Similarity of Molecular Surfaces:
Tales of Multiple Scales

Ajay N. Jain, PhD
Associate Professor, Cancer Research Institute, Dept. of Biopharmaceutical Sciences, and Dept. of Laboratory Medicine
University of California, San Francisco
ajain@jainlab.org www.jainlab.org

Copyright © 2006, Ajay N. Jain, All Rights Reserved
If we can compare molecules sensibly, what can we show?

1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not

2. N-molecule optimal superpositions can be used as the object of virtual screening; enrichment as good as in docking

3. Distances between vectors of similarities can be used as a fast surrogate for similarity, supporting large computations

4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology

5. Models that cover ¼ of approved drugs work well in virtual screening; surprising degree of selectivity

6. Large-scale modeling of drug targets suggests the potential for a rational pharmacology
Answers to the critical questions about chemicals lie in molecular surfaces

Protein-ligand interactions
- Potency
- Specificity
- Active transport
- Efflux
- Metabolism

Ligand-solvent interactions
- Solubility
- Passive diffusion through lipid bilayers
Specific interactions occur at the junction of molecular surfaces: not between points

Biotin/Streptavidin
- $K_d = 10^{-13.4}$
- Complete set of hydrogen-bonding or salt-bridging interactions
- Extensive hydrophobic packing

2D structure is uninformative

3D pharmacophore is too sparse

Molecular surface interactions explain the binding event
We want to approximate a 3D surface
But we need to compute distances quickly

We will approximate molecules as collections of spheres with fixed radii

- H = 1.2
- C = 1.6
- N = 1.5
- O = 1.4
- S = 1.95
- P = 1.9
- F = 1.35
- Cl = 1.8
- Br = 1.95
- I = 2.15

Hard plastic balls is *not* necessarily a reasonable representation

- But it is fast to compute the distance to a set of spheres
- Hard to compute the distance to a more realistic molecular envelope
- Implication for scoring function softness
Molecules have directional polar contact preferences

We will mark atoms as follows

- **Polar positive:**
  - H-bond donors
  - Formally positively charged atoms

- **Polar negative**
  - H-bond acceptors
  - Formally negatively charged atoms

- **Polar atoms have directional preferences**
  - Defined on a local coordinate system
  - Up to three preference vectors
We have used this representation in scoring functions for docking.

**Co-crystal data used to tune the function**
- 34 structures, ranging from $10^{-3}$ to $10^{-14}$ in $K_d$
- 16 different proteins (heavy on enzymes)

**Linear combination of non-linear functions of protein-ligand atomic surface distances**
- Steric term: Gaussian + sigmoidal
- Polar term: Gaussian + sigmoidal
  - Polar term is influenced by directionality and formal charge
- Entropic term: # rot bonds, log(MW)
- Solvation term: “missed h-bonds”

Empirically derived scoring function learns reasonable magnitudes and geometries

Steric and polar terms dominate

Steric term
- Peaks at about 0.1 log units per ideal contact
- Ends up dominating the energy function because there are so many such contacts

Polar term
- H-bond: peaks at 1.25 log units (H-O distance of 2.0 Å)
- Formal charge scales the polar term: 2.25 units for a tert-amine proton to a carboxylate oxygen

But from what do we compute distance with no protein?
We don’t know where to put the points, so we put them everywhere (spacing $\lambda$)
We can’t have an infinite number of points, so we will light up points at distance $\gamma$. 

\[-0.5, -0.5\]
We define a Gaussian function of distance to weight the pseudo-protein points.

\[ W = e^{-(d-\gamma)^2 / \sigma} \]

We cut off points with weight < 0.1.

This gives us points spaced at about \( \gamma \) from the molecule in question.

We can control the sloppiness of the weight using \( \sigma \).

\[ \lambda = 2.0, \quad \gamma = 4.0, \quad \sigma = 0.2 \]
Molecule pairs: Alignment and similarity computation

Computation of molecular similarity is done based on comparing distances from observer points to the molecular surfaces once the molecules have been brought into an aligned coordinate frame. The similarity is defined on a scale of 0 to 1.

Output: 0.94

Similarity is a Gaussian function of the difference in distances.
We can compare molecular surfaces.
So what can we show?

1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not
2. N-molecule optimal superpositions can be used as the object of virtual screening; enrichment as good as in docking
3. Distances between vectors of similarities can be used as a fast surrogate for similarity, supporting large computations
4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology
5. Models that cover ¼ of approved drugs work well in virtual screening; surprising degree of selectivity
6. Large-scale modeling of drug targets suggests the potential for a rational pharmacology
Scales: computational and physical

1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not.

