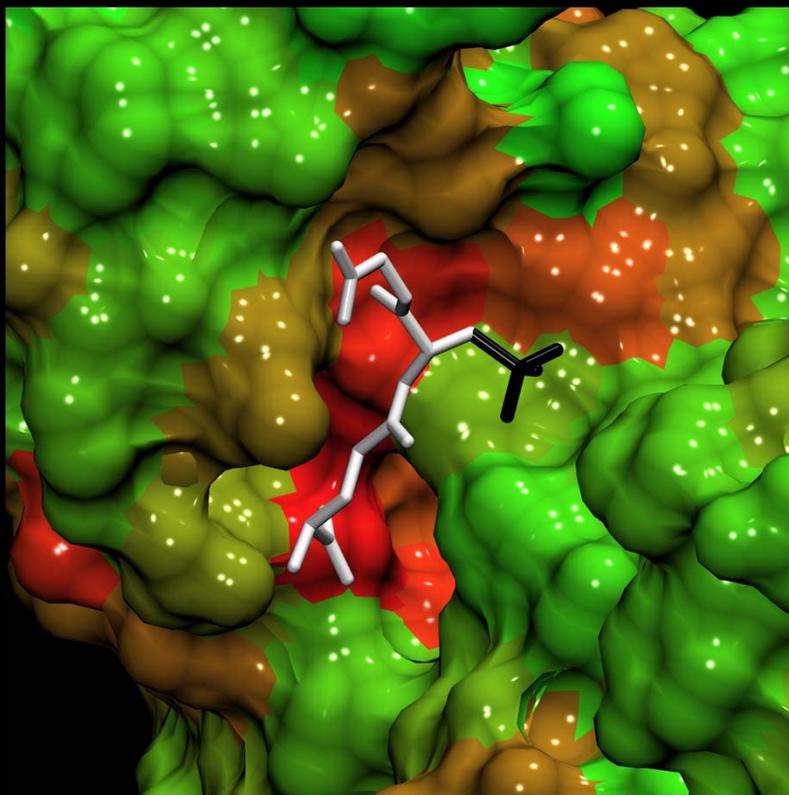
Sequence entropy
(red – low, green – high)
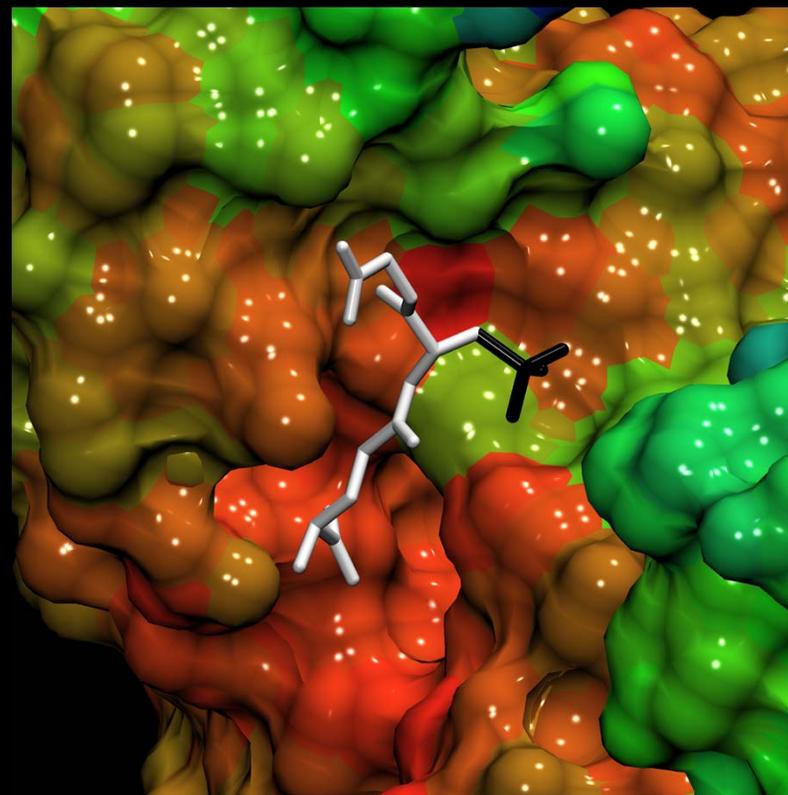
Experimental B-factors
(red – low, green – high)

FINDSITE^{LHM}

- A fast, <u>similarity-based</u> docking approach
- Uses conserved common ligand substructures
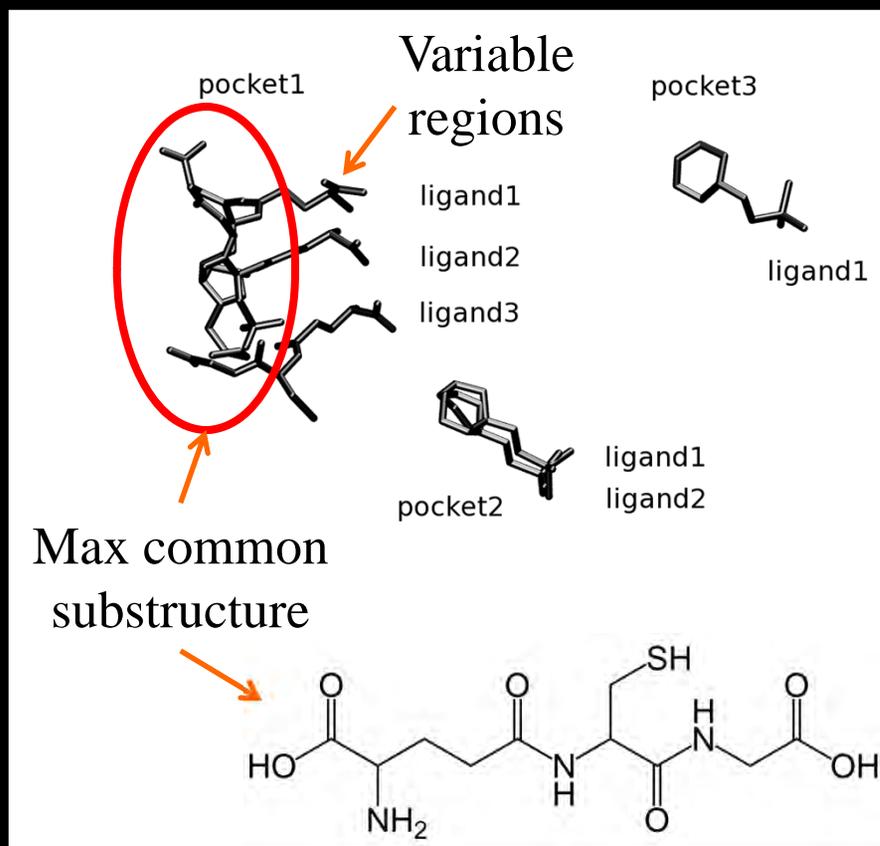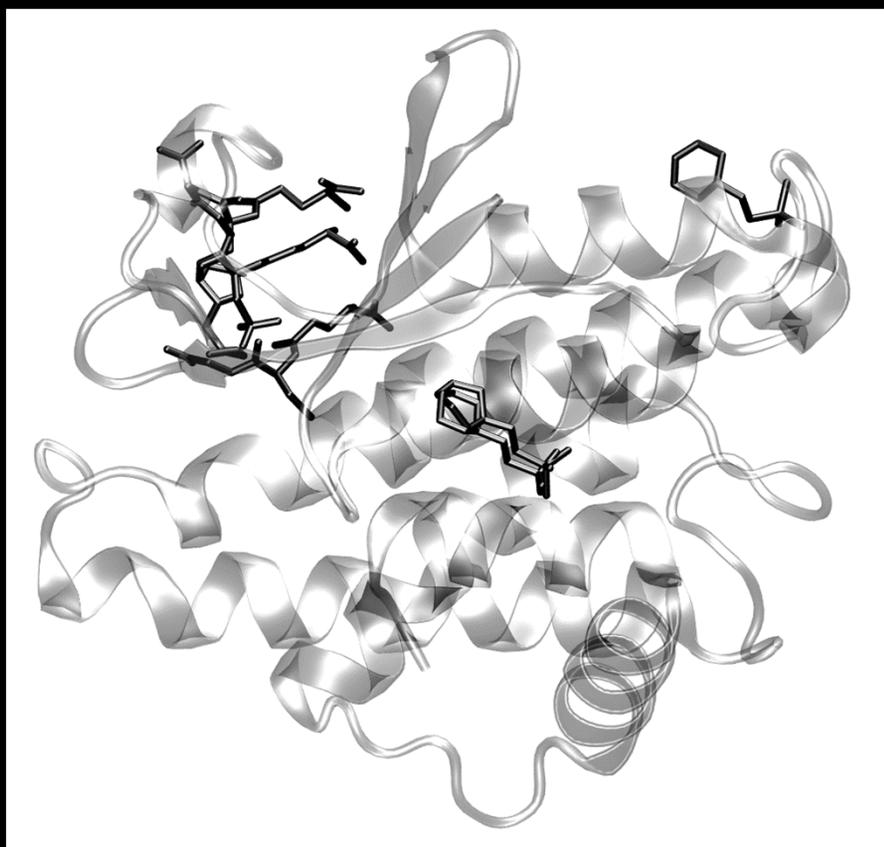
LIGAND HOMOLOGY MODELING: FINDSITE[LHM]

Glutathione S-transferase from *E. Coli* (PDB-ID: 1a0f)

Similarity-based docking by FINDSITE[LHM]
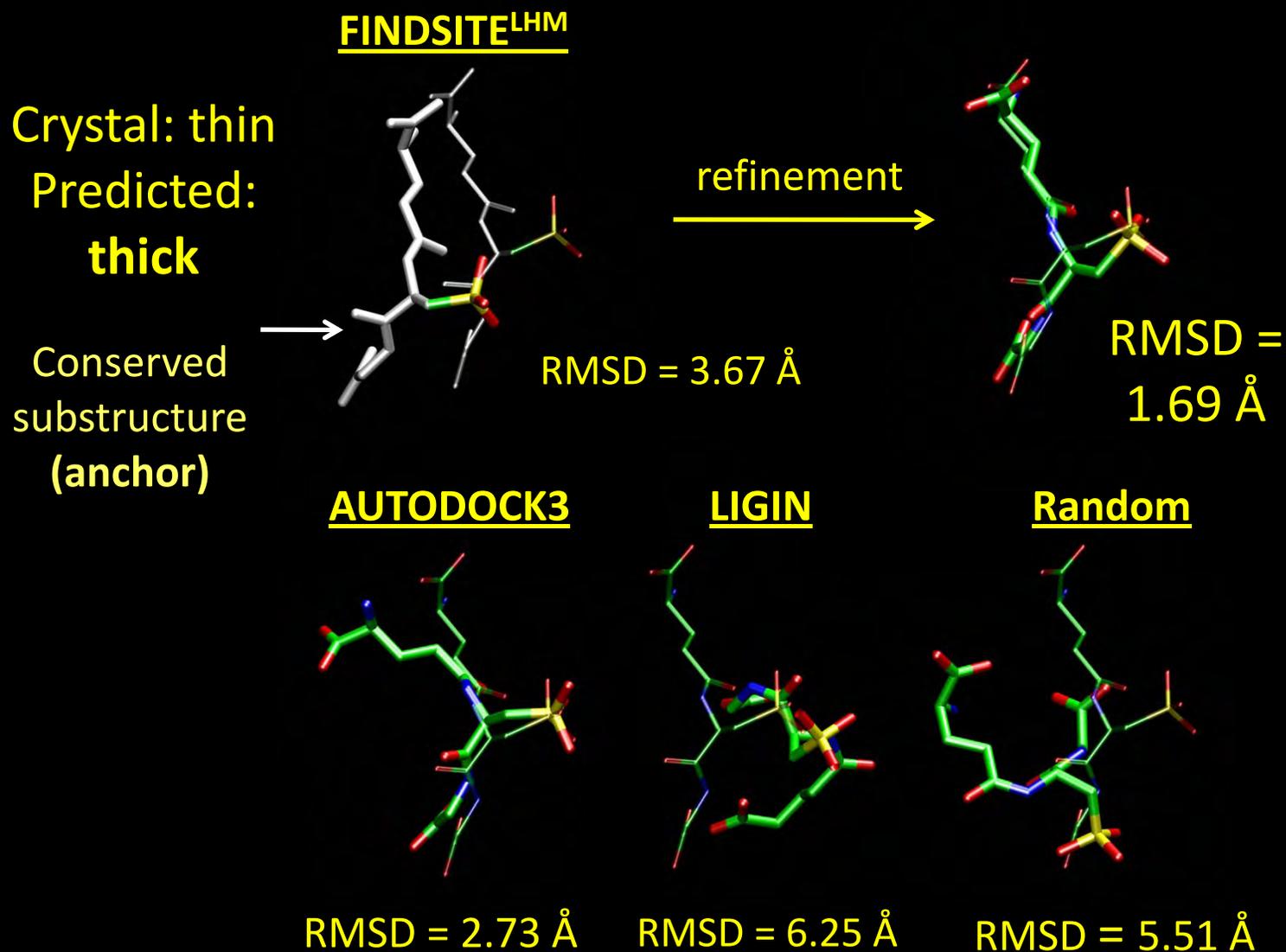
LIGAND HOMOLOGY MODELING: Q-DOCK^{LHM}

FINDSITE
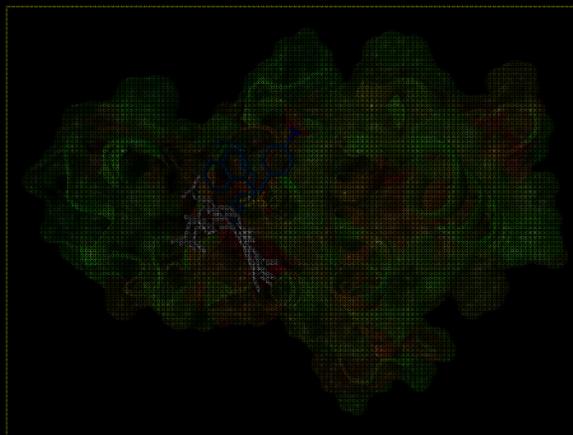
FINDSITE^{LHM}

Q-Dock^{LHM}

Ligand binding site prediction

Similarity-based ligand docking
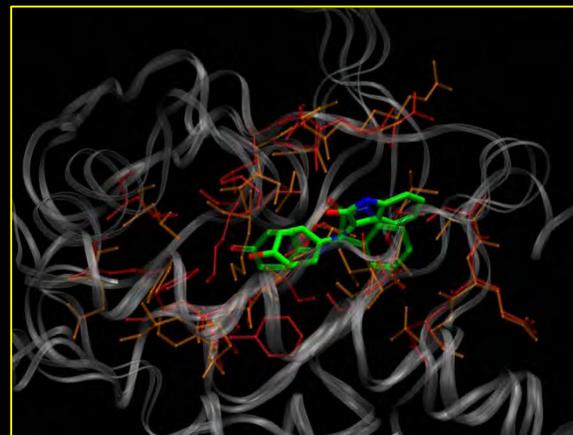
Low-resolution ligand docking/refinement

KINOME^{LHM}

X-React^{KIN}

Virtual screening of the human kinome

In silico drug profiling
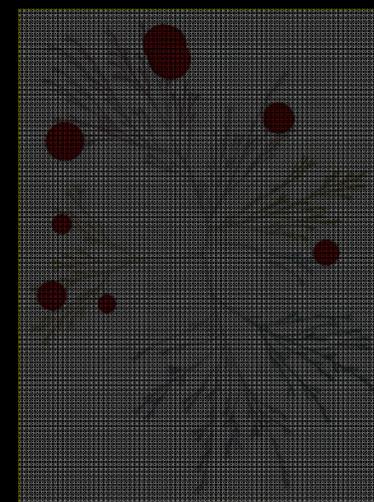
LIGAND HOMOLOGY MODELING: Q-DOCK[LHM]

Q-Dock[LHM]:
Low-resolution modeling

RECEPTOR: Cα and side chain centers of mass
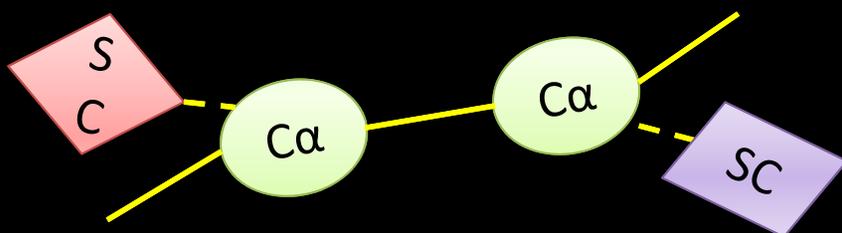
LIGANDS: quasichemical building blocks (functional groups)
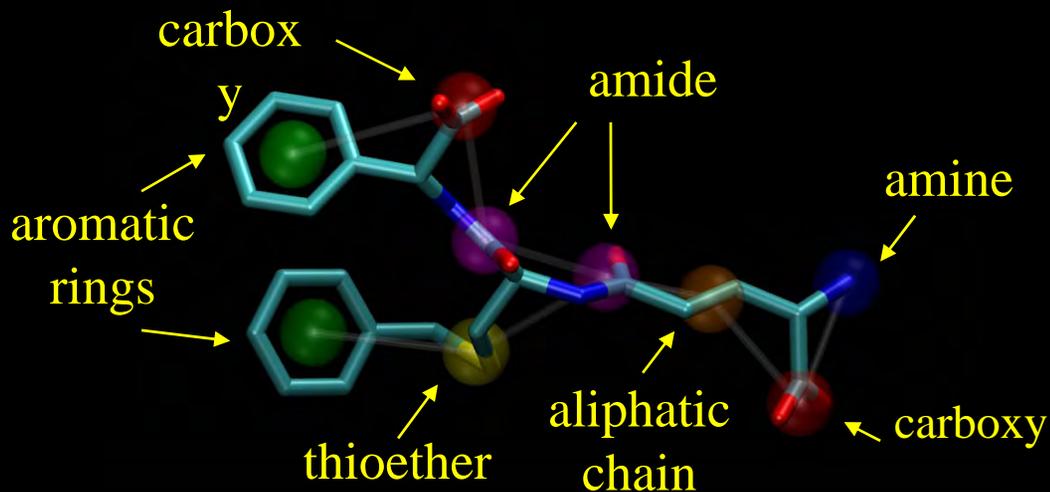
Description	Symbol/formula
Aromatic rings	mono-, heterocyclic
Ether	–C–O–C–
Thioether	–C–S–C–
Carbonyl	>C=O
Thiocarbonyl	>C=S
Halogene	–Cl; –Br; –F; –I
Guanidine	$-NHC(NH_2)NH$
Amide	–CONH–
Carboxyl	–COOH
Amine (1°,2°,3°)	$-NH_2$; >NH; >N–
Phosphate	$-PO_4$
Sulphate	$-SO_4$
Nitro group	$-NO_2$
Metals	Fe; Zn; Mg; Ca
Hydroxyl group	–OH
Thiol group	–SH
Aliphatic chain	$-(C–C)_x-$

