

Scoring Functions and Docking

Keith Davies

Treweren Consultants Ltd

26 October 2005



Overview

- Applications
- Docking Algorithms
- Scoring Functions
- Results
- Demonstration



Docking Applications

Drug Design

- Lead Generation
- Lead Optimisation
- Library Design
- Compound Purchases

Academic

- Predicting Crystal Structure Complexes



Lead Generation

- Alternative to Experimental HighThroughput Screening
- Issues
 - Availability of Crystal Structures
 - False Positives
- Often Difficult to Demonstrate Cost Benefit



Lead Optimisation

- Ranking Derivatives of Known Active Molecules
- Subject to Systematic Failures
- Not Sufficiently Discriminating for Many Optimisation Series
- Changing Chemical Series is really Lead Generation



Library Design

- Aim to Eliminate Molecules which are not Credible
- Receptor Shape is Important
- Can be Lead Optimisation when only Substituents on an Active core vary



Compound Purchases

- Millions of Catalogue Molecules
 - Diverse
 - Approximately 0.5 Million Drug-like
- Issues
 - False Negatives
 - Delivery Timescales
- Cost Benefit is Demonstratable



Drug-like Filtering

- **Properties**
- **Substructures**
 - Unstable
 - Reactive
 - Undesirable
(eg toxic)

Centres	2:9
Mass	150:800
XSA	20:240
Rot-Bonds	≤ 10
Conformers	≤ 1000000



Classifying Docking Algorithms

- Ligand Conformations
 - Rigid
 - Fixed Sample ~ 100
 - Flexible
- Constraints
 - Residues and Pockets
 - Pharmacophores
- Charges and Tautomers
- Water Molecules



Conformations

- Small Molecules are NOT Rigid
- Literature on Conformational Generation
 - Use of Constraints
 - Speed Enhancements
- Often 000's Representative Conformations
- Suitable Problem for Parallel Computing



Constraints: Binding Site

- Identified by Crystal Structure
 - Activity associated with another site
 - Different binding mode at same site
- Confirmed by Mutagenesis Studies
- Suggested by Software
 - Size of cavity/cleft
 - Binding Potential



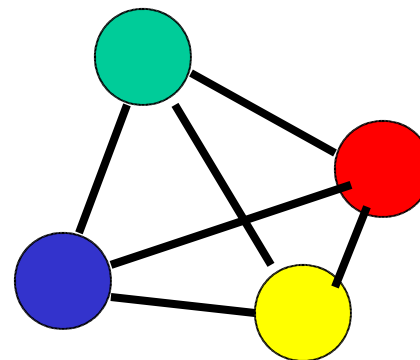
Binding Site Identification

- Place Protein in 3-D Grid
- Remove Grid Points Inside Protein
- Remove Grid Points Inside 8 Å Sphere Positioned on the Grid Outside the Protein
- Binding Site
 - Minimum of 3 Grid Points
 - Minimum of 3 Residues



Constraints: Pharmacophores

- Centre Types
 - **H-bond Donor**
 - **H-bond Acceptor**
 - Acid
 - Base
 - **Positive Charge**
 - Negative Charge
 - **Aromatic Ring**
 - Lipophile
 - Lewis Base
 - 2 User Definable