2. N-molecule optimal superpositions can be used as the object of virtual screening; enrichment as good as in docking.

3. Distances between vectors of similarities can be used as a fast surrogate for similarity, supporting large computations.

4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology.

5. Models that cover 1/6 of approved drugs work well in virtual screening; surprising degree of selectivity.

6. Large-scale modeling of drug targets suggests the potential for rational pharmacology.

How can we compute millions of pairwise similarities fast?

What if we model all relevant biological effectors?

Can we do this directly to big molecules like proteins?

Hundreds of pairs of molecules

4 target models each with 2-3 ligands, screened against thousands of non-ligands

All-by-all comparison of 48 targets and their cognate drugs (10^6 comparisons)

22 target models, screened against thousands of non-ligands and against hundreds of other drugs.

Hundreds of pairs of molecules

4 target models each with 2-3 ligands, screened against thousands of non-ligands

All-by-all comparison of 48 targets and their cognate drugs (10^6 comparisons)
1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not

Compute a separation test

- Set of ligand pairs known to bind the same sites
- Set of ligand pairs thought not to bind the same sites
- Compute the degree of overlap between the distributions


- 134 protein/ligand complexes (> 20 different proteins with multiple ligands)
- 74 related pairs of molecules
- 680 unrelated pairs

1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not.

Morphological similarity
- 3D surface-based method
- Possible to rapidly optimize conformation and alignment

Daylight Tanimoto similarity
- Widely used benchmark
- Computes 2D similarity very efficiently
1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not

At a false positive rate of 0.05, MS yields a 47% reduction in the number of related pairs that are lost.

At a true positive rate of 0.70, MS yields a 7-fold better elimination of false positives.

![Graph showing similarity scores](image)

**Pairs work. Can we build models for virtual screening?**
1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not.

2. N-molecule optimal superpositions can be used as the object of virtual screening; enrichment as good as in docking.

3. Distances between vectors of similarities can be used as a fast surrogate for similarity, supporting large computations.

4. Known drugs and their biological targets are modelled based on similarity; this parallels known pharmacology.

5. Models that cover ¼ of approved drugs work well in virtual screening; surprising degree of selectivity.

6. Large-scale modeling of drug targets and their potential effector molecules like proteins?

- **Hundreds** of pairs of molecules
- All-by-all comparison of 48 targets and their cognate drugs ($10^6$ comparisons)
- 22 target models, screened against thousands of non-ligands and against hundreds of other drugs
- What if we model all relevant biological effectors?

Can we do this directly to big molecules like proteins?
2. N-molecule superpositions can be used as the object of virtual screening; enrichment as good as in docking

Many classes of targets are not currently tractable by crystallography

GPCR and ligand-gated ion channel ligands

♦ Molecules A–B: 5-HT1a ligands
♦ Molecules C–E: Muscarinic antagonists
♦ Molecules F–H: Histamine receptor antagonists
♦ Molecules I–K: GABA<sub>A</sub> receptor agonists

We can induce binding site models by making use of molecular similarity methods (surflec-sim)

Use these models for screening libraries (essentially docking putative ligands to the joint superposition)

How do we make the models?
Ligand-based target models: Induction of model from N ligands

**Input:** Molecule list

The algorithm seeks a joint superposition of all input molecules that maximizes similarity and minimizes overall volume of the superposition.

The alignment optimization method and molecular similarity computation from before is used in the internal search of this procedure.

**Output**

The output is a set of superposition hypotheses and corresponding scores. Each superposition is a specific conformation and alignment of each input molecule within the same coordinate frame.

Hypo 48: score 0.448
Ligand-based target models: Virtual screening using a ligand-based model

Each input molecule is aligned to the molecular superposition (model). Each molecule within the model is used as an object of alignment separately. The score for each alignment is the mean similarity to all molecules within the model. The maximum such score is defined as the score for the input molecule.

For each input molecule, the score is reported, along with information about the ligand (number of atoms and number of rotatable bonds).

- **Pentazocine**: 49 atoms, 3 rot: 0.702
- **Methadone**: 51 atoms, 7 rot: 0.697
- **Mephenytoin**: 30 atoms, 2 rot: 0.551
- **ACD20031**: 26 atoms, 4 rot: 0.564
Cumulative distributions of random molecules are shifted far left of true ligands.

True positive rates of 60% are possible with a FP rate of 2-3%.