Q-Dock^{LHM}

Docking accuracy

Ligand ranking

➢ Combined knowledge-based potentials:

- Generic energy terms (derived from PDB)
- Harmonic restraints on conserved ligand groups
- Pocket specific contact potential
- Optimized weight factors

➢ Replica Exchange Monte Carlo sampling
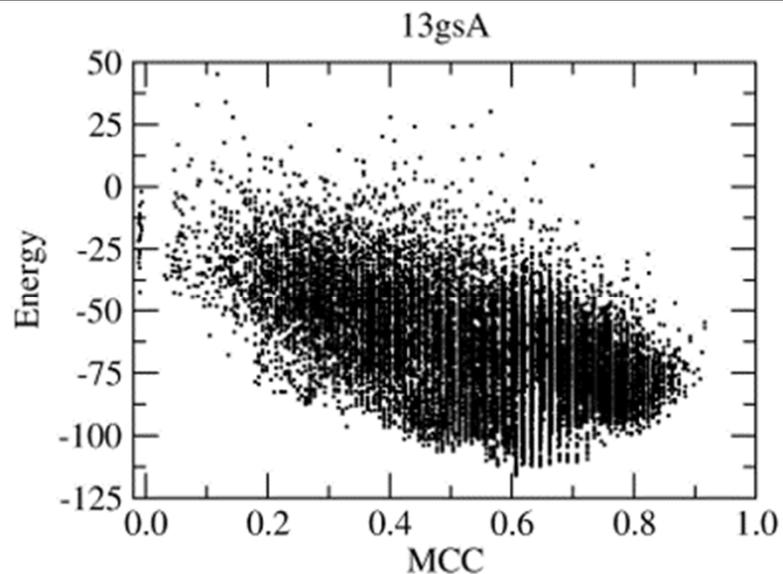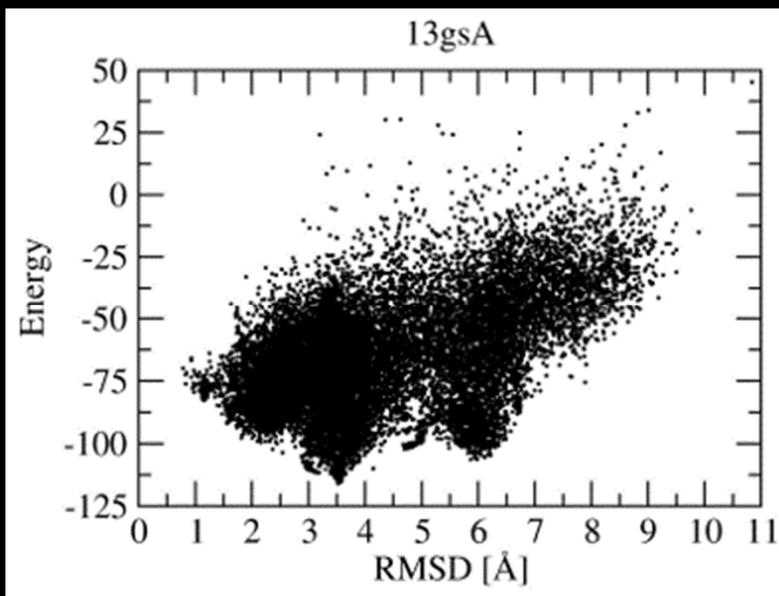
204 pharmacologically relevant targets from CCDC/Astex dataset
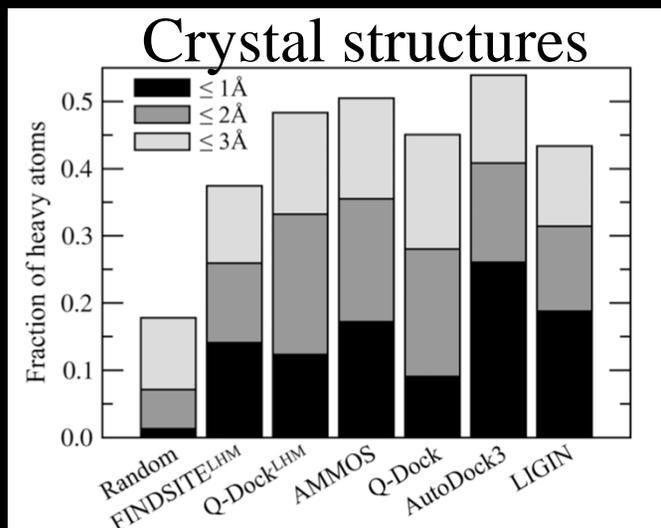
Average fraction of ligand heavy atoms predicted within 1, 2 and 3 Å
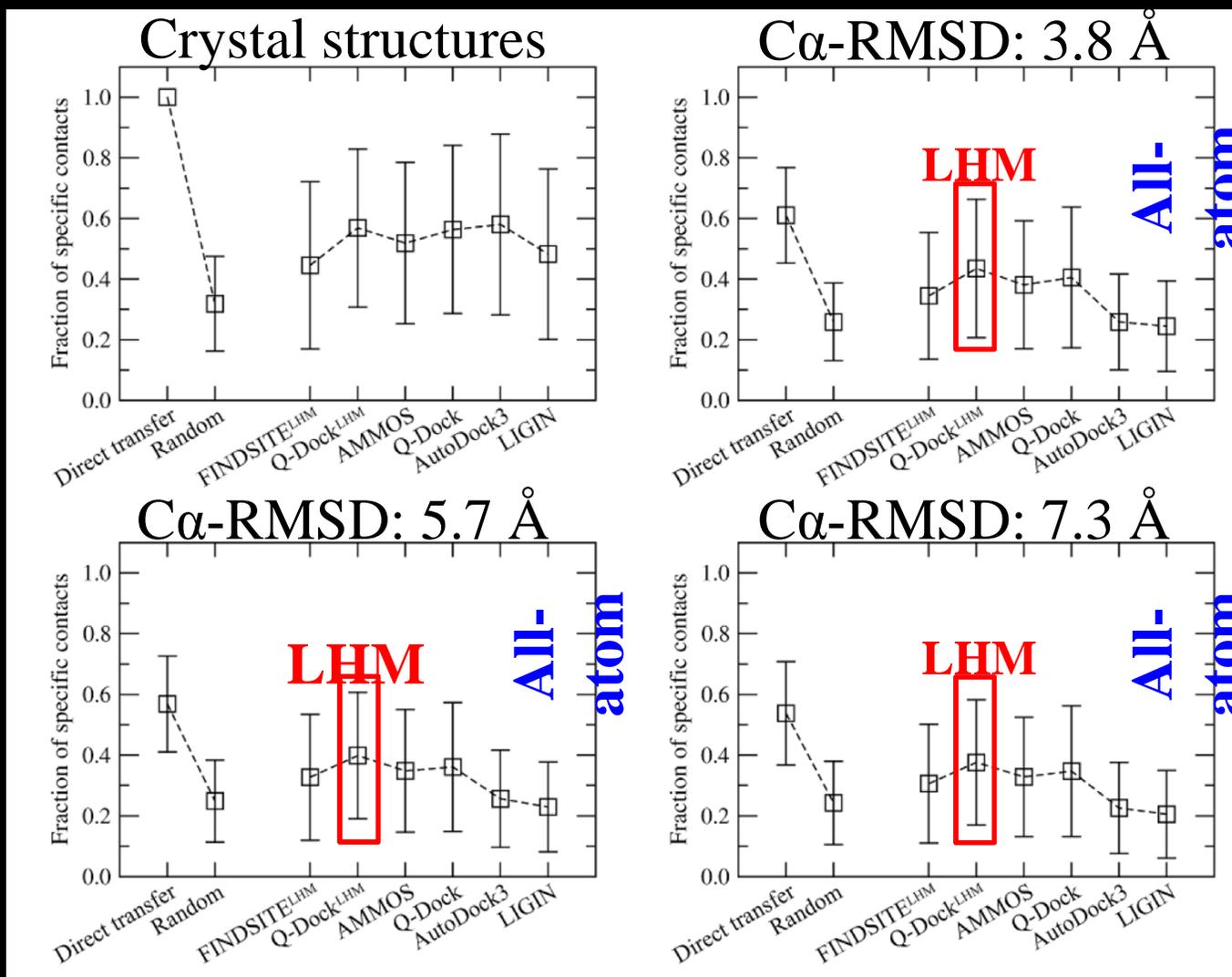
Average fraction of correctly predicted specific protein-ligand contacts

Docking times on 2.0 GHz AMD Opteron processor
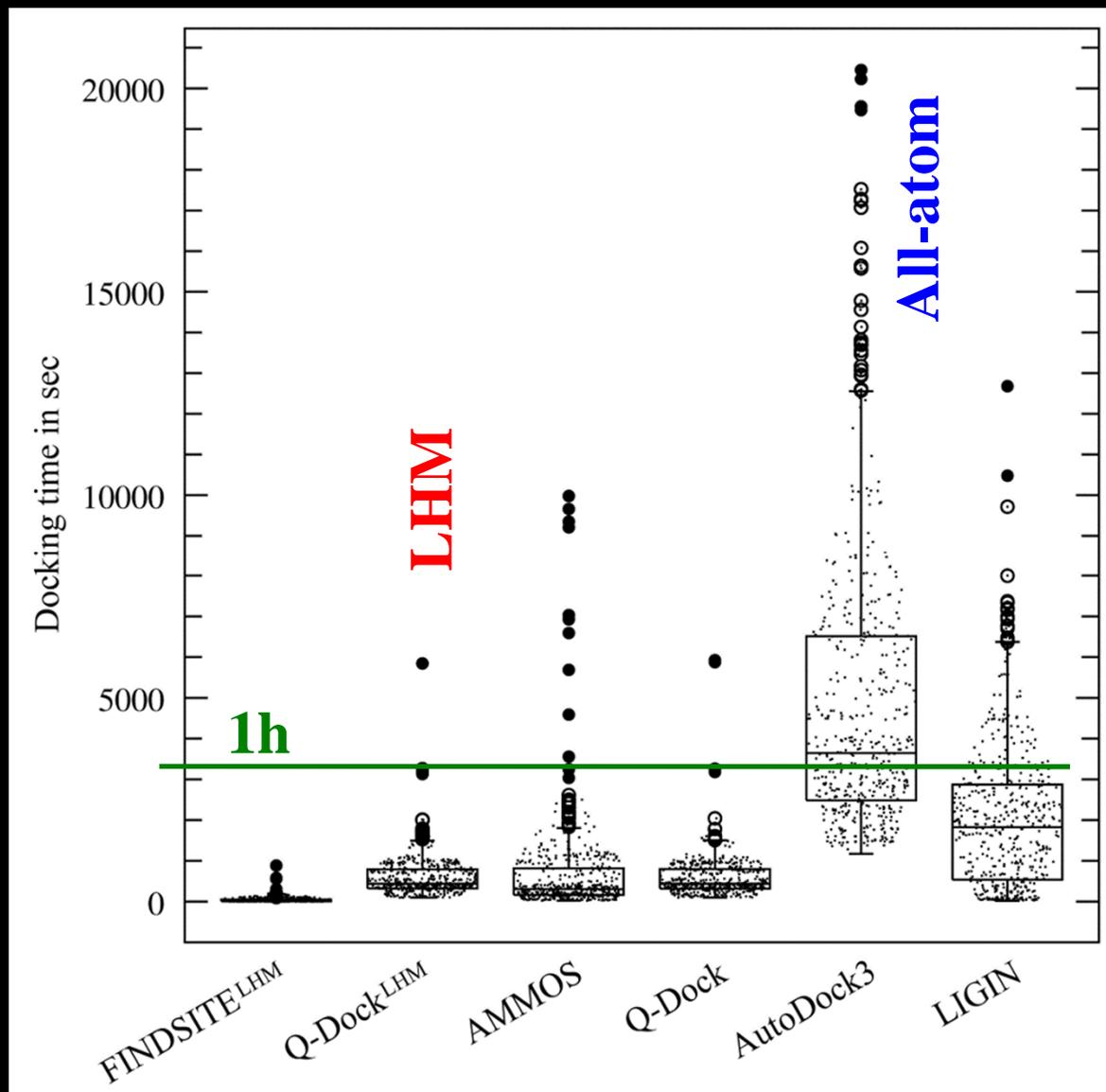
LHM is more accurate and much faster

For AMMOS, AutoDock3 and LIGIN, the default sets of parameters were used and the docking protocols have not been optimized with respect to the accuracy and simulation time.

LIGAND HOMOLOGY MODELING: KINOME^{LHM}

FINDSITE
Ligand binding site prediction
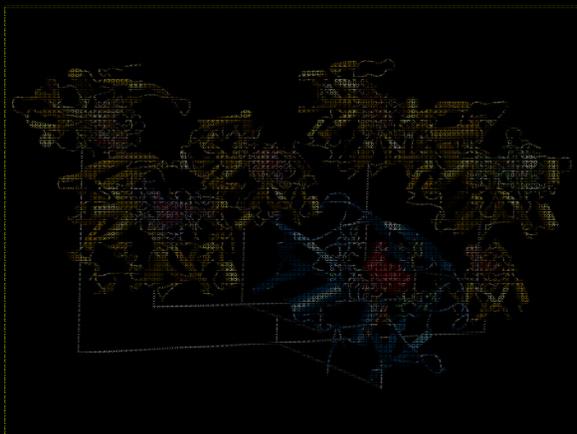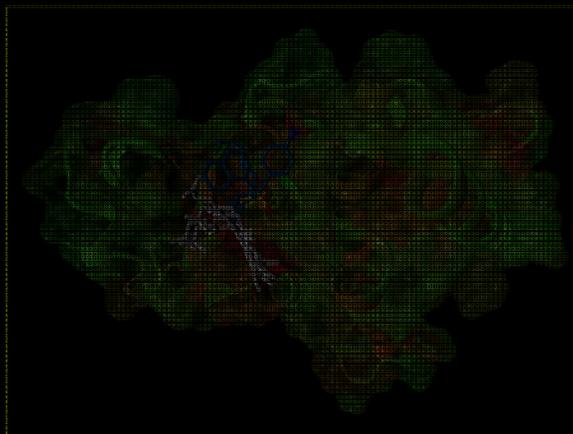
FINDSITE^{LHM}
Similarity-based ligand docking

Q-Dock^{LHM}
Low-resolution ligand docking/refinement
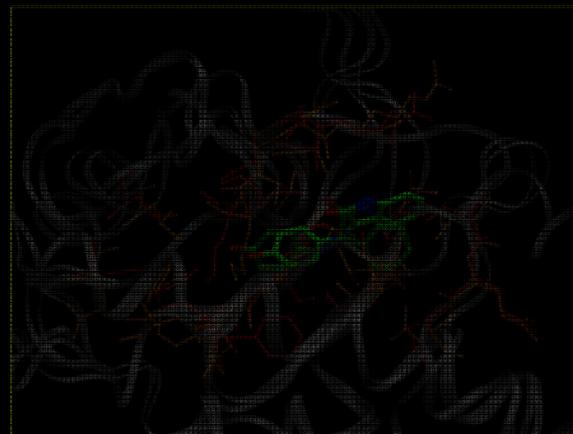
KINOME^{LHM}
Virtual screening of the human kinome