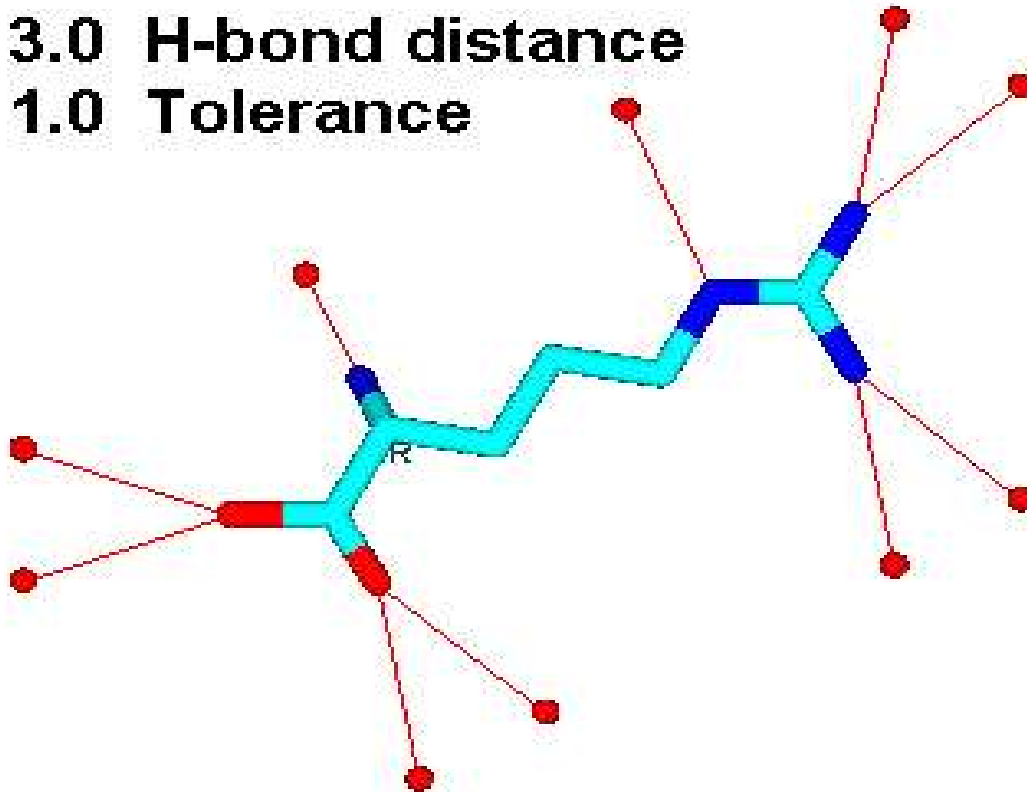


- Options
 - 3 or 4 **Centre Types**
 - User Definable Bins
 - Occurrence Frequency
- Output to file



Centre Positions

3.0 H-bond distance
1.0 Tolerance



Charges and Tautomers

- Docking Doesn't Need Hydrogen Positions
- Charges Important for Energy Functions
 - Sometimes Alternative Charge Models
 - Rarely Multiple Same Charges Within Site
- Tautomer State
 - Sometimes Too Many Alternatives
- Charges and Tautomeric State Must Complement Ligand