Not as good as the best docking results, but competitive with many docking methods.

Much better than 2D indexing methods.

2. N-molecule superpositions can be used as the object of virtual screening; enrichment as good as in docking

This is not simply an artifact of inductive bias from the molecules used for model construction
1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not.

2. N-molecule optimal superpositions can be used as the object of virtual screening; enrichment as good as in docking.

3. Distances between vectors of similarities can be used as a fast surrogate for similarity, supporting large computations.

4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology.

5. Models that cover 1/4 of approved drugs work well in virtual screening; surprising degree of selectivity.

6. Large-scale modeling of drug targets suggests the potential for a rational pharmacology.

- **Hundreds** of pairs of molecules
- 4 target models each with 2-3 ligands, screened against **thousands** of non-ligands
- How can we compute **millions** of pairwise similarities fast?
- All-by-all comparison of 48 targets and their cognate drugs (10^6 comparisons)
- 22 target models, screened against **thousands** of non-ligands and against **hundreds** of other drugs
- What if we model **all** relevant biological effectors?
- Can we do this directly to **big** molecules like proteins?
3. Distances between vectors of similarities are a surrogate for similarity, supporting large computations

**Input:** Molecule list

**Molecular Imprinting**

**Basis Set**

- **Basis 1**
- **Basis 2**
- **Basis 3**
- **Basis 20**

**Output**

Each input molecule is aligned (using the procedure above) to a fixed conformation of each of the basis set molecules. This yields a value from 0-1 for each such alignment and similarity computation. The basis set consists of 20 molecules chosen for diversity. The procedure results in a 20-dimensional vector for each input molecule.

<table>
<thead>
<tr>
<th></th>
<th>B-01</th>
<th>B-02</th>
<th>B-03</th>
<th>B-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>clomipramine:</td>
<td>0.41</td>
<td>0.32</td>
<td>0.38</td>
<td>0.62</td>
</tr>
<tr>
<td>benzphetamine:</td>
<td>0.41</td>
<td>0.38</td>
<td>0.42</td>
<td>0.62</td>
</tr>
<tr>
<td>sertraline:</td>
<td>0.37</td>
<td>0.44</td>
<td>0.50</td>
<td>0.53</td>
</tr>
</tbody>
</table>

A fast surrogate of molecular similarity between these molecules can be computed by computing the Euclidean distance between their corresponding vectors and reporting that difference from 1.
3. Distances between vectors of similarities are a surrogate for similarity, supporting large computations

Correlation coefficient = 0.78

Imprint distance (similarity) computation is very fast

Similar imprints → high similarity

Dissimilar imprints → low similarity
1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not.

2. N-molecule optimal superpositions can be used as the object of virtual screening; enrichment as good as in docking.

3. Distances between vectors of similarities can be used as a fast surrogate for similarity, supporting large computations.

4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology.

5. Models that cover ¼ of approved drugs work well in virtual screening; surprising degree of selectivity.

6. Large-scale modeling of drug targets and their potential selective effector targets can we do this directly to big molecules like proteins?
4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology

Classical pharmacology describes the relationship between 2D molecular subgraphs and pharmacological effects

*Antihistamines*

**Ethanolamines (Prototype: Diphenhydramine):**

The drugs in this group possess significant antimuscarinic activity and have a pronounced tendency to induce sedation. (Goodman and Gilman’s)

The post-genomic era of biology coupled with advances in structure determination demands more specific descriptions of the drugs and the biological targets that underlie pharmacological effects

♦ We should be able to group drug structures based on computable properties instead of English

♦ We should be able to describe and group targets based on specific biology and on computable properties instead of phenomenology

Reference: Cleves, A.E. and Jain, A.N. (in revision). Robust Ligand-Based Modeling of the Biological Targets of Known Drugs
4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology.

About 500 drugs are hierarchically clustered on this axis.

There appear to be sensible groupings.

We know that the muscarinic and histamine receptors share ligands.
4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology.

We observe blocks of drugs sharing the same targets (vertical black stripes).

We observe blocks of targets sharing the same drugs (horizontal black stripes).
4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology.

We observe blocks of drugs sharing the same targets (vertical black stripes).

We observe blocks of targets sharing the same drugs (horizontal black stripes).

But methadone is an opioid agonist!
4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology.

Drugs that share a common target have higher similarity than those that don’t.

Targets with increasing overlap in drugs are increasingly similar based on ligand similarities (correcting for shared ligands).