X-React^{KIN}
In silico drug profiling
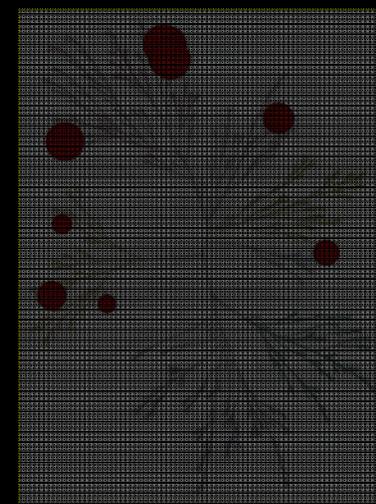
Availability of high-resolution crystal structures for the human kinome
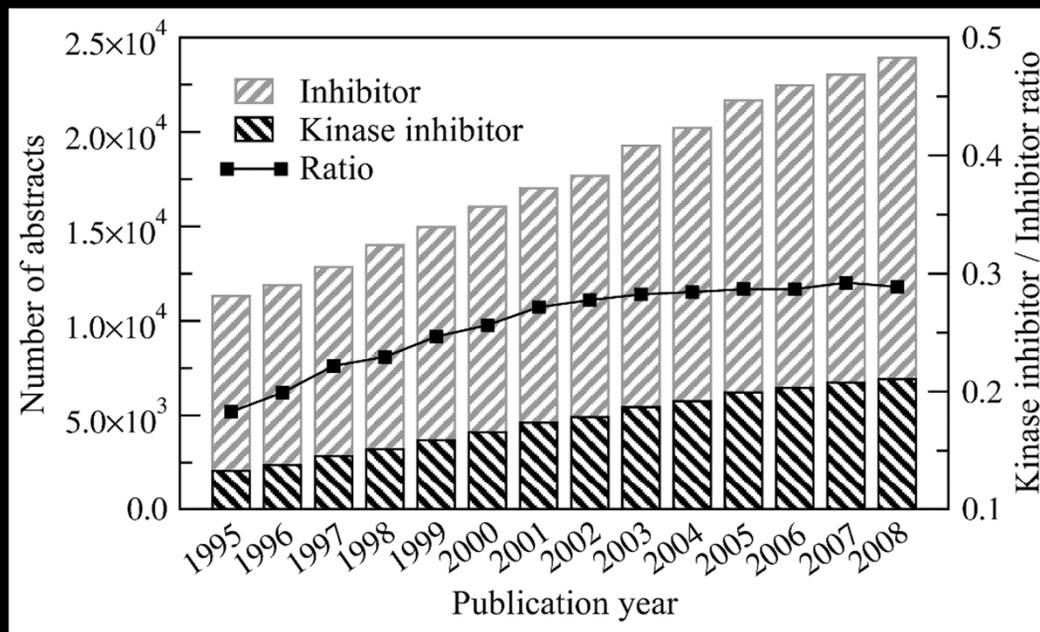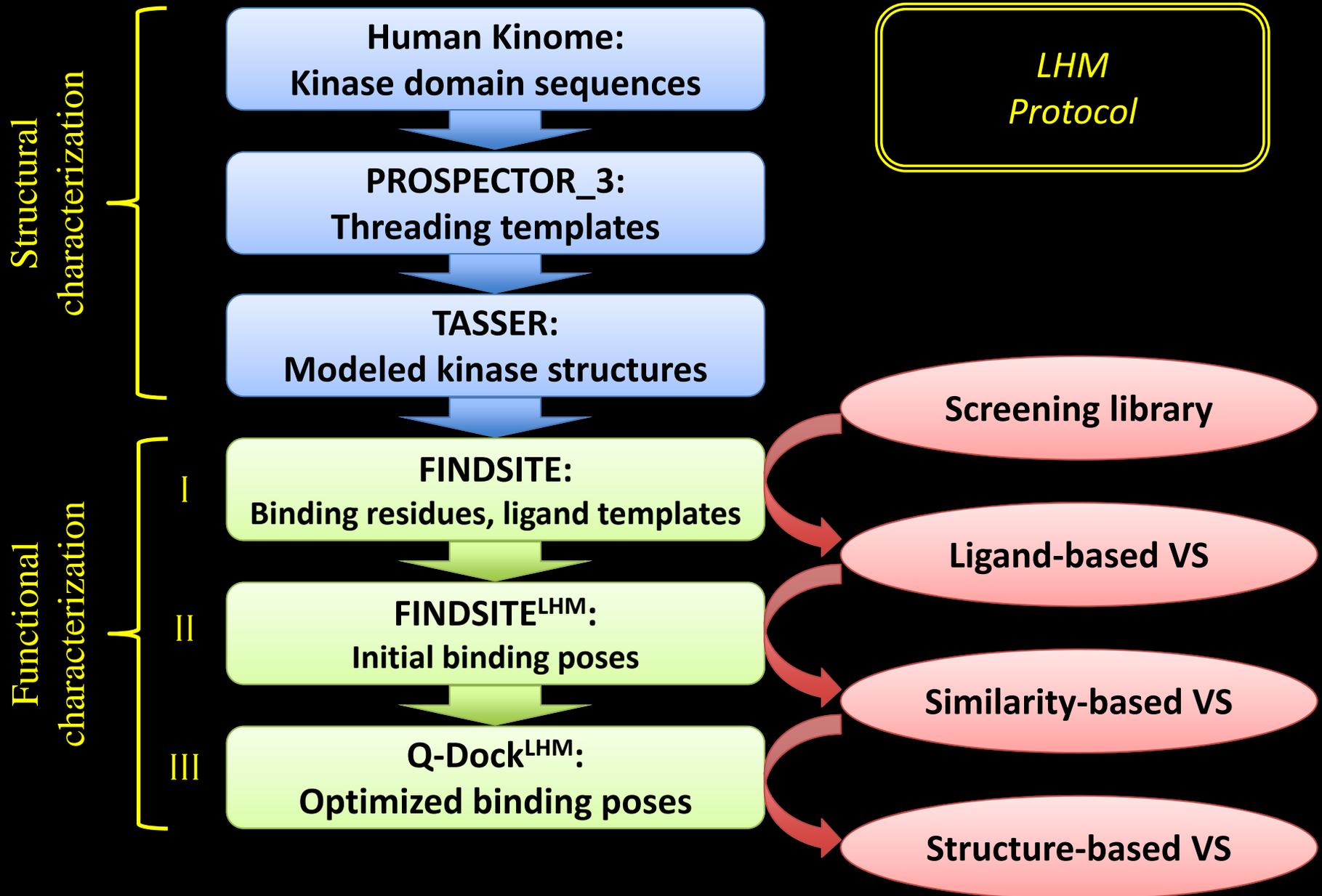
LIGAND HOMOLOGY MODELING: KINOME^{LHM}

Structural characterization

- Human Kinome: Kinase domain sequences
- PROSPECTOR_3: Threading templates
- TASSER: Modeled kinase structures

Functional characterization

- I. FINDSITE: Binding residues, ligand templates
- II. FINDSITE^{LHM}: Initial binding poses
- III. Q-Dock^{LHM}: Optimized binding poses

LHM Protocol

- Screening library
- Ligand-based VS
- Similarity-based VS
- Structure-based VS

LIGAND HOMOLOGY MODELING: KINOME^LHM

516
kinase structures

2,095,759
ZINC7 compounds
http://zinc.docking.org

108,141,164
fingerprint scores for kinase-drug pairs

5,134,483
predicted ligand binding poses, structure-based scores

Simulation time
150-200 years on a single core

Structure modeling of kinase domains
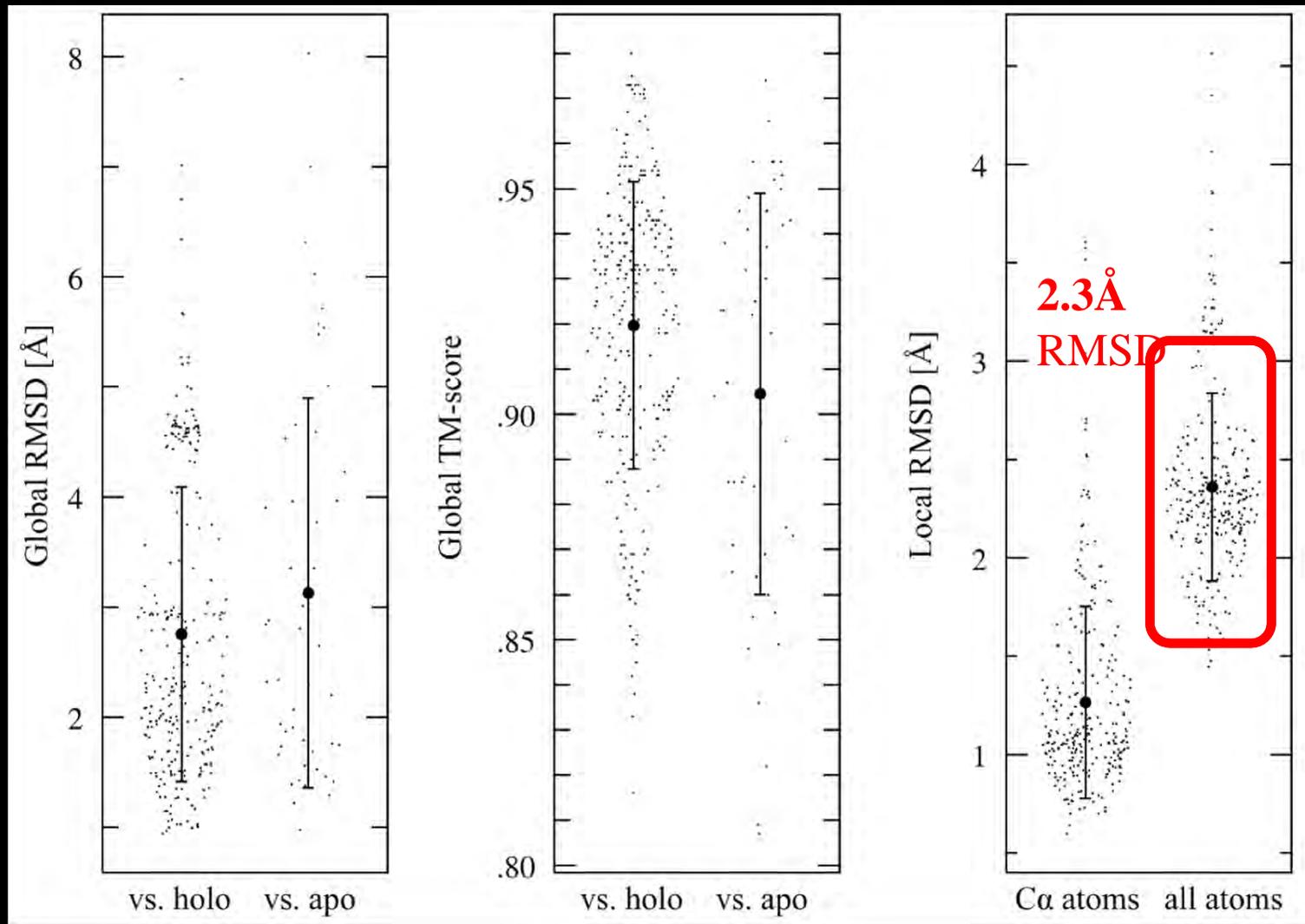
Binding site/residue prediction
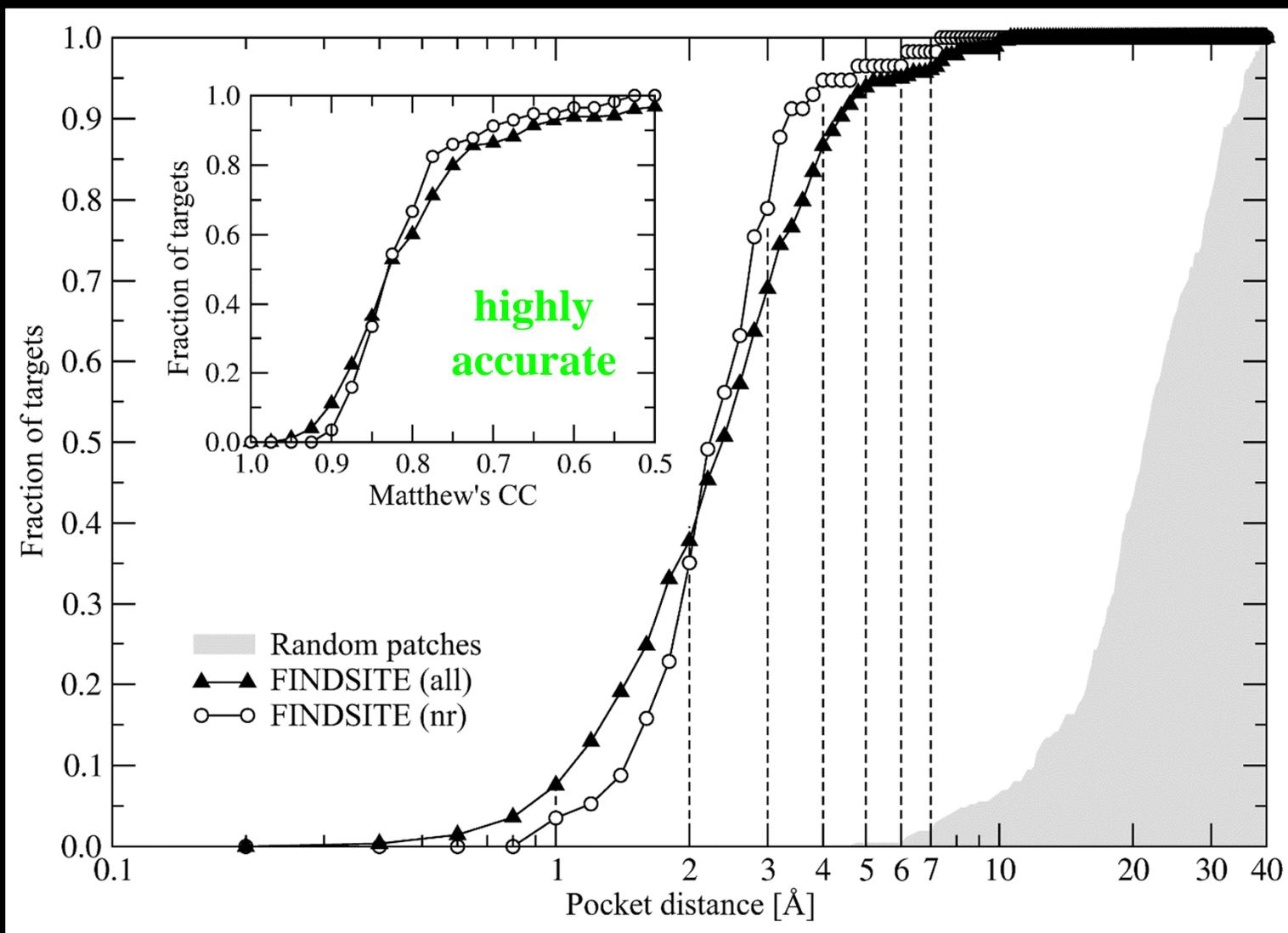
Docking accuracy for kinase inhibitors

Non-specific contacts
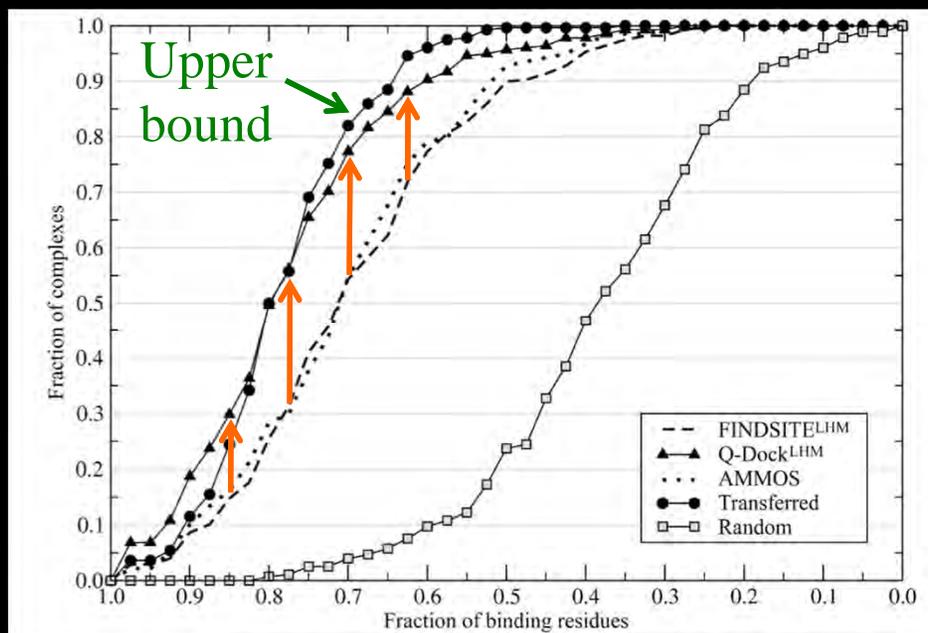
Low-resolution refinement by Q-Dock^{LHM} improves binding poses over FINDSITE^{LHM}