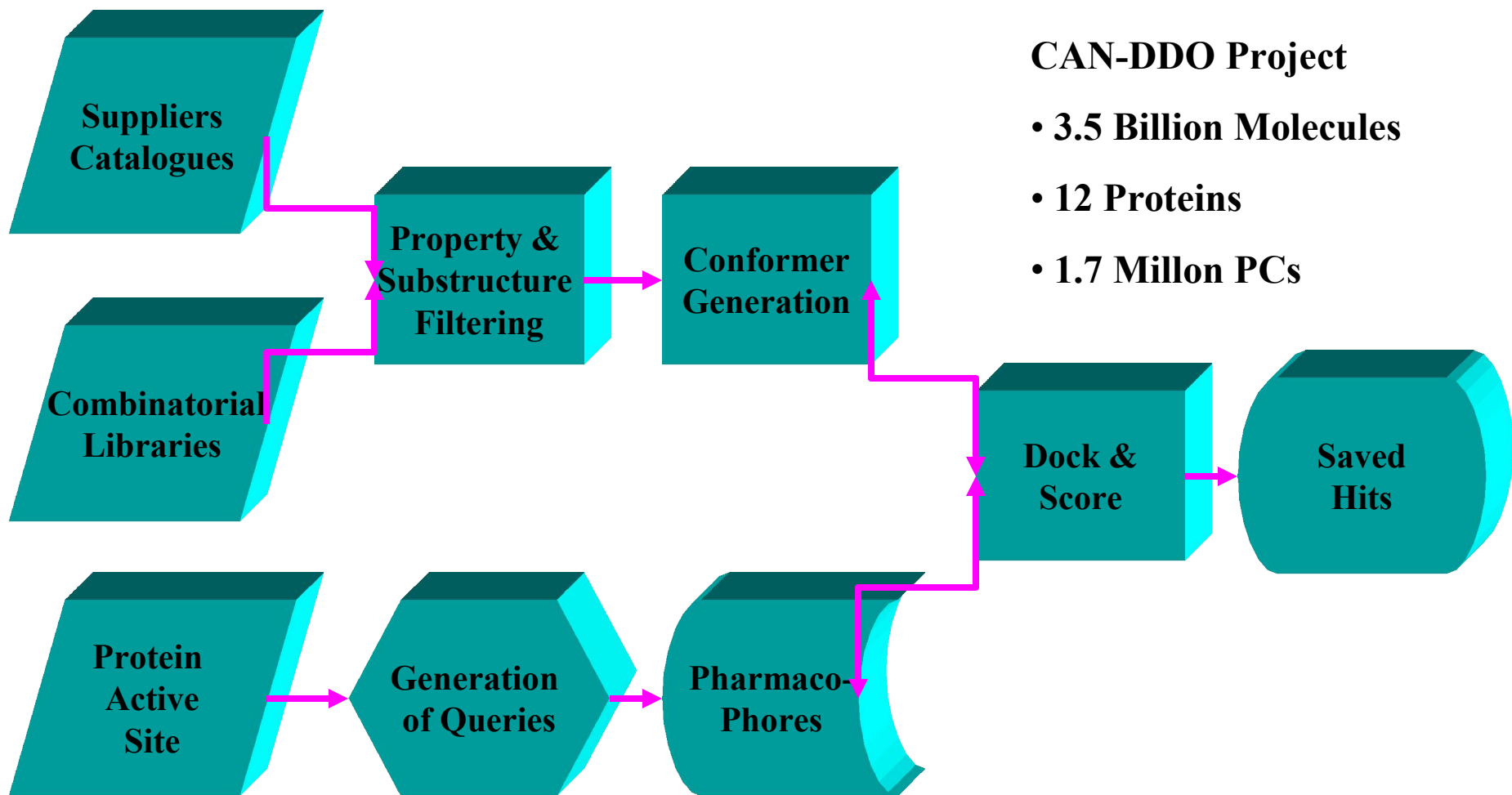


Water Molecules

- Essential
 - Mediates binding for all Ligands
- Optional
 - Presence required by some Ligands
 - Inhibits binding of other Ligands
- Solvation & Desolvation Free Energy
Critical for Scoring Function