The blocks in the two-way clustering come from these properties.
Scales: computational and physical

1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not

2. N-molecule optimal superpositions can be used as the object of virtual screening; enrichment as good as in docking

3. Distances between vectors of similarities can be used as a fast surrogate for similarity, supporting large computations

4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology

5. Models that cover 1/4 of approved drugs work well in virtual screening; surprising degree of selectivity

6. Large-scale modeling of drug targets suggests the potential for a rational pharmacology

How can we compute millions of pairwise similarities fast?

All-by-all comparison of 48 targets and their cognate drugs ($10^6$ comparisons)

What if we model all relevant biological effectors?

Can we do this directly to big molecules like proteins?
5. Models that cover ¼ of approved drugs work well in virtual screening; surprising degree of selectivity

We chose 22 targets with the largest number of annotated drugs

♦ Covers about ¼ of approved therapeutics
♦ A–E: proteins within bacterial, viral, and fungal pathogens
♦ F, H, I: involved in cardiac indications
♦ L–O: steroid receptor targets
♦ J, K, P: involved in analgesia
♦ I, J, S, T: GPCR targets
♦ Many confusible groups

![Chemical Structures](image-url)
5. Models that cover ¼ of approved drugs work well in virtual screening; surprising degree of selectivity

We screened against a random set of screening molecules, as with the four targets before

- Very similar performance across the board
- In all but three cases, >70% TP rate with FP < 5%
- Enrichment > 100-fold for 20/22 cases
- Performance as good as the best docking results
- Note: inductive bias makes this easier in some cases, but many have diverse drugs
5. Models that cover ¼ of approved drugs work well in virtual screening; surprising degree of selectivity

We screened all models against the drugs of other targets:

♦ Remarkably similar performance to that from virtual screen
♦ Histamine model perfectly segregates antihistamines away from antimuscarinics that have no histamine receptor effects
♦ Excellent selectivity among the steroid receptor ligands
♦ But what about methadone from before?
5. Models that cover ¼ of approved drugs work well in virtual screening; surprising degree of selectivity

Methadone, along with other opioid ligands are well segregated from non-ligand and non-cognate drugs in the opioid model.

But, methadone shows up very high on the list of predicted muscarinic ligands:
- 90th percentile of drugs against the muscarinic receptor model
- 97th percentile on the screening ligand ranked list
- Methadone is an antimuscarinic: dry-mouth, urinary retention, loss of bowel motility!

So here we see target overlap predicted by the models.
1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not.

2. N-molecule optimal superpositions can be used as the object of virtual screening; enrichment as good as in docking.

3. Distances between vectors of similarities can be used as a fast surrogate for similarity, supporting large computations.

4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology.

5. Models that cover 1/4 of approved drugs work well in virtual screening; surprising degree of selectivity.

6. Large-scale modeling of drug targets suggests the potential for rational pharmacology.

---

**Hundreds** of pairs of molecules

4 target models each with 2-3 ligands, screened against **thousands** of non-ligands.

How can we compute **millions** of pairwise similarities fast?

All-by-all comparison of **48** targets and their cognate drugs (10^6 comparisons).

22 target models, screened against **thousands** of non-ligands and against **hundreds** of other drugs.

What if we model all relevant biological effectors?

Can we do this directly to big molecules like proteins?
6. Large-scale modeling of drug targets suggests the potential for a rational pharmacology

Case 1: Primary Target Overlap
- Classic side-effect: Benadryl makes you sleepy
- Protein structure focused methods will find these if we have lots of structures and good docking methods

Case 2: Tertiary Target Overlap
- Drug of Target A and drug of Target B both hit Target C
- Tricky effect because protein structures won’t necessarily help you here since you won’t screen against the “third shooter”

Case 3: Drug Transport Overlap
- Drugs of Targets A and B share a transporter and affect uptake or excretion of one another
- Classic drug interaction: verapamil increases serum concentration of digoxin

Case 4: Drug Metabolism Overlap
- Drugs of Targets A and B share a metabolic enzyme and affect concentrations of bioactive drug forms
- Classic drug interactions around the CYP enzymes
6. Large-scale modeling of drug targets suggests the potential for a rational pharmacology

Case 1: Primary Target Overlap
- GABAr barbiturate model
  - Finds cognate drugs but also finds SCN antagonists
  - Turns out that the two drug classes have been shown to target one another’s receptors
- Similar story for muscarinic, histamine and mu opioid receptors

Case 2: Tertiary Target Overlap
- GABAr_barb and SCN drugs also hit the NMDA receptor
- Prediction that NSAIDs and nucleoside analogs may share targets in the human apoptotic pathway: Bcl-2 or Bcl-XL