Virtual screening benchmarks using **362** kinase inhibitors from BindingDB (*http://www.bindingdb.org*)
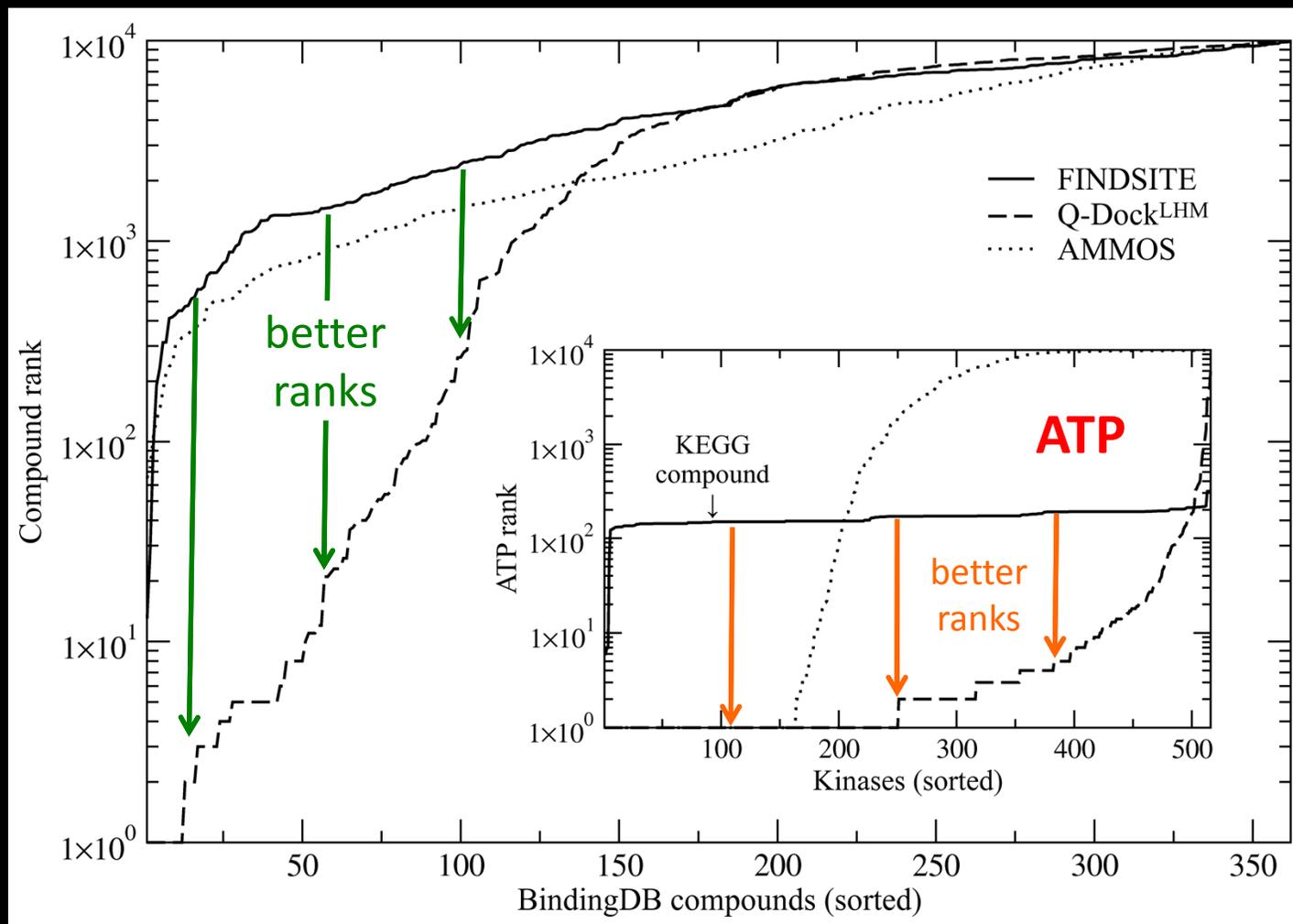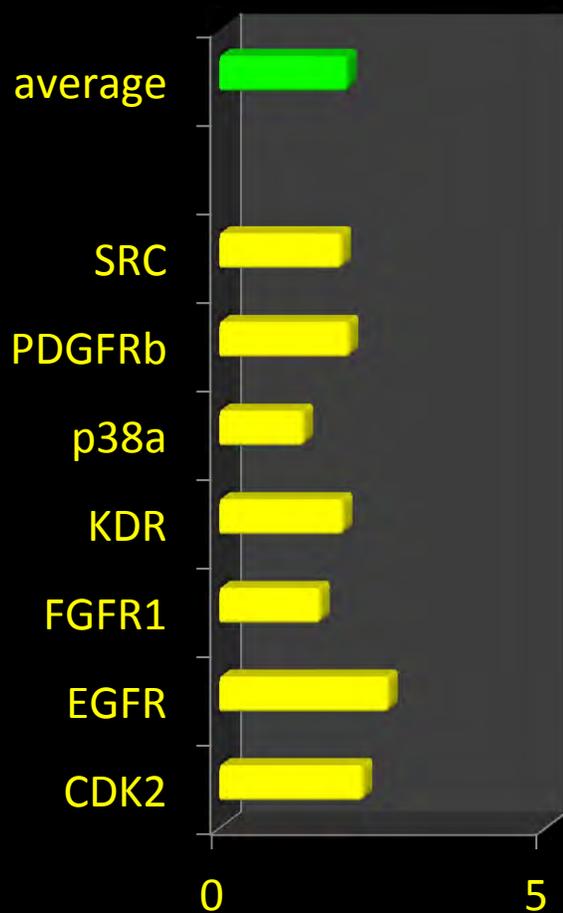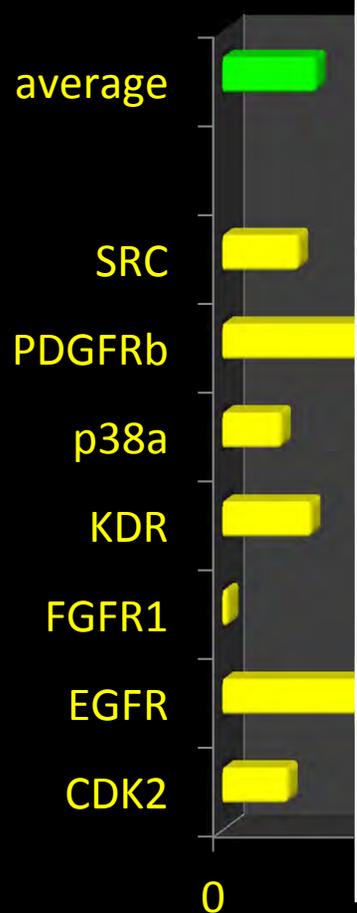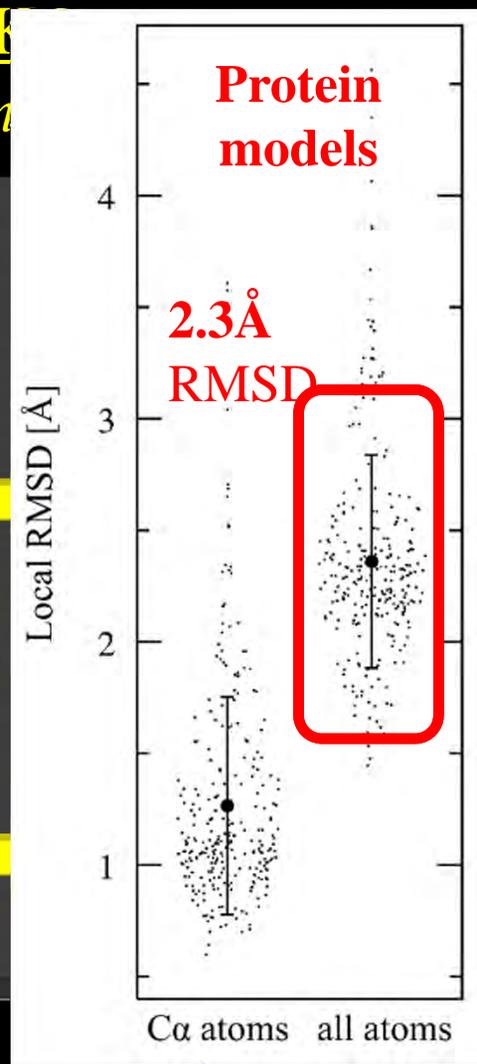
LIGAND HOMOLOGY MODELING: X-React^KIN

FINDSITE
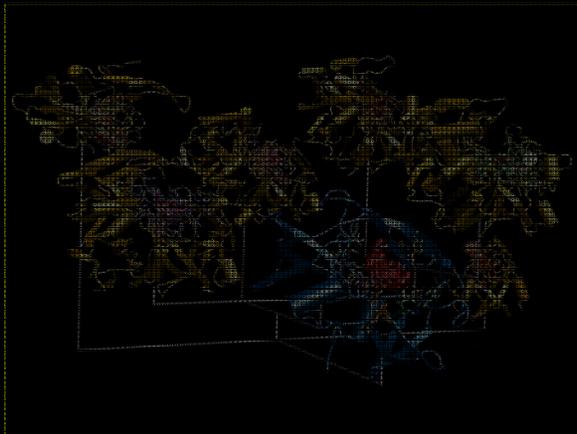
Ligand binding site prediction
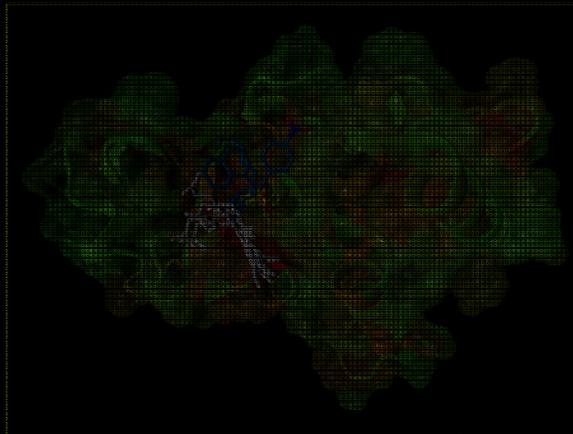
FINDSITE^LHM

Similarity-based ligand docking

Q-Dock^LHM
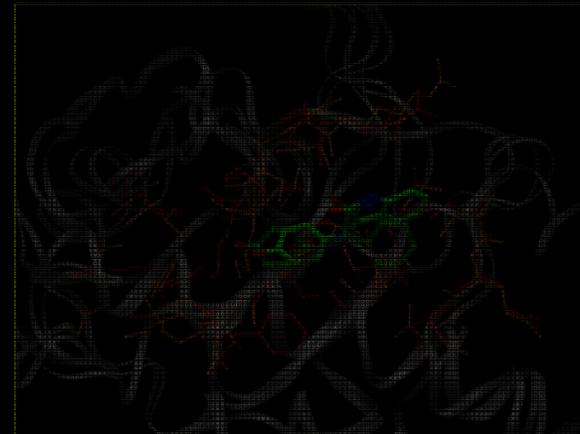
Low-resolution ligand docking/refinement

KINOME^LHM

Virtual screening of the human kinome

X-React^KIN

In silico drug profiling
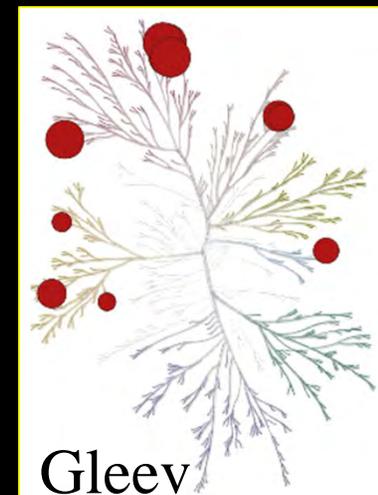
Gleev

LIGAND HOMOLOGY MODELING: X-REACT^KIN
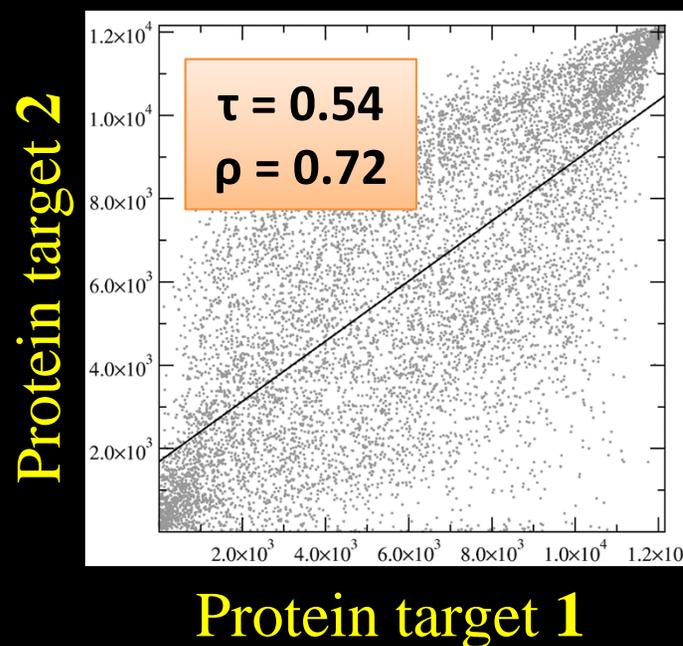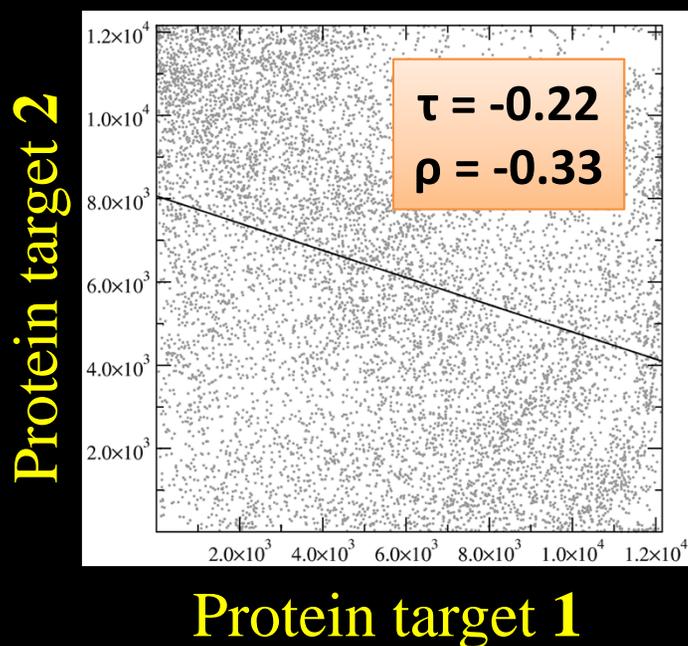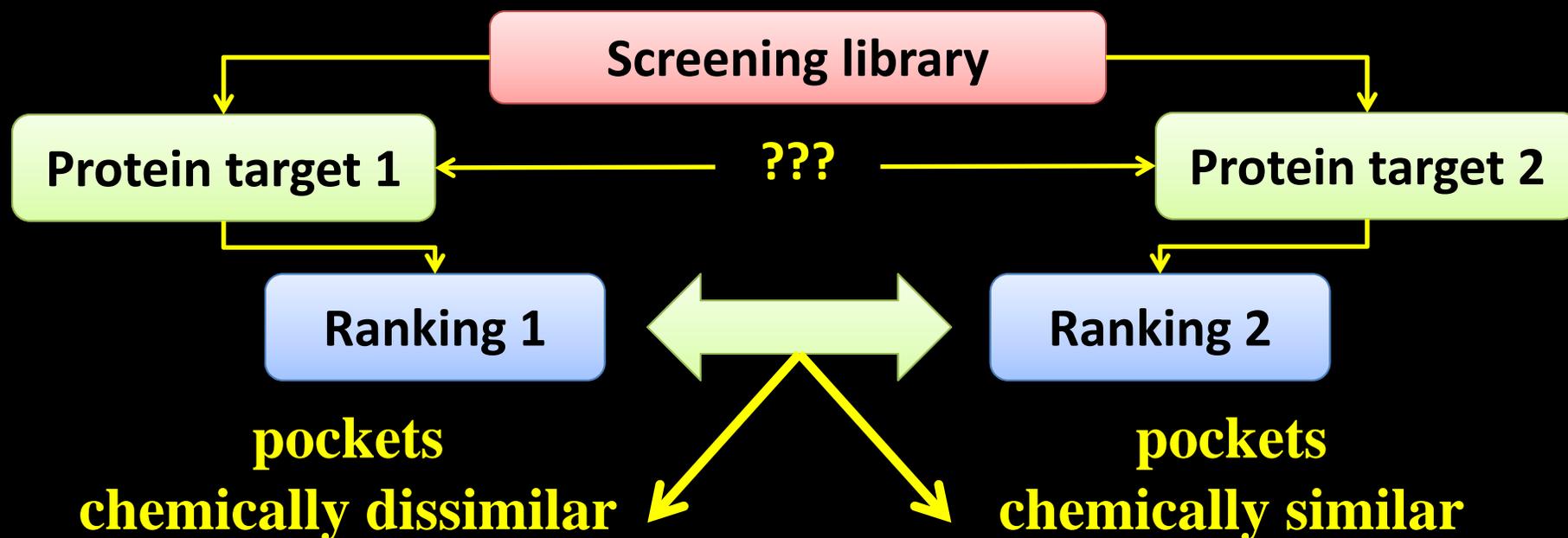
LIGAND HOMOLOGY MODELING: X-REACT^KIN

Sequence-based
Pocket SID
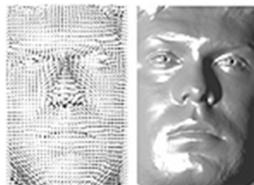
Structure-based
Geometric hashing
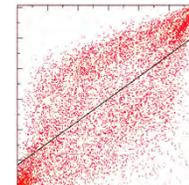
Ligand-based
Chemical correlation

Human Kinome^LHM
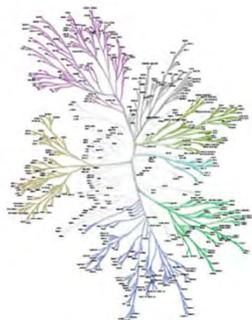
X-React^KIN
Machine learning

Bioassay data
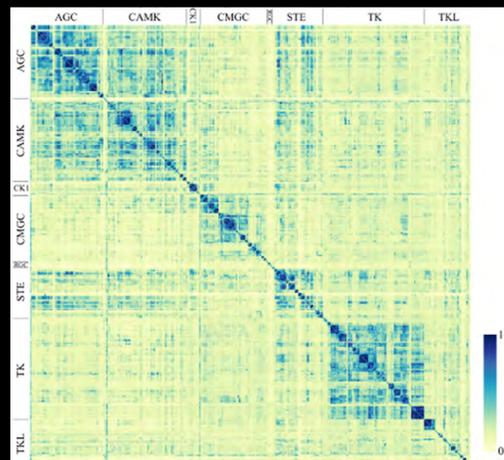
Off-target identification

Prediction of inhibitor cross-reactivity for the human kinome

21 (#1), 112 (#2), 275 (#3) kinases
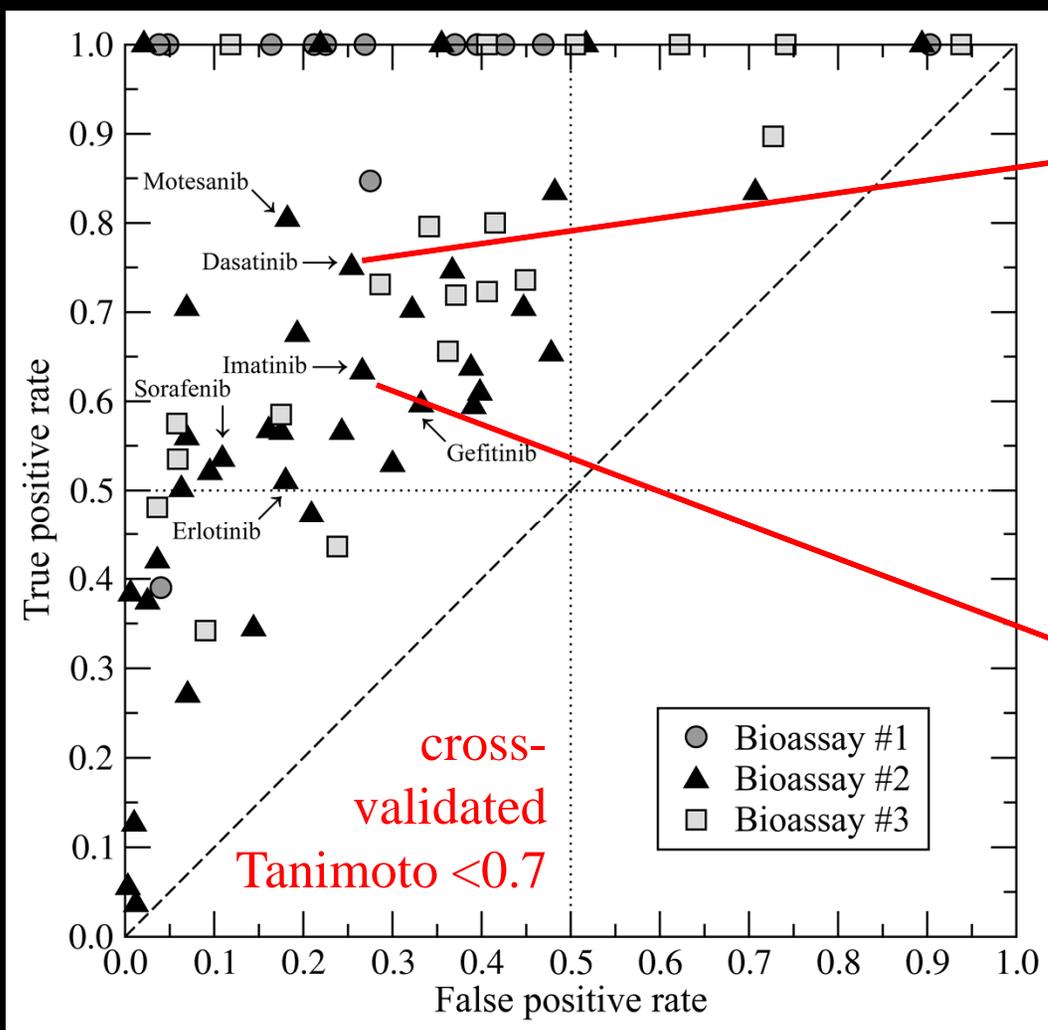
Dasatinib
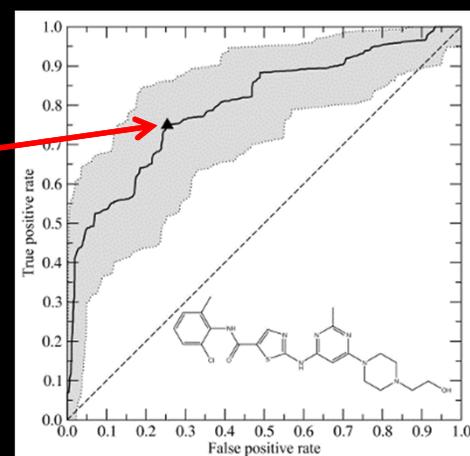
Imatinib
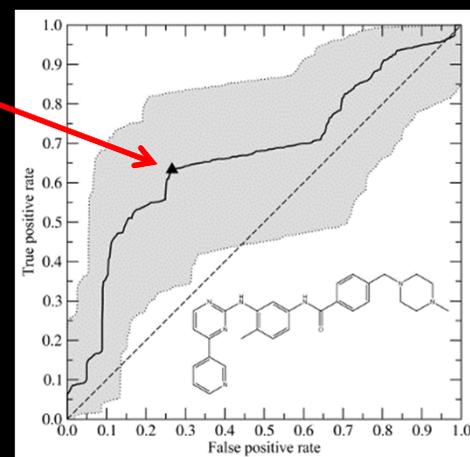
LIGAND HOMOLOGY MODELING: X-REACT^{KIN}