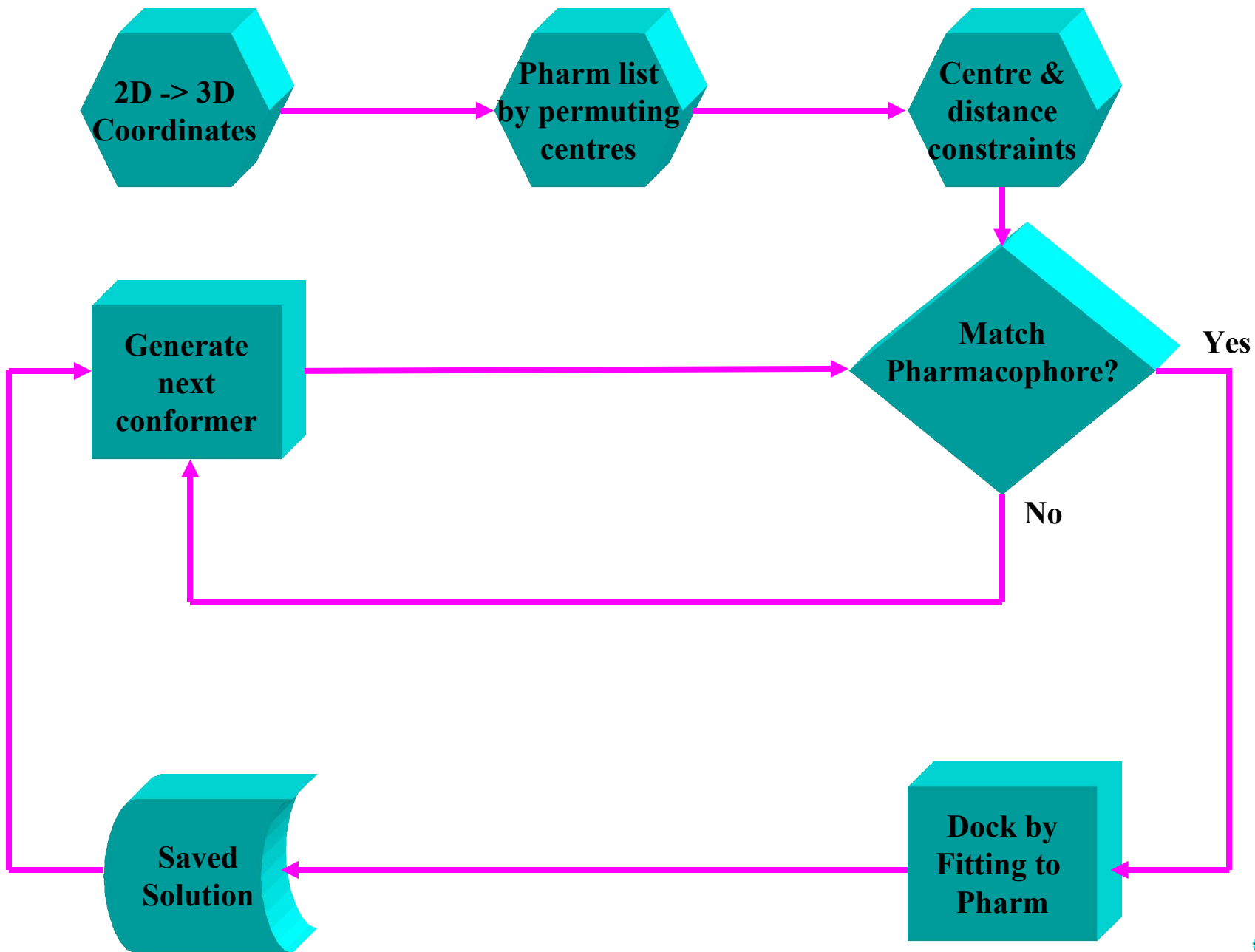
Structure-based Virtual Screening

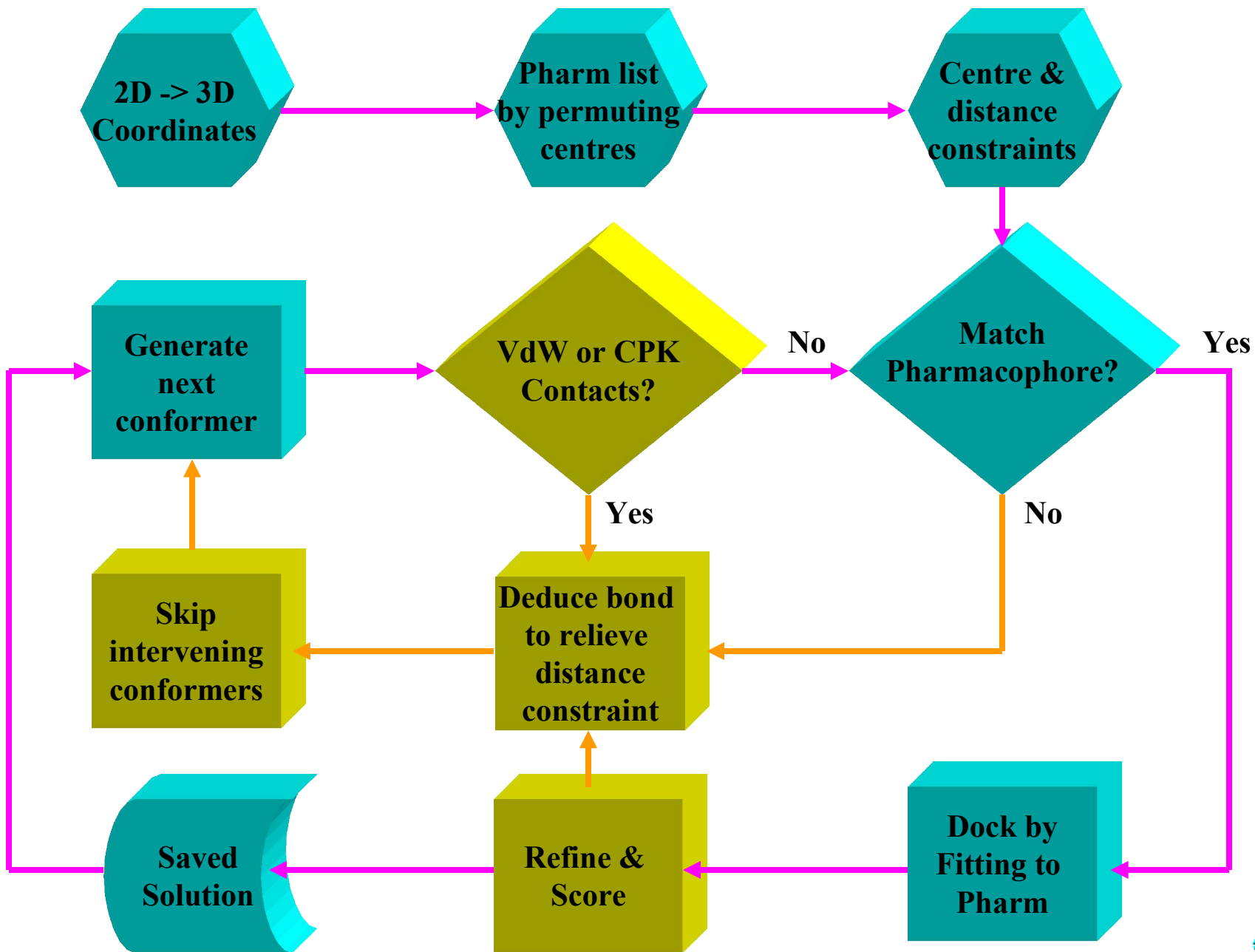


CAN-DDO Project

- 3.5 Billion Molecules
- 12 Proteins
- 1.7 Million PCs





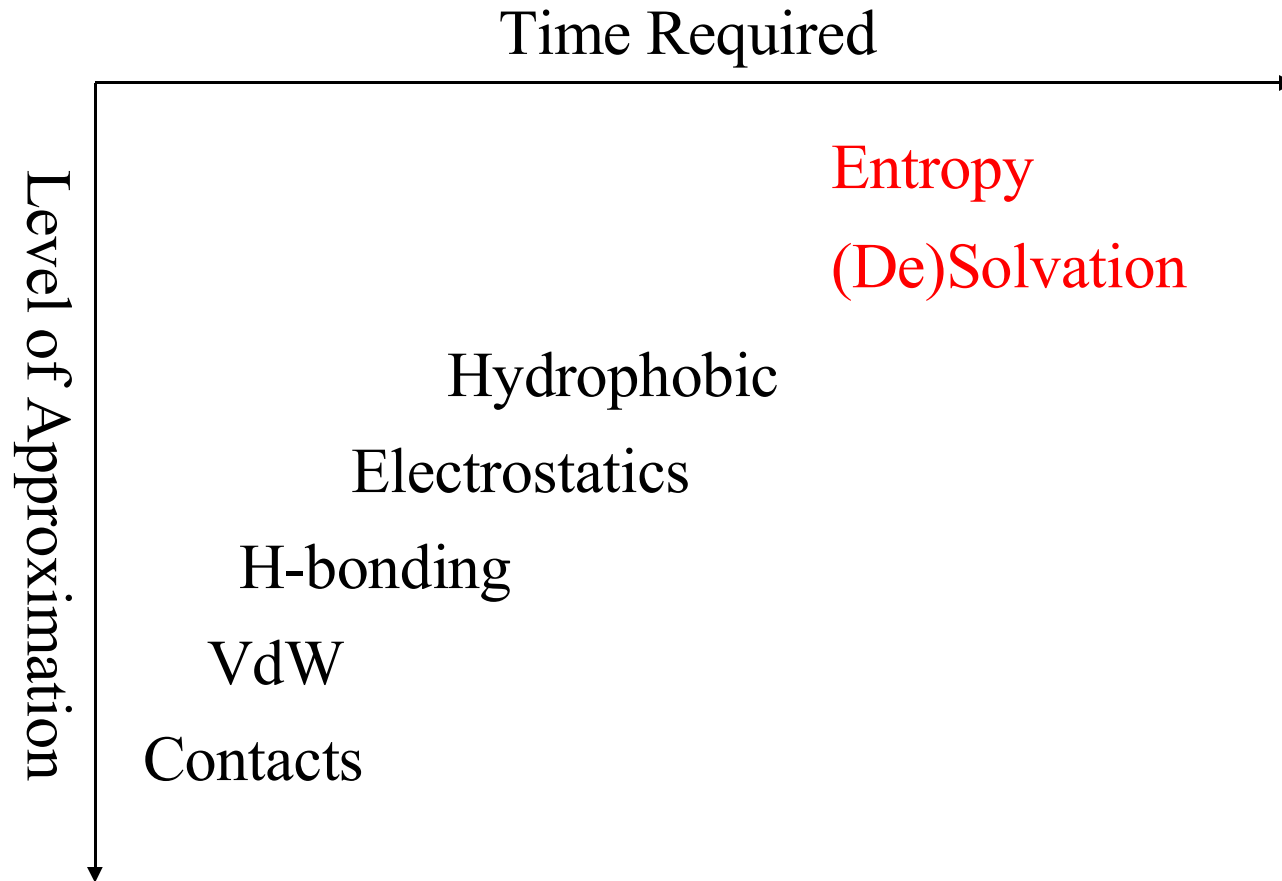


Limitations of Current Methods

- Rigid Protein Side-Chains
- Binding Relevance for Biological Activity
- Scoring Functions



Terms in Scoring Functions



Classifying Scoring Functions

- Knowledge-based (Atom pairs in contact)
 - DrugScore, PMF
- Energy
 - GOLD, DOCK, LigandFit, MOE
- Energy + Parameterised Solvation
 - ChemScore (Glide, THINK)
- Free Energy Perturbation



Knowledge-Based Functions