Case 3: Drug Transport Overlap
- β-AR model
  - Finds cognate ligands such as nadolol effectively
  - Also turns up several beta-lactam antibiotics
  - Ample evidence of drug interaction between things like ampicillin and timolol
  - We predict that there is an overlap in absorptive transport

Case 4: Drug Metabolism Overlap
- These are very common and center around promiscuous enzymes, so we see many such effects
- Example: SCN and GABAr_barb appear to converge on CYP2C19
1. Molecule pairs with high similarity are likely to bind the same protein; low similarity not

2. N-molecule optimal superpositions can be used as the object of virtual screening; enrichment as good as in docking

3. Distances between vectors of similarities can be used as a fast surrogate for similarity, supporting large computations

4. Known drugs and their biological targets segregate based on similarity; this parallels known pharmacology

5. Models that cover ¼ of approved drugs work well in virtual screening, suggesting surprising degree of selectivity

6. Large-scale modeling of drug targets suggests the potential for rational pharmacology

---

How can we compute millions of pairwise similarities fast?

All-by-all comparison of 48 targets and their cognate drugs ($10^6$ comparisons)

22 target models, screened against thousands of non-ligands and against hundreds of other drugs

What if we model all relevant biological effectors?

Can we do this directly to big molecules like proteins?
So now suppose we want to compare proteins? We don’t care what the back-side looks like!
We must find concavities and focus on them. We must tweak our grid parameters.

\[ \lambda = 1.0, \gamma = 0.5, \phi = 0.02 \quad \text{(small mol: } \lambda = 2.0, \gamma = 4.0, \phi = 0.2) \]
Protein comparison: Eukaryotic trypsin-like serine proteases (SCOP B.47.1.2)

Procedure

♦ Take 95% similarity reduced Astral SCOP DB
♦ Find B.47.1.2 domains
♦ Perform all vs. all comparison using morphological similarity:
  \[ \lambda = 1.0, \gamma = 0.5, \omega = 0.02 \]
♦ Induce a clustering based on the inferred distances
Trypsin and Factor Xa are highly similar in their active sites

Alignment at right

- Factor Xa onto trypsin
  Sequence identity: 37%
- Generated by morphological similarity
  - Scores 0.90
  - Nearly atom for atom match in P1 pocket, identical positions of catalytic machinery
- Comparison using CE tool for structural superposition
  - ZScore 7.1
Trypsin and Chymase are much less similar in their active sites

Alignment at right

- Chymase onto Trypsin
  Sequence identity: 37%
- Generated by morphological similarity
  - Scores 0.78
  - Note serine vs. aspartate in P1 pocket, no disulphide on pocket lip, proline at bottom
- Comparison using CE tool for structural superposition
  - ZScore 7.2
Dendrogram suggests sensible grouping in terms of ligand specificity
Concluding Remarks

Computational tools based on molecular similarity can be used to understand the relationship between molecular structure and biological activity

- Virtual screening
- Lead identification
- Scaffold hopping

Methodological development of these tools benefits from curated biological activities of small molecules

Very high-speed 3D chemical indexing enables new computations

- All-by-all comparisons of drug structures to establish the correlation between 3D index distance and biological activity
- Target comparison based on ligand structures

Systematic organization of biological activity information can influence the modeling paradigm

- Predictive annotations
- Early identification of possible side-effects
- Toward a rational pharmacology
Surflex-Sim and Surflex-Dock are available free of charge to academics for non-commercial research purposes. Contact: ajain@jainlab.org

Researchers with commercial interests may license Surflex software through BioPharmics LLC (www.biopharmics.com), which retains commercialization rights to the software and underlying intellectual property. Demo copies are easily available for evaluation. Contact: ajain@biopharmics.com
Jain Lab: Chris Kingsley, Jane Fridlyand, Taku Tokuyasu, Lawrence Hon, Barbara Novak, Tuan Pham

CNRS
- Didier Rognan
- Bruno Villoutreix

Other UCSF
- Frank McCormick
- Ken Dill
- Patsy Babbitt

Arris Pharmaceutical (now Celera)
- Mike Ross
- Will Welch
- Jim Ruppert
- John Mount

MetaXen (now Exelixis)
- Mike Ross
- Anwar Ghuloum

BioPharmics LLC
- Ann E. Cleves (Jain)

Ajay N. Jain, PhD    ajain@jainlab.org    http://www.jainlab.org
www.biopharmics.com