Comparison to experimental SAR profiles

577 cmps
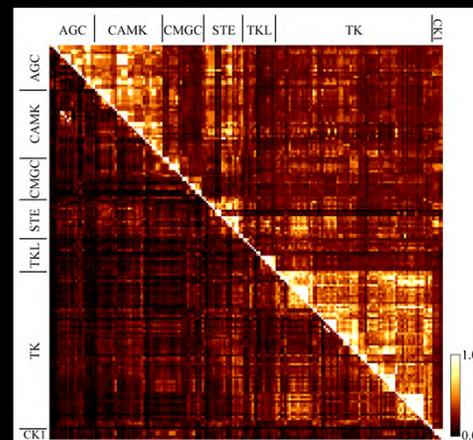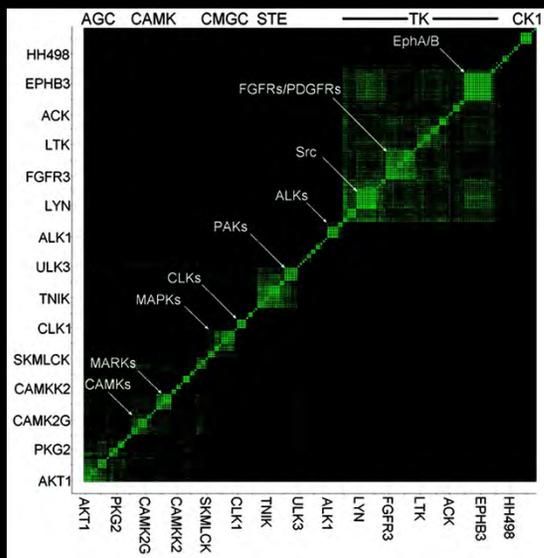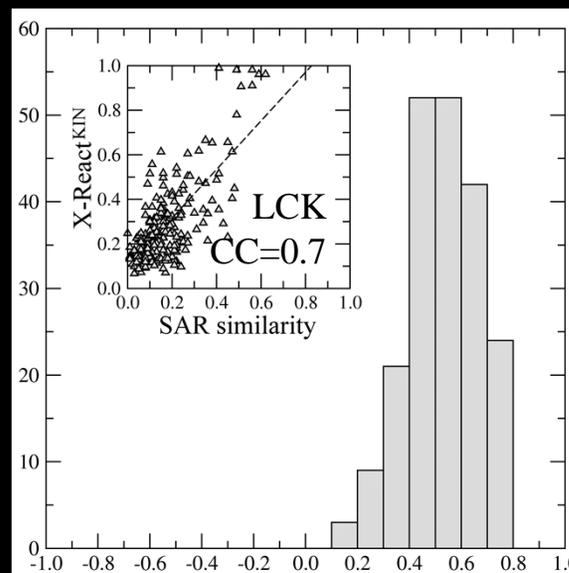
X-React^{KIN}

Bamborough *et al.* (2008) *J Med Chem* **51**, 7898-7914

LIGAND HOMOLOGY MODELING: SUMMARY OF FINDSITE

- FINDSITE is a powerful threading based approach for the prediction of protein function, binding site location and ligand screening.

- Based on the insight that across evolution the location of the binding site is conserved as well as common features of the bound ligands.

- Method does not require a crystal structure, but also works for low resolution predicted models; tolerates inaccuracies up to a global RMSD of 8-10Å because the binding site is often < 2 Å.

- For approximate models, predicts the binding site within 4 Å for 67% of target proteins.

- Have also demonstrated promising results for GO-based functional inference. Average MCC=0.64

LIGAND HOMOLOGY MODELING: SUMMARY OF FINDSITE[LHM]

- Have developed FINDSITE[LHM], an automated approach to the prediction of the ligand anchor and variable regions, ligand binding pose and virtual ligand screening that is applicable to experimental structures and lower resolution predicted models.

- Has significant implications as to how protein function evolves. Have conserved anchor region and variable region that imparts specificity.

- For binding pose prediction, method works acceptably for protein models with a RMSD from native of 4-5 Å.

- Encouraging results shown for the use of low resolution models in virtual ligand screening.

- Ligand ranking is not simply correlated with its molecular weight. Unfortunately, many ligand docking algorithms give results that are strongly correlated with molecular weight.

LIGAND HOMOLOGY MODELING: SUMMARY KINOME^{LHM}/X-REACT^{KIN}

- Have provided structure predictions for the entire human kinome

- For each KINASE, have screened ZINC7 library of

  ~2 million compounds. Have ranking and binding pose predictions for each compound.

- Can exploit these screening results to prioritize ligands that might be specific or general kinase inhibitors.

- Have predicted the structure and cross-reactivity of all proteins in the Human Kinome, results are strongly correlated with experimental SAR results.

- A database of all predictions of the entire human kinome is at Kinome^{LHM} is at http://cssb.biology.gatech.edu/kinomelhm.

LIGAND HOMOLOGY MODELING: PUBLICATIONS

FINDSITE

Brylinski M, Skolnick J (2008) *PNAS* **105**:129
Skolnick J, Brylinski M (2009) *Brief Bioinform* **10**:378
Brylinski M, Skolnick J (2010) *Proteins* **78**:118

FINDSITE[LHM]

Brylinski M, Skolnick J (2009) *PLoS Comput Biol* **5**:e405

Evaluated by Rainer Merkl & Reinhard Sterner

Q-Dock[LHM]

Brylinski M, Skolnick J (2008) *J Comput Chem* **29**:1574
Brylinski M, Skolnick J (2010) *J Comput Chem* **31**:1093

Kinome[LHM] / X-React[KIN]

Brylinski M, Skolnick J (2010) *J Chem Inf Model* **50**:1839
Brylinski M, Skolnick J (2010) *Mol Pharm* **7**:2324

Kinome[LHM]
October 2010
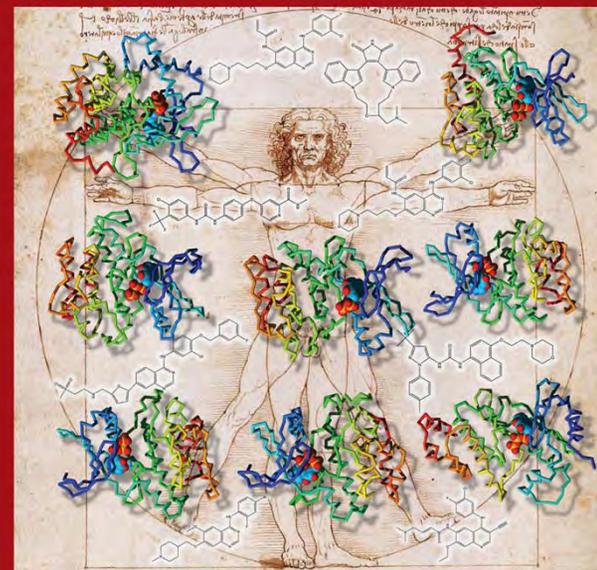
LIGAND HOMOLOGY MODELING: AVAILABILITY

http://cssb.biology.gatech.edu

User-friendly webservers

Software & databases

Manuals & documentation

SERVICES

PSIFR PSIFR

Protein Structure Prediction
chunk-TASSER
MetaTASSER
pro-sp3-TASSER
TASSER
BSR

Protein Structure Alignment
Fr-TM-align
iAlign

DNA-binding Prediction
DBD-Threader
DBD-Hunter
DP-dock

Protein Function Prediction
FINDSITE
FINDSITE-metal
EFICAz^2

Ligand Docking/Screening
FINDSITE^{LHM}
Q-Dock^{LHM}

SOFTWARE

EFICAz^2
FINDSITE
FINDSITE^{LHM}
FINDSITE-metal
Fr-TM-align
iAlign
PULCHRA
TASSER-Lite

Databases
Structure Prediction Libraries
Human GPCRs
E. coli Proteome
PDB-like Structures

Function Prediction Libraries
Kinome^{LHM}
X-React^{KIN}
Human Proteome
Human DNA Binders

Reference Libraries
Active Site Templates
Apo/Holo Pairs

MANUALS

FINDSITE

FINDSITE-metal
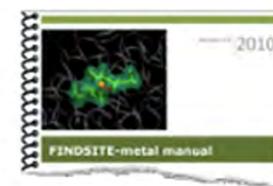
FINDSITE^{LHM}
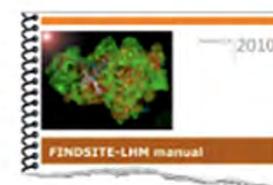
CROWDING AND HYDRODYNAMIC INTERACTIONS LIKELY DOMINATE *IN VIVO* MACROMOLECULAR MOTION

INTRODUCTION

The total concentration of macromolecules inside an *E. coli* cell is in the range of 300-400 mg/ml. This crowding greatly affects both the kinetics and equilibria of biochemical reactions in living systems.
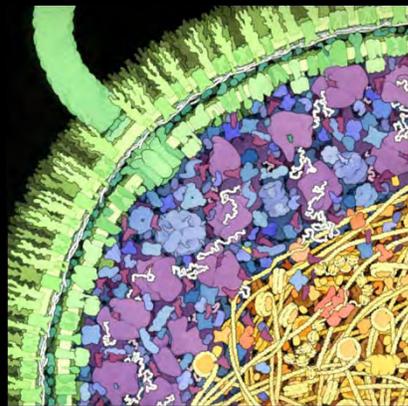
(Illustrated by D. Goodsell)
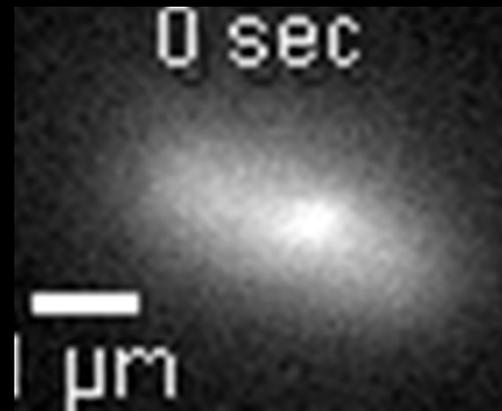
Motion of a tagged RNA inside an *E. coli* cell (Golding and Cox, 2006)

To improve our understanding of a living cell, development of computational methods that can provide reliable predictions of crowding effects on biological reactions is necessary.

Need to develop and apply simulation methods, e.g. Brownian Dynamics

What are the features dominating macromolecular diffusion within an *E. coli* cell?

- Crowding alone?

- How important is macromolecular shape?

- Are hydrodynamic interactions important?

- What are the differences in dynamic behavior when HI or attractive interactions dominate?

- What is needed to reproduce experiment?

PRELIMINARY PROTOCELL STUDY

The "crowded environment" inside a cell can alter and modify the overall behavior of biological systems.

Modeling the crowded cellular environment is an important first step toward whole cell simulation.

We have developed a Brownian dynamics method for simulating the motion of molecules in a cell that can include HI .

PROPERTIES OF AN *E. COLI CELL*

Description	Data
Cell total volume	1 fL
Cytoplasm volume	0.67 fL
Nucleoid (DNA + protein) volume	0.16 fL
# of cytoplasmic proteins (excluding ribosomal proteins):	1,000,000 (3 mM)[†]
# of ribosomes	18,000 (58 µM) [†]
# of tRNAs	200,000 (650 µM) [†]
# of mRNAs	4,000 (13 µM) [†]
Volume occupancy of macromolecules	~30% (300 – 400 mg/mL)
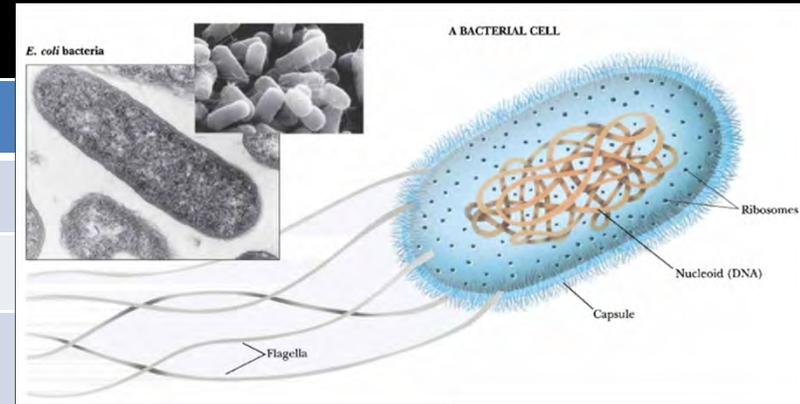
(R. H. Garrett and C. M. Grisham, *Biochemistry*, 1999)

[†]The value of [(Cytoplasmic volume) – (Nucleoid volume)] was used as the volume for the denominator of the concentration calculation.

(From the CyberCell database, CCDB)
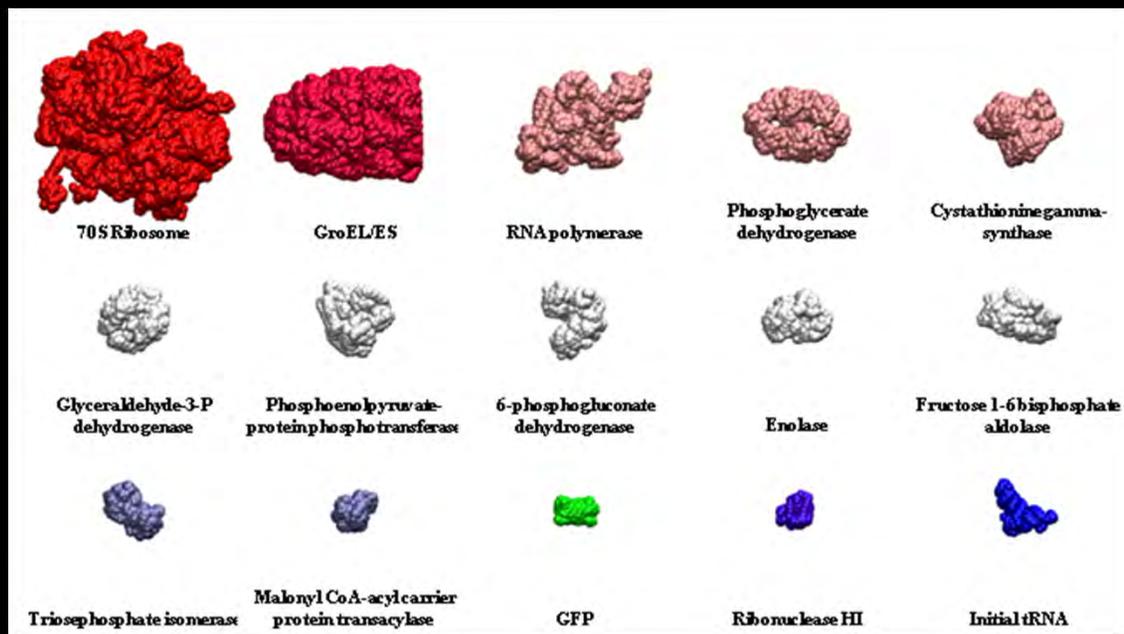
VIRTUAL CYTOPLASMIC SYSTEM
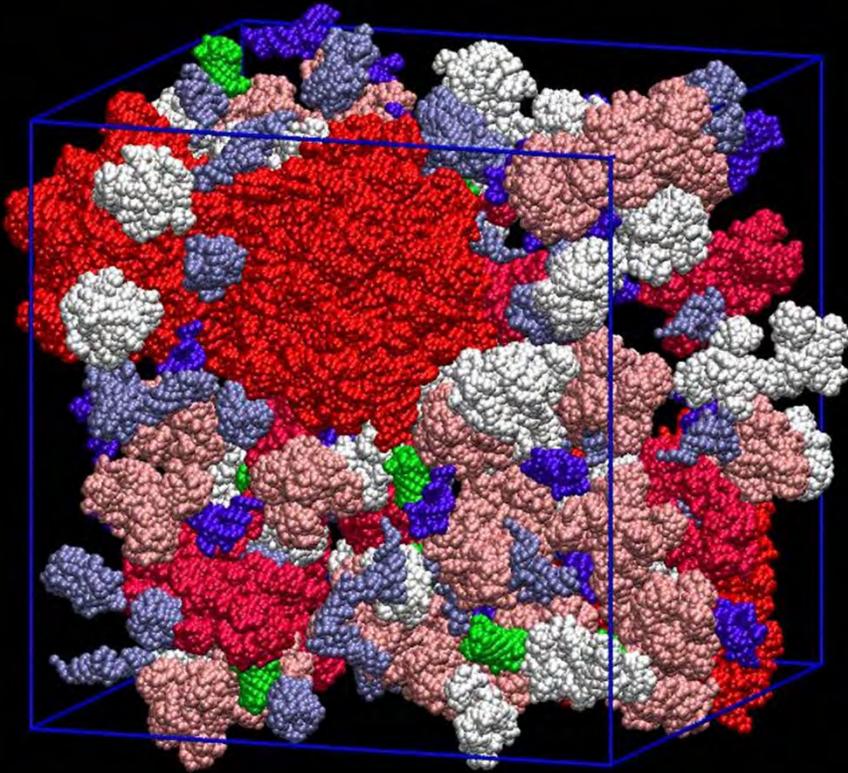
- Proteins are represented by Cα beads. For nucleic acids, C4 and P atoms were used to describe each nucleotide.