$$\text{Score} = \sum_{r < \text{cutoff}} A_{ij} (r)$$

- Potentials of Mean Force (PMF)
 - J Med Chem (1999) 42 p791-804
- DrugScore
 - J Med Chem (2005) 48 p6296-6303
- Less Confused by Crystal Structure Precision



Energy

- Lennard Jones

$$\sum_{ij} (A / r_{ij}^{12} - B/r_{ij}^6)$$

- Torsion Term

$$\Sigma (1 - \cos 2\omega) \text{ Conjugated}$$

$$\Sigma (1 + \cos 3\omega) \text{ Non-conjugated}$$

- Electrostatics

$$\sum_{ij} q_i q_j / \epsilon r_{ij}$$



Enhanced ChemScore

$$\Delta G = \Delta G_0 + \Delta G_{\text{hbond}} * N_{\text{hbond}} + \Delta G_{\text{lipo}} * N_{\text{lipo}} + \Delta G_{\text{bad}} * N_{\text{bad}} + \Delta G_{\text{rot}} * N_{\text{rot}} + E$$

where

ΔG_0 ΔG_{hbond} ΔG_{lipo} ΔG_{bad} ΔG_{rot} are constants
(-5.48; -3.34; -0.117; 0.058; 2.56)

N_{hbond} is the number of interactions (using geometric criteria)

N_{lipo} is the number of lipophilic contacts (cf PMF, DrugScore)

N_{bad} is the number of lipophilic-hydrophilic contacts (extension)

N_{rot} is the number of frozen rotatable bonds in the ligand

E is the VdW interaction energy and ligand torsional energy
(extension)



Free Energy Perturbation

$$\Delta G_{\text{sol}} = - \Delta G_{\text{sol}} (\text{ligand}) - \Delta G_{\text{sol}} (\text{protein}) \\ + \Delta G_{\text{gas}} + \Delta G_{\text{sol}} (\text{complex})$$