- Box size is 100 nm x 50 nm x 50 nm (1.25×10^{-4} fL).

- This system contains 29 ribosomes, 528 glycolytic enzymes, 299 tRNA's, and 113 GFP's. Total concentration is 300 mg/mL.

- The macromolecules were placed within the box by random translation and rotation without any steric clashes.

- Stokes-Einstein relationships for spherical particles:

$$D^{\mathrm{T}} = \frac{k_{\mathrm{B}}T}{6\pi\eta a} \text{ and } D^{\mathrm{R}} = \frac{k_{\mathrm{B}}T}{8\pi\eta a^3},$$

where η is the viscosity of the solvent and a is the Stokes radius of the sphere.

An arbitrarily shaped object undergoing Brownian motion is expressed by a 6 × 6 diffusion tensor, **D**, which is related to a resistance tensor, , through the generalized Einstein relationship . The translational diffusion tensor D_{tt} is related to the translational diffusion coefficient, D_o by

$$D_o = 1/3 Tr(\mathbf{D}_{tt})$$

For a system with hydrodynamic interactions, the hydrodynamic interaction tensor is calculated using the Rotne, Prager, Yamakawa formalism:

$$\mathbf{T}_{ij} = \begin{cases} \dfrac{1}{8\pi\eta r_{ij}}\left[\left(\mathbf{I}+\dfrac{\mathbf{r}_{ij}\mathbf{r}_{ij}}{r_{ij}^2}\right)+\dfrac{2a^2}{r_{ij}^2}\left(\dfrac{1}{3}\mathbf{I}-\dfrac{\mathbf{r}_{ij}\mathbf{r}_{ij}}{r_{ij}^2}\right)\right] & r_{ij}\geq 2a, \\[3ex] \dfrac{1}{6\pi\eta a}\left[\left(1-\dfrac{9}{32}\dfrac{r_{ij}}{a}\right)\mathbf{I}+\dfrac{3}{32}\dfrac{\mathbf{r}_{ij}\mathbf{r}_{ij}}{r_{ij}a}\right] & r_{ij}< 2a. \end{cases}$$

With η the viscosity of the solvent and $\mathbf{r}_{ij}$ is the distance vector between beads i and j. *Note that the radius of bead is the only parameter to be optimized to reproduce hydrodynamic properties in dilute conditions.*

Now, consider a 3N × 3N supermatrix, **B**, consisting of N × N **B**$_{ij}$ blocks at an arbitrary origin O

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathrm{L} & \mathbf{B}_{1N} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \mathbf{B}_{N1} & \mathrm{L} & \mathbf{B}_{NN} \end{pmatrix},$$

$$\mathbf{B}_{ij} = \delta_{ij} \frac{1}{6\pi\eta a} + \left(1 - \delta_{ij}\right)\mathbf{T}_{ij}.$$

Here, δ_{ij} is the Kronecker delta function. This supermatrix is then inverted to obtain a 3N × 3N supermatrix, **C**,

$$\mathbf{C} = \mathbf{B}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathrm{L} & \mathbf{C}_{1N} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ \mathbf{C}_{N1} & \mathrm{L} & \mathbf{C}_{NN} \end{pmatrix},$$

and

$$\Xi_{tt} = \sum_i \sum_j \mathbf{C}_{ij}$$

For *N* particles system in a Newtonian fluid and in absence of an external shear flow, the hydrodynamic forces acting on particles, **F**, are related to the particle velocities, **U**, through the Stokes equation

$$\mathbf{F} = \mathbf{R} \bullet \mathbf{U}$$

where **R** is the resistance matrix and is the inverse of the mobility matrix, **M**.

When torque-angular velocity is not considered, the so-called "F version" in Ref. (11), **F** and **U** are $3N \times 1$ vectors and **R** and **M** are $3N \times 3N$ matrices. Then, the diffusion matrix of the system is simply given by

$$\mathbf{D} = k_B T \mathbf{M}$$

The resistance tensor **R**, which contains both near-field lubrication effects and far-field many-body interactions, is calculated as

$$\mathbf{R} = \left(\mathbf{M}^{\infty}\right)^{-1} + \mathbf{R}_{2B} - \mathbf{R}_{2B}^{\infty}$$

- $(\mathbf{M}^{\infty})^{-1}$, represents the contribution of many-body, far-field interactions.

- $\mathbf{R}_{2B}$ represents the exact two-body HI, which includes both near-field and far-field interactions.

- is the resistance tensor that represents two-body far-field interactions. The far-field part has already been included on $(\mathbf{M}^{\infty})^{-1}$. Thus, in order not to count these interactions twice, we must subtract off the two-body interactions. This is the standard method to correct for the lubrication effects in the resistance tensor.

$$\mathbf{x}_i = \mathbf{x}_i^0 + \frac{\Delta t}{k_B T} \mathbf{D}_i \cdot \mathbf{F}_i^p + \mathbf{G}_i \left(\Delta t \right)$$

where Δt is the time step and $\mathbf{x}_i$ is the vector describing the position of the center of diffusion and orientation of the i-th object. $\mathbf{F}^p$ is a generalized force and $\mathbf{G}_i(\Delta t)$ is a 6×1 random displacement vector during time step Δt due to the Brownian noise, which satisfies

$$\left\langle \mathbf{G}_i \left(\Delta t \right) \right\rangle = 0, \left\langle \mathbf{G}_i \left(\Delta t \right) \mathbf{G}_j \left(\Delta t \right) \right\rangle = 2 \mathbf{D}_i \Delta t \delta_{ij}$$

When HI are considered, the diffusion tensor depends in principle on the configuration of the entire system. The propagation eq is

$$\mathbf{r} = \mathbf{r}^0 + \left(\nabla \cdot \mathbf{D}\right)\Delta t + \frac{\mathbf{D} \cdot \mathbf{F}^p}{k_B T}\Delta t + \mathbf{G}\left(\Delta t\right)$$

$$= \mathbf{r}^0 + k_B T\left(\nabla \cdot \mathbf{M}\right)\Delta t + \left(\mathbf{M} \cdot \mathbf{F}^p\right)\Delta t + \mathbf{G}\left(\Delta t\right),$$

Where **r** is the particle's position vector and **G**(Δt) is the random displacement due to Brownian motion, which has the following properties

$$\left\langle \mathbf{G}\left(\Delta t\right)\right\rangle = 0, \left\langle \mathbf{G}\left(\Delta t\right)\mathbf{G}\left(\Delta t\right)\right\rangle = 2k_B T\mathbf{M}\Delta t$$

In contrast to a BD algorithm with constant diffusion tensors, we need to evaluate the spatial gradient of the mobility tensor in the BD simulation with HI, in which the explicit computation of is a $O(N^3)$ task. To avoid this, we used a method introduced by Banchio and Brady, based on Fixman's idea, the so-called "mid-point scheme".

$$\langle \mathbf{F}^{\mathrm{B}} \rangle = 0, \langle \mathbf{F}^{\mathrm{B}}(0)\mathbf{F}^{\mathrm{B}}(t) \rangle = 2k_{\mathrm{B}}T\mathbf{R}/\Delta t$$

Here, $\mathbf{F}^{\mathrm{B}}$ is the Brownian force, obtained by the Cholesky decomposition method. The procedure is the following:

(1) Compute the velocity $\mathbf{U}^0$ using an initial configuration $\mathbf{r}^0$

$$\mathbf{U}^0 = \left(\mathbf{R}^0\right)^{-1} \cdot \left(\mathbf{F}^{\mathrm{p},0} + \mathbf{F}^{\mathrm{B},0}\right)$$

(2) Move the particles to intermediate positions **r'** by a small
fraction of a time step, $\Delta t/m$

where m is 100.
$$\mathbf{r'} = \mathbf{r}^0 + \frac{\Delta t}{m}\mathbf{U}^0$$

(3) Calculate a new velocity **U'** at the intermediate positions
using the forces evaluated at $\mathbf{r}^0$

$$\mathbf{U'} = \left(\mathbf{R'}\right)^{-1} \cdot \left(\mathbf{F}^{p,0} + \mathbf{F}^{B,0}\right)$$

(4) Calculate the drift velocity, $\mathbf{U}^{drift}$,

$$\mathbf{U}^{drift} = \frac{m}{2}\left(\mathbf{U'} - \mathbf{U}^0\right)$$

(5) Finally, update the positions of the particles for time step Δt.

$$\mathbf{r} = \mathbf{r}^0 + \left(\mathbf{U}^0 + \mathbf{U}^{drift}\right)\Delta t$$

For short times, the mean square displacement is linear and the short time diffusion constant is defined as

$$D_i^S = \frac{1}{3N_i} \sum_{\alpha \in i}^{N_i} \mathrm{tr}\left(\mathbf{D}^{\alpha\alpha}\right)$$

N_i is the number of type i particles in the system and $\mathbf{D}^{\alpha\alpha}$ is the 3 × 3 matrix of the self part of the diffusion tensor.

Similarly, the long time diffusion constant is given by

$$D^L = \lim_{t \to \infty} \left\langle \left| \mathbf{r}(t) - \mathbf{r}(0) \right|^2 \right\rangle \Big/ 6t$$

SIMULATION CONDITIONS

- Periodic boundary conditions were used.

- Simulation temperature was 298 K.

- Generated ten different random configurations of the system.

- 30 independent simulations were done.

- For BD simulations of repulsive and non-specific, attractive binding models without HI, 30 and 50 µs simulations were performed with a time step of 0.5 and 0.1 ps, respectively.

- For BD simulations with HI, we ran 15 µs simulations with a time step of 2 ps.

- The first 5, 30, and 5 µs of simulations of repulsive, non-specific binding models, and HI models were ignored entirely. To estimate the long-time diffusion coefficients, MSD values after a relative time interval of 5 µs was used.

RESULTS
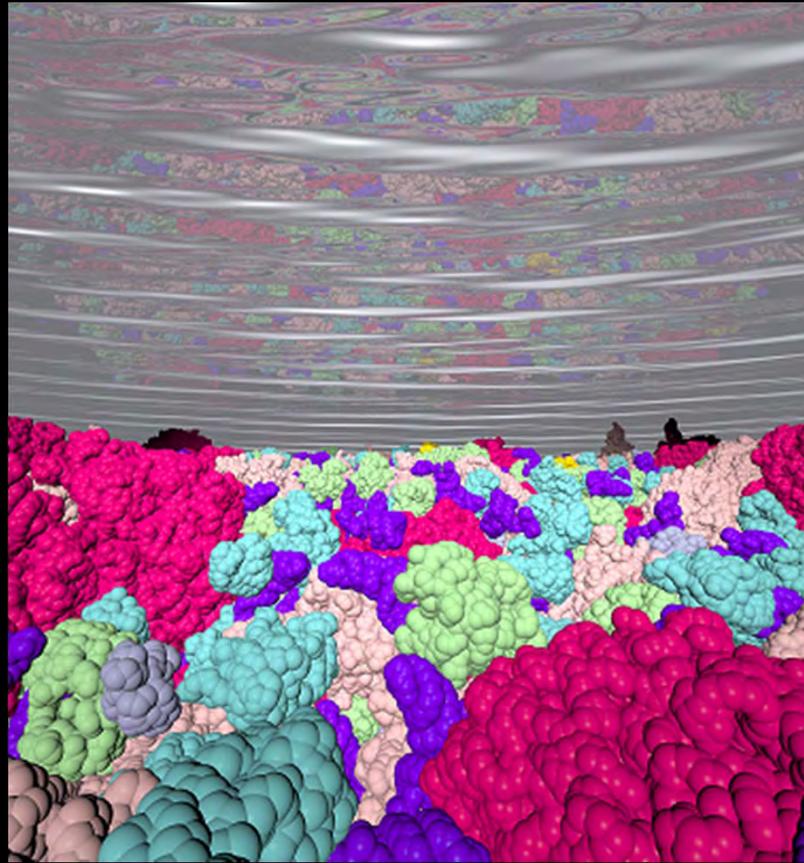
See: T. Ando and J. Skolnick. Crowding and hydrodynamic interactions likely dominate *in vivo* macromolecular motion. Proc Natl Acad Science 2010:**107**: 18457-18462.

ESTIMATION OF INFINITE DILUTION MACROMOLECULAR DIFFUSION CONSTANT FROM ATOMIC STRUCTURE

Fit to infinite dilution diffusion constant of a representative set of proteins and t-RNA using the rigid-particle formalism (see Garcia De La Torre J, Huertas ML, & Carrasco B (2000) *Biophys J* 78(2):719-730). Fit the data with a bead radius of 6.1 Å

Gives GFP diffusion constant of 8.9 Å^2 vs experiment of 8.7 Å^2.
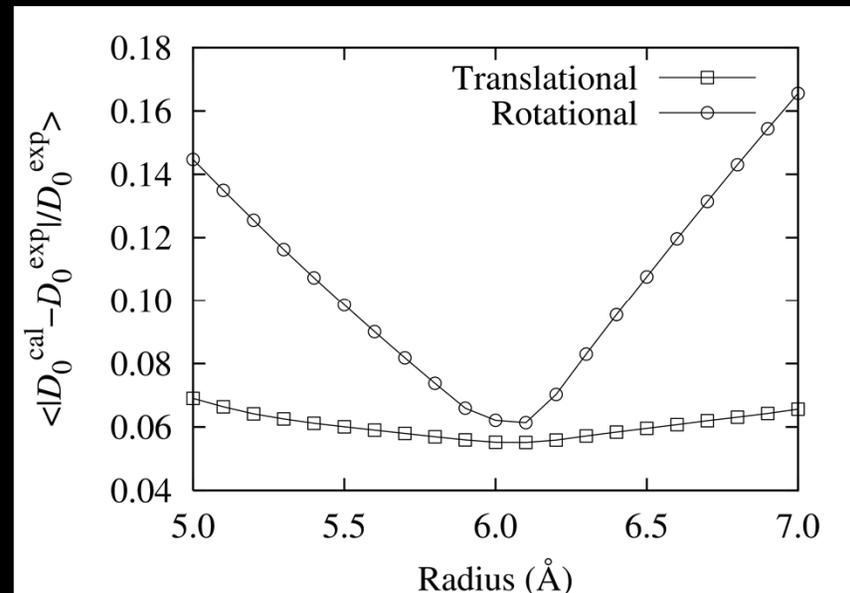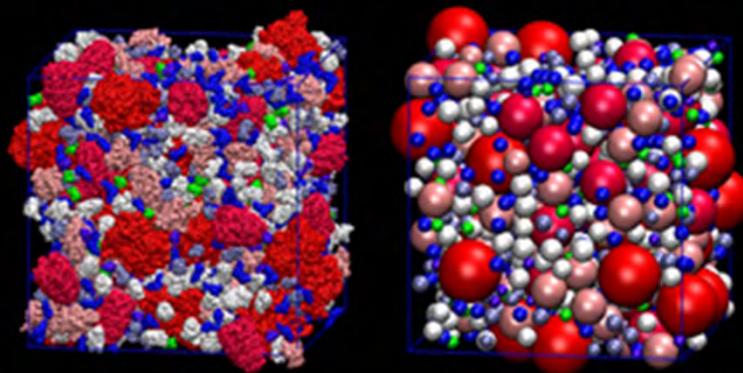
Molecular-shaped (*left*) and sphere (*right*) systems at 300 mg/ml. Macromolecules are represented in different colors.

DIFFUSION CONSTANTS IN DENSE SYSTEMS OF MOLECULES AND EQUIVALENT SPHERE SYSTEMS
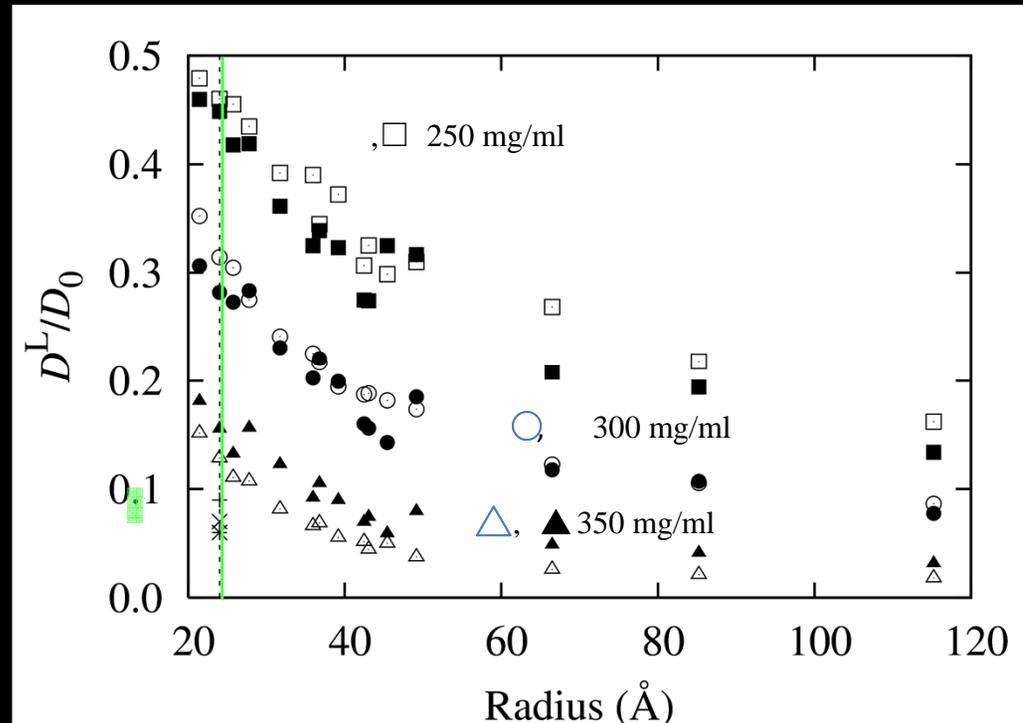
Long time diffusion constant ratio as a function of macromolecule radius in the sphere (open symbols) and molecular-shaped systems (filled syReduction in diffusion constant of GFP measured *in vivo* of DH5α, BL21(DE3), and K-12 *E. coli*. are shown by plus, cross, and asterisk, respectively. Green line is GFP's radius.mbols).

- Below 350 mg/ml, the results of explicit shape and equivalent sphere systems are very close.
- At 350 mg/ml, D_L is somewhat smaller for the sphere systems.
- So we can use the equivalent sphere approximation to explore the role of HI.
- But for both representations, crowding cannot explain the D_L/D_o of 0.06-0.0.09 for GFP. For example, at 300 mg/ml, simulated D_L/D_o = 0.31.
- Something is missing!

COMPARISON OF SPHERICAL SYSTEM WITH HYDRODYNAMIC INTERACTIONS VS JUST REPULSIVE INTERACTIONS
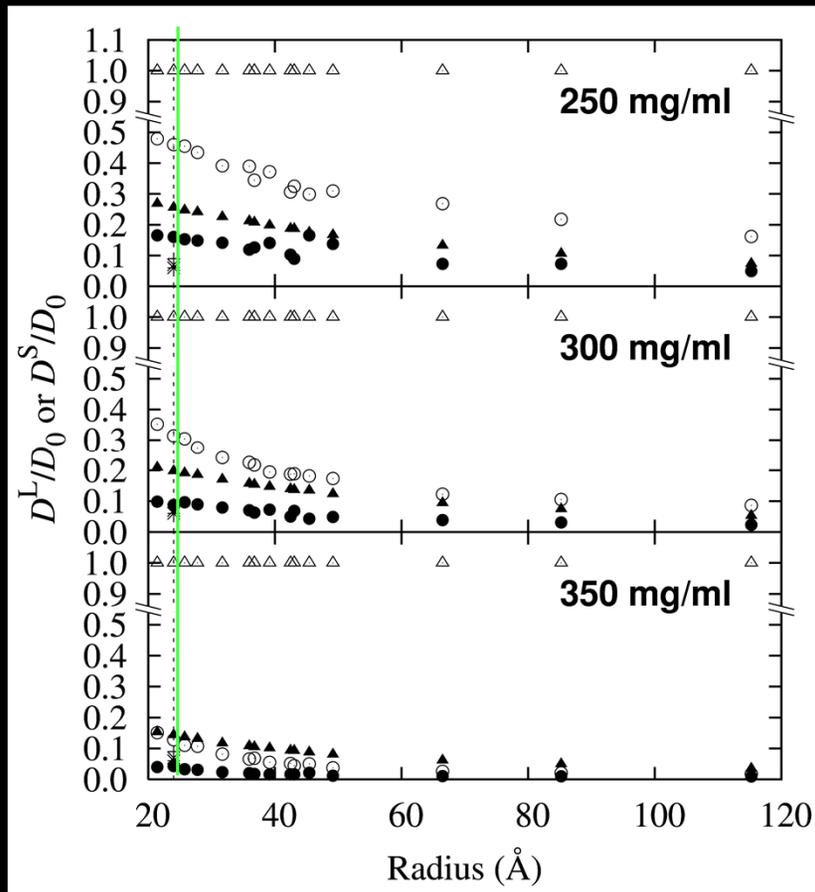
Reduction in diffusivity as a function of radius at three different concentrations. Triangles and circles represent D^S/D_0 and D^L/D_0. Open (filled) symbols are values in the sphere model with repulsive (HI) interactions. Plus, cross, and asterisk symbols exp. Green line is GFP's radius.
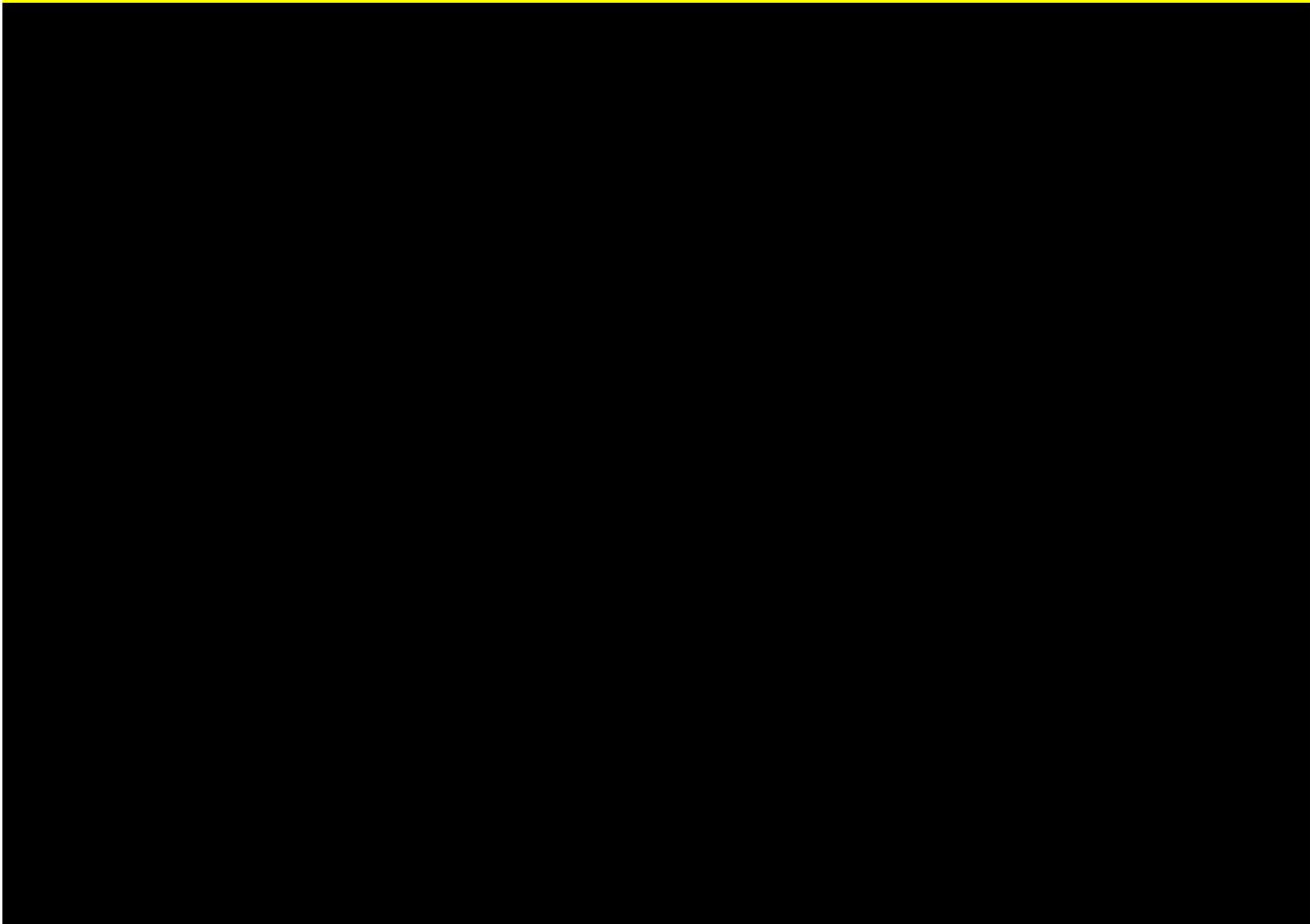
- HI greatly reduces the short time diffusion constant, D^S from the infinite dilution value, D_o. In contrast, $D^S = D_o$ when HI are ignored.

- For GFP, without adjustable parameters, simulations reproduce the experimental reduction in D_L.

- Implies that crowding and HI are major factors responsible for the slow down in diffusion in intracellular environments.

- Consider each macromolecule to be a rough sphere filled with van der Waals particles with a 3 Å diameter.

- Surface roughness is estimated as the difference between the Stokes radius and the radius of gyration of the macromolecule.

- Ignore HI and adjust the strength of the van der Waals attraction to give the experimentally observed reduction in GFP's diffusion constant.
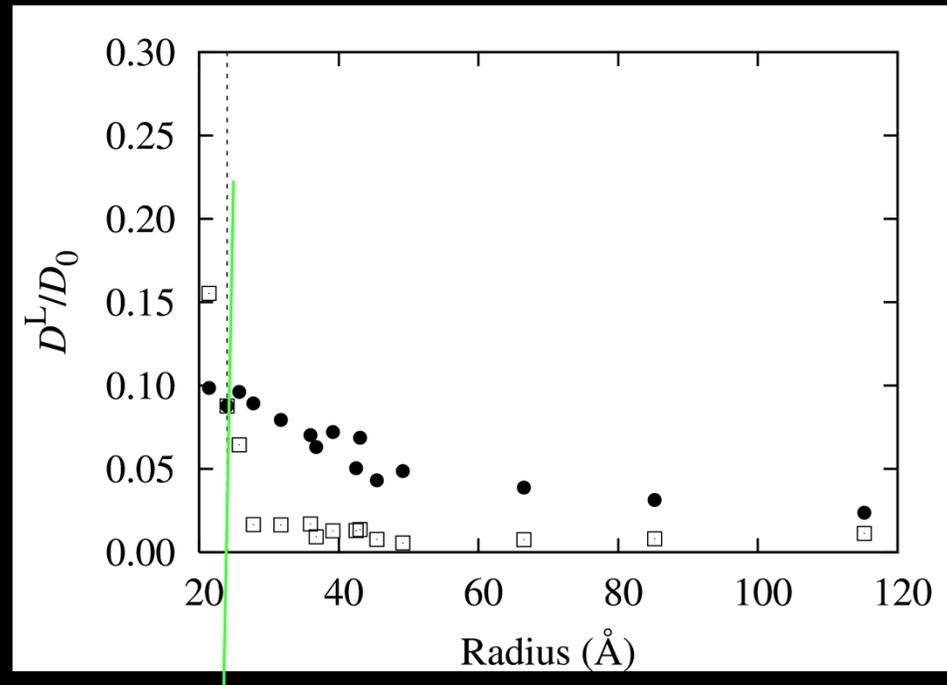
At 300 mg/ml, long-time diffusion constant ratio, D^L/D_0, as a function of radius in the non-specific, van der Waals interaction (HI) model is represented by squares (filled circles). Green line is GFP's radius.

- If non specific attractions dominate *in vivo* diffusion, then reduction in diffusion constant is very strongly dependent on molecular radius.

- In contrast, if HI dominate, reduction in diffusion constant is much less sensitive to molecular radius.

- ## Calculate normalized pair correlation function between molecules *i* and *j:*

$$C_{ij}\left(d_0, \tau\right) = \frac{\Sigma\left[\left(\Delta \mathbf{r}_i(\tau) \cdot \Delta \mathbf{r}_j(\tau)\right) \delta\left(d_0 - d_{ij}\right)\right]}{\sqrt{\Sigma\left|\Delta \mathbf{r}_i(\tau) \delta\left(d_0 - d_{ij}\right)\right|^2} \sqrt{\Sigma\left|\Delta \mathbf{r}_j(\tau) \delta\left(d_0 - d_{ij}\right)\right|^2}}$$

Where d_0 is a specified the surface distance between particles *i* and *j*, and τ is the time interval. $\delta(d_0 - d_{ij})$ is the Dirac delta function. d_{ij} is the surface distance between particles *i* and *j* at time *t*.
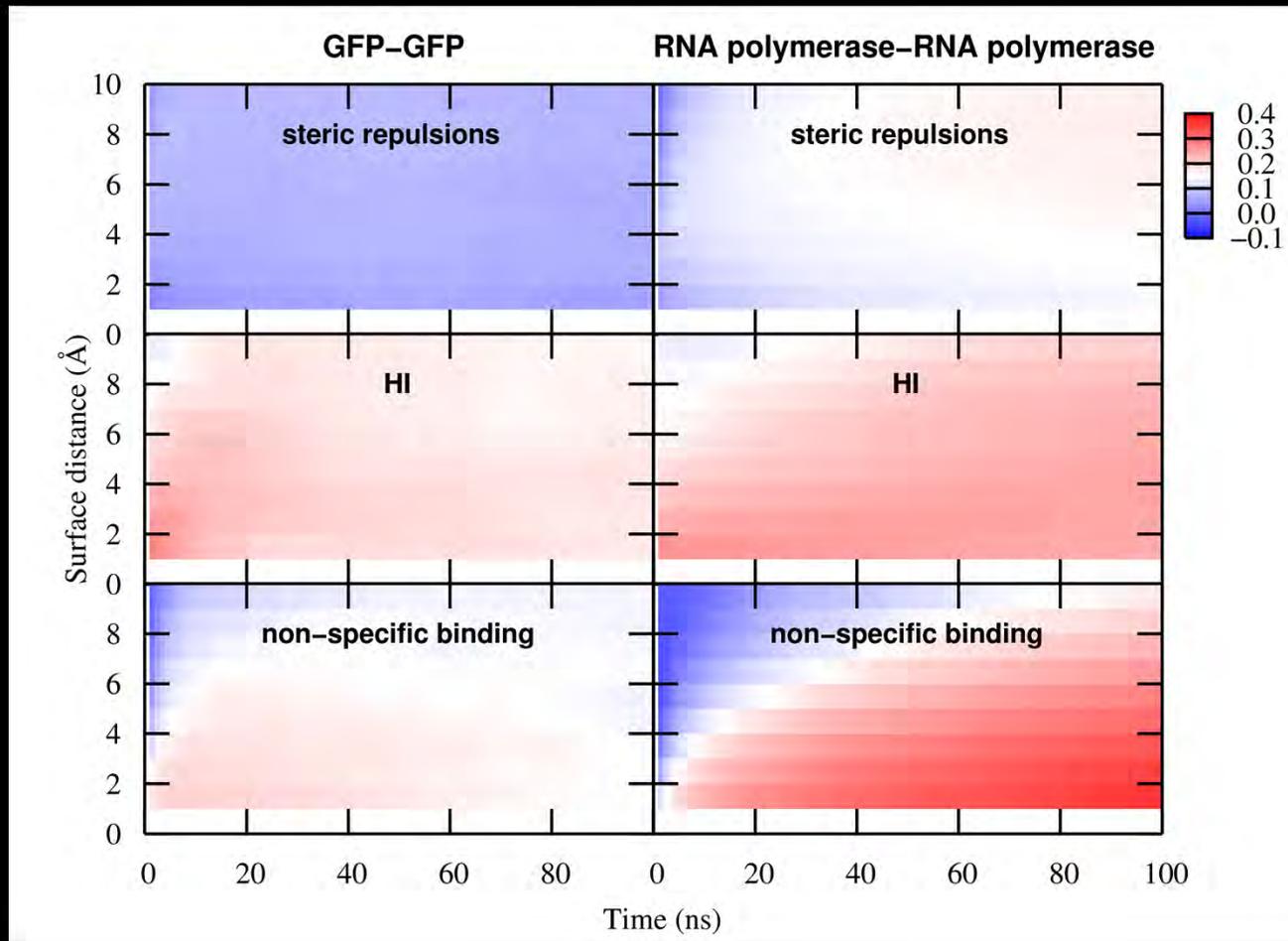
DIFFERENCES OF THE THREE MODELS

Hard sphere model

- Little spatial or temporal correlation.

- $C_{ij} < 0.1$ even at short times.

- True for all size pairs of molecules.

HI model

- For both small pairs and large pairs, see significant, but weak, $C_{ij} < 0.3$, intermolecular correlation that persists up to quite long times (>100 ns) and distances (at least 10 Å).

Non specific binding model

- For large molecules, see positive correlation for distances<5 Å that are long lived in time. See long lived clusters.

- For small molecules, this effect is greatly reduced.

CONCLUSIONS

- Equivalent sphere model is a good description of the motion of macromolecules in intracellular environments.

- HI dynamics likely exert a significant effect on intercellular dynamics.

- Crowding and HI can quantitatively reproduce the experimentally observed diffusion constant of GFP *without any adjustable parameters.*

HOW TO EXPERIMENTALLY DIFFERENTIATE BETWEEN HI AND NON SPECIFIC BINDING MODELS:

If HI dominate:

- Decay like $1/r$; lubrication forces give repulsion on approach and attraction as pairs of move apart.

- Short time diffusion constant, D_s is significantly reduced from D_o.

- D_s/D_0 depends on molecule radius.

- Long time diffusion constant D_L has much weaker dependence on particle radius.

- Have significant spatial and temporal correlations for all size macromolecules.

If nonspecific binding dominates:

- Decay like $1/r^2$ for spherical macromolecules filled with small van der Waal spheres.

- Short time diffusion constant is the same as in infinite dilution, D_o.

- D_s/D_0 is independent of molecule radius.

- D_L is strongly size dependent, with long lived clusters formed with larger macromolecules.

- Significant, radius independent spatial and temporal correlations are absent.

OTHER FACTORS THAT COULD AFFECT INTRACELLULAR DIFFUSION:

- *Electrostatics*- but inclusion in simulations only slightly reduces GFP diffusion constant.

- *Viscosity of the cytoplasm- In vivo* viscosity is essentially the same as bulk water; < 2CP.

- *GFP dimerization*- GFP dimerizes at <100nm ionic strength.

But we expect these factors to be small.

OUTLOOK

- Have a useful model to elucidate some general futures of intracellular dynamics at the molecular level.

- Exploring differences in metabolic fluxes *in vivo* from that at infinite dilution.

- Have generalized this class of models to include a schematic membrane. Using this to examine very simple models of cellular division.