- Error Prone due to Subtraction of Large Numbers
- Solvent Accessible Surface Area (SASA) approximation (cf ChemScore)
- J Med Chem (2004) 47 p3065-74



General Observations

- Single Electrostatic and Tautomer models have Systematic Failures
- Energy Inadequate for Ranking Series of Diverse Molecules
- Free Energy Perturbation Methods Slow



Excuses for Inaccuracies

- Rigid Side-Chains
- Estimation of Solvation Effects
- Precision of Force Fields
- Biologically Non-relevant Binding
- Kinetics vs Thermodynamics



Ostriches

- Ignoring Fundamental Theory
 - PMF
 - DrugScore
- Omitting Geometry Refinement
 - J Med Chem (2004) 47.12 p3032-47
- Rigid and Semi-Rigid Ligands
 - Dock
 - LigandFit
- Biased Validations
 - J Med Chem (2005)



Performance

- THINK 1.03 (used for CAN-DDO)
 - 42,000,000,000 molecules
 - 126,000 years
 - 900 molecules per CPU day (excluding redundancy)
- THINK 1.30 (current release)
 - Optimised with assistance from Intel
 - Up to **100** times faster
 - More centres useful for larger sites
 - Refinement of docked geometry
 - About **500,000** molecules per 2GHz CPU day



Validation and Results

- Reproduce Ligand-Protein Crystal Structures
 - RMS Deviation of non-H Atoms
 - Docking Score
- Dock Actives
 - Used for Developing Scoring Functions
- Prediction
 - Enrichment over Random
 - Percentage of False Positives



Selection Criteria

- Possible Kinases Cancer Targets
- X-ray Crystal Structures in PDB
- Resolution (1.9-2.8)
- All Atoms (1IAN C α only)
- Ligand Flexibility ≤ 10 Rotatable Bonds
(excludes 1GAG, 1IR3, 2FGI, 5TMP, 1LCK)
- 21 Structures Processed



PDB ID	Score	RMS	Notes
2KI5	-52.2 (-39.8)	5.88 (5.32)	3 A S
1QHI	-78.4 (-73.5)	1.03 (0.98)	3 A
1STC	-105.7	0.59	2
1E8Z	-86.8	1.77	2
1AQ1	-87.0	0.89	2
1AGW	-60.0 (-53.9)	3.45 (0.65)	2
1FGI	-83.8 (-69.4)	0.56 (0.52)	2
1FVV	-81.3 (-73.3)	0.61 (0.56)	3
1FVT	-69.8 (-58.5)	1.44 (0.72)	3

A All Site Points

S Single Bond Increment

C Conjugated Bond Increment

1 Two Centre Fit

3 Three Centre Fit

W Water Site Point



PDB ID	Score	RMS	Notes
1DI8	-55.2 (-52.9)	0.98 (0.70)	2 W
1DI9	-39.5	1.97	2 W C
1YDR	-60.1 (-43.9)	3.00 (2.33)	2
1YDS	-55.5 (-48.1)	2.80 (0.76)	2 W
1YDT	-36.8	2.86	2
1PME	-28.4	1.30	3 W*2 C
4ERK	-17.9 (-11.8)	7.83 (3.33)	2
1IEP	-90.6 (-84.8)	0.73 (0.69)	3
1FPU	-47.1	0.75	3 C
1DM2	-54.0	0.61	3
1CKP	-43.2 (-43.0)	1.99 (1.10)	2 C
1QCF	-58.8	1.33	2



Find-a-Drug Cancer Results

PDB Code	Protein	Number Tested	Number Active
1FLT	VEGFr	3	3
821P	RAS	48	4
1C1Y	RAF	3	1
1FGI	FGFr-1	44	6
1E7U	PI3K	47	5



Find-a-Drug Project

- Processed
 - 270+ protein targets
 - 60+ billion molecules (40-500 million per query)
- Validation using NCI Cancer Data
 - 20% true positive
 - >10x enrichment





Find-a-Drug

- Download THINK software
www.treweren.com
- Virtual Screening
Keith.Davies@treweren.com