ACKNOWLEDGMENTS

Tadashi Ando

CORRELATION BETWEEN STOKES RADIUS AND THE AVERAGE VALUE OF R_G AND R_{MAX}

Molecule	Stokes radius (Å)	Rg (Å)	R_{max} (Å)	Average of Rg and R_{max} (Å)
Ribonuclease	18	14.2	22.4	18.3
GFP	24.7	17.0	27.6	22.3
Phe-tRNA	28.8	22.3	45.9	34.1
Ovoalbumin	31	21.7	38.5	30.1
Hemoglobin	31	23.7	33.0	28.3
Enolase	36	26.5	41.2	33.8
Aldolase	46	34.9	59.0	47.0
Ribosome	126	85.0	154.9	120.0
(RMSD)	0	17.5	14.2	3.5

Ligand Homology Modeling: Overview

PDB

Template-based protein structure prediction

- Structure templates
- Initial protein model
- Low-resolution refinement *TASSER*
- Improved protein model
- High-resolution refinement *ATASSER*
- **Final protein model**
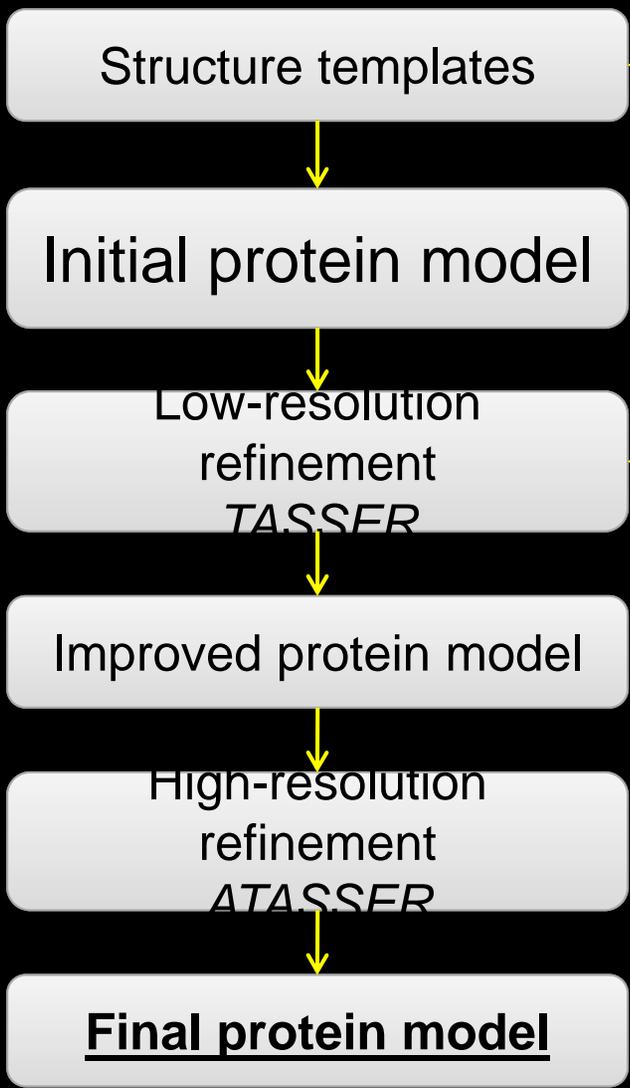
(meta) Threading
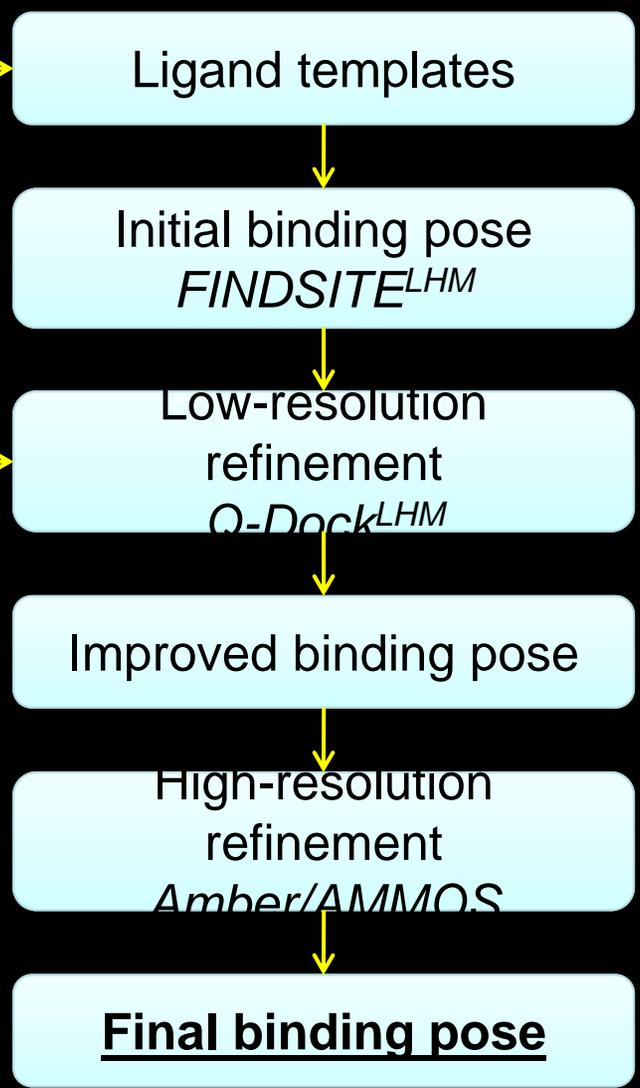*PROSPECTOR_3, SP3, Sparks*

Specific potentials & restraints

Template-based protein-ligand modeling

- Ligand templates
- Initial binding pose *FINDSITE^{LHM}*
- Low-resolution refinement *Q-Dock^{LHM}*
- Improved binding pose
- High-resolution refinement *Amber/AMMOS*
- **Final binding pose**
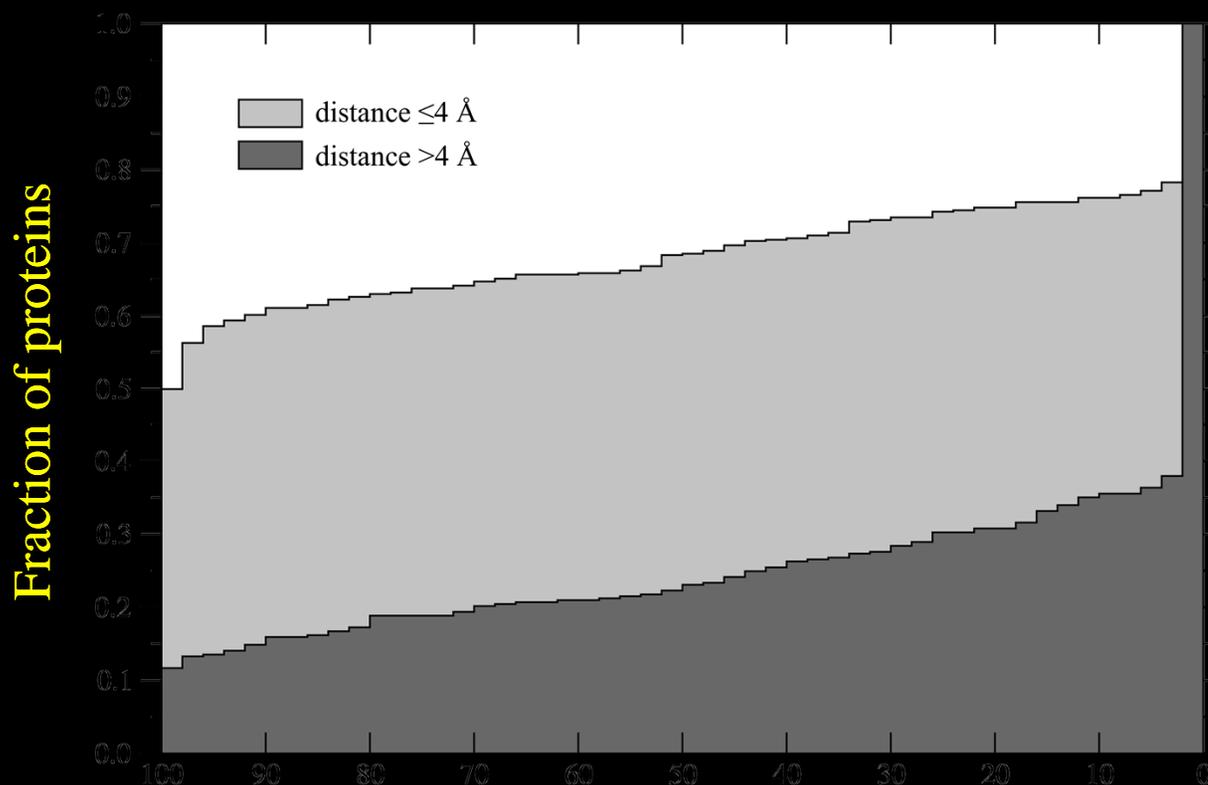
Ligand Homology Modeling: Challenges
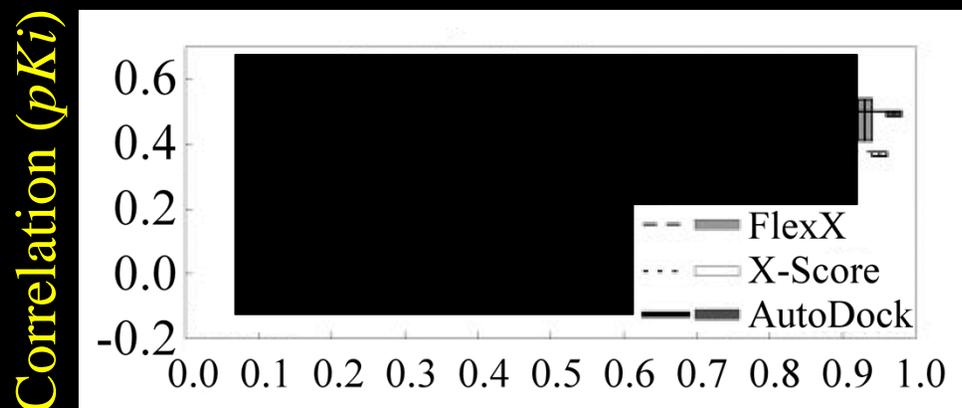
Most ligand docking/scoring algorithms lack specificity

HIV-1 protease
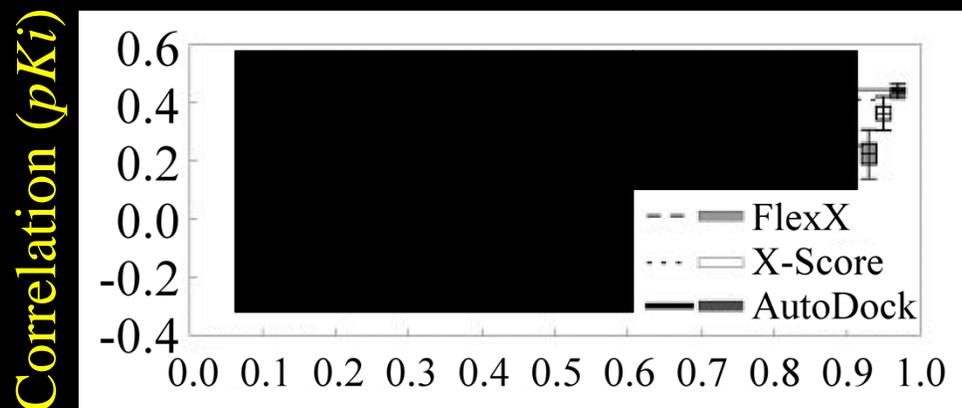
Tanimoto coefficient

Ribonuclease A

Tanimoto coefficient

Calculated for ~1,000 randomized ligands (decoys)

➢ Predicted binding affinity is

 highly correlated with MW

➢ Docking programs cannot distinguish native ligands from their randomized decoys
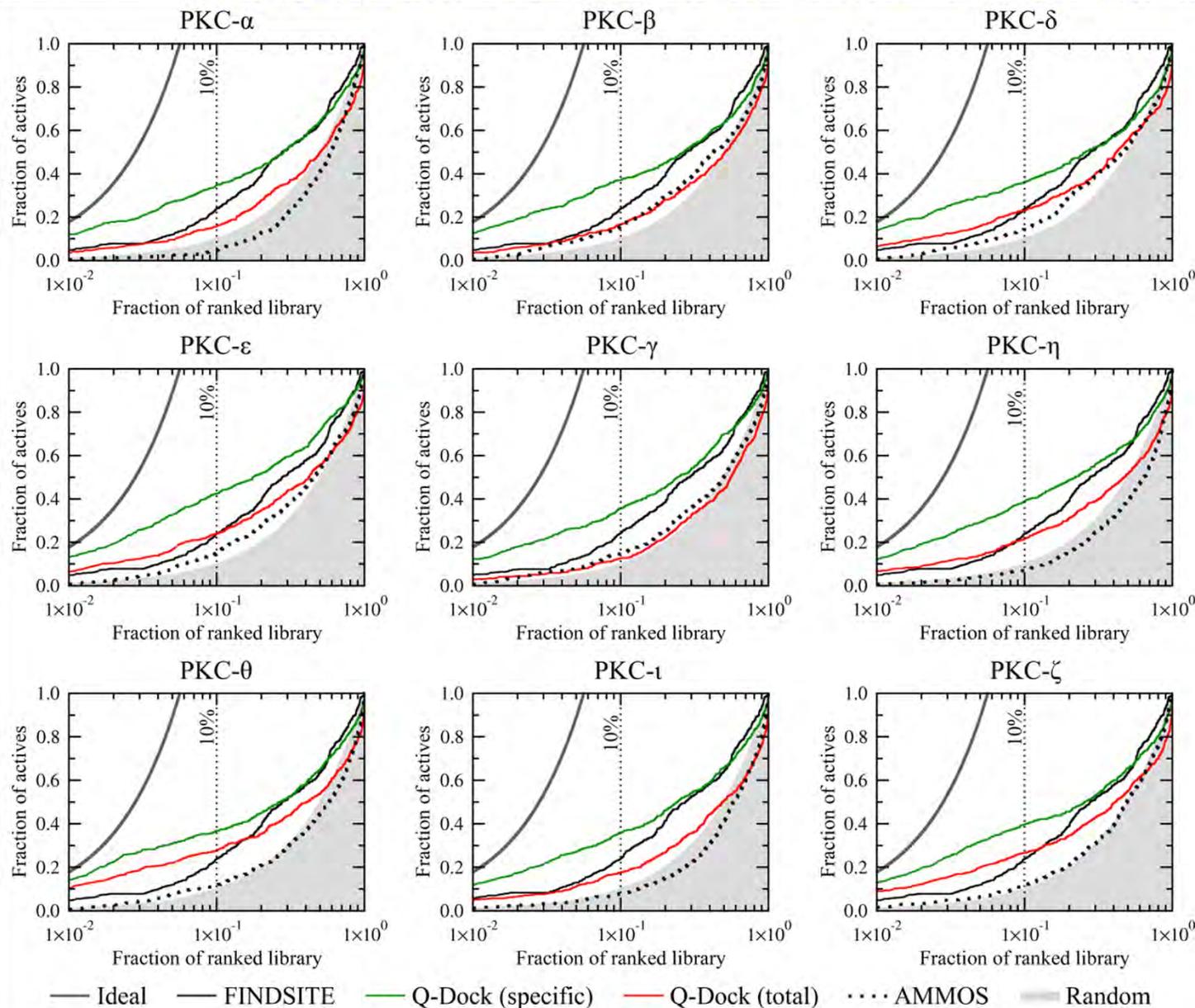
Kim & Skolnick (2008) *J Comput Chem* **29:** 1316-1331

Ligand Homology Modeling: Q-Dock^{LHM}

204 pharmacologically relevant targets from CCDC/Astex dataset

Crystal structures   Cα-RMSD: 3.8 Å

Cα-RMSD: 5.7 Å   Cα-RMSD: 7.3 Å
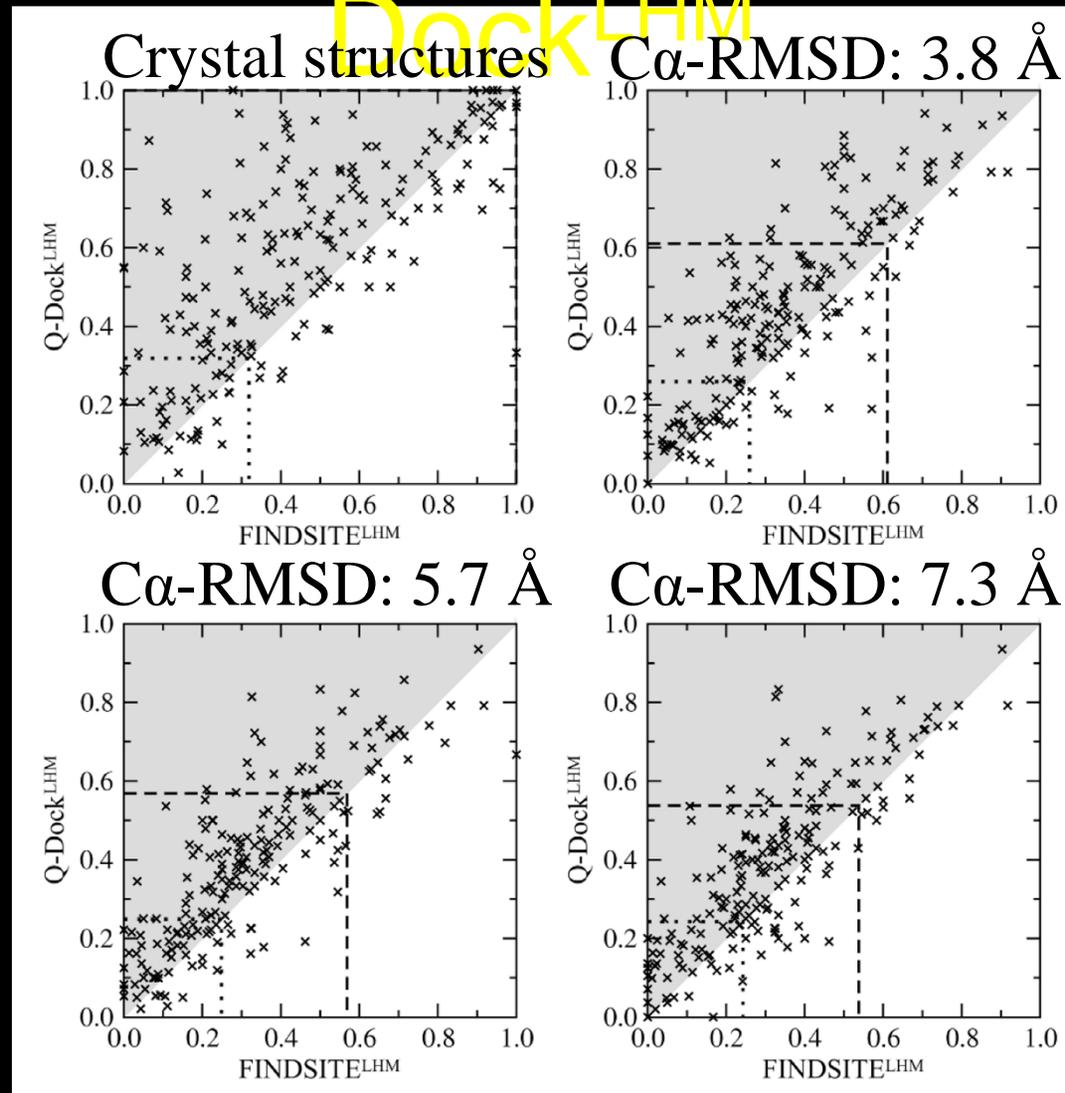
Fraction of correctly predicted specific protein-ligand contacts

Ligand Homology Modeling:

Virtual screening against PKC isoforms

562 known inhibitors (*MDDR*)

10,000 decoy cmps (*ZINC7*)