Fingerprints in the RDKit

Gregory Landrum

NIBR IT

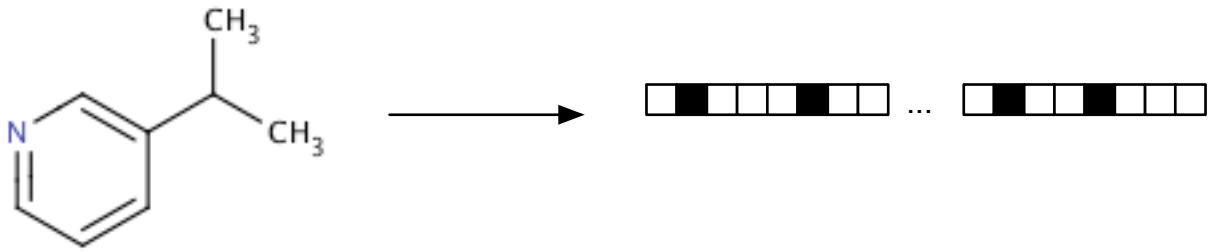
Novartis Institutes for BioMedical Research

Basel

RDKit UGM 2012, London

Molecular Fingerprints

- Idea : Apply a kernel to a molecule to generate a bit vector or count vector (less frequent)



- Typical kernels extract features of the molecule, hash them, and use the hash to determine bits that should be set
- Typical fingerprint sizes: 1K-4K bits.

Calculating similarity between fingerprints

- Most common approach is Tanimoto similarity:

$$Tani(V_i, V_j) = \frac{V_i \bullet V_j}{\sum_b V_{ib} + \sum_b V_{jb} - V_i \bullet V_j}$$

- Shorthand for that:

  $Tani(V_i,V_j) = |V_i \& V_j| / (|V_i| + |V_j| - |V_i \& V_j|)$
- A more general form, Tversky similarity:

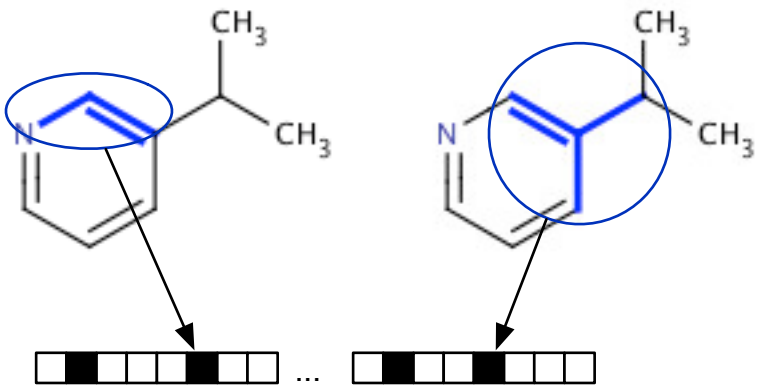
  $Tversky(V_i,V_j,a,b) = |V_i \& V_j| / (a*|V_i| + b*|V_j| + (1-a-b)*|V_i \& V_j|)$
- $Tani(V_i,V_j) = Tversky(V_i,V_j,1,1)$
- $Dice(V_i,V_j) = Tversky(V_i,V_j,0.5,0.5)$

These metrics and others are described and compared here:
JW Raymond, P Willett *JCAMD* **16:**59-71 (2002)

NOVARTIS

Fingerprint similarity == molecule similarity?

- Each fingerprint bit corresponds to a fragment of the molecule

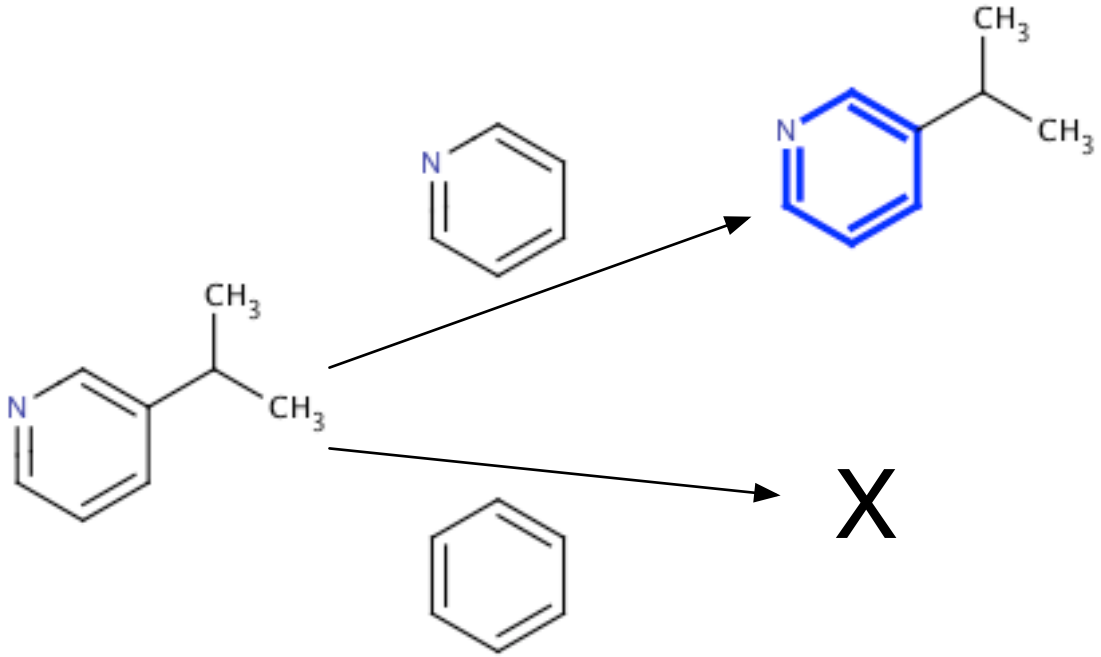


- Assumption: molecules that are similar have a lot of fragments in common
- No “right” answer for defining similarity: there’s no canonical definition of “molecular similarity”
- Lots of experience shows that the best fingerprint for activities like virtual screening (finding active similar molecules in a database) depends strongly on the data set.
- So: there are many different fingerprints available

NOVARTIS

Substructure Searching

- Find a subgraph isomorphism between two molecules (or: find whether or not an isomorphism exists)



- Problem is NP complete, but there are clever algorithms and heuristics to make it tractable (e.g. vf2)

Cordella, L.P., Foggia, P., Sansone, C. & Vento, M. A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**, 1367-1372 (2004).
Ehrlich, H.-C. & Rarey, M. Systematic benchmark of substructure search in molecular graphs - From Ullmann to VF2. *J Cheminf* **4**, 13 (2012).

Substructure Searching at Scale using Fingerprints

- Problem: Practical SSS algorithms work for small numbers of searches, but are too slow for general database querying
- Solution: Use a fingerprinting algorithm to minimize the number of calls to the subgraph isomorphism library:

  Mol_j is a subgraph of Mol_i if and only if every bit set in FP_j is also set in FP_i (i.e. $|FP_j \& FP_i| == |FP_j|$)

- Places severe constraints on the nature of the fingerprinting algorithm
- The more accurate this fingerprint-based screenout is, the faster the overall substructure search will be

NOVARTIS

“RDKit” fingerprints

yet another implementation of a Daylight-like fingerprint

- Substructure fingerprint
- Atom types: set by atomic number and aromaticity
- Bond types: set by atom types and bond types
- Algorithm:

For each subgraph (or path, if `branchedPaths` is false) of length `minPath`-`maxPath` bonds:

1. generate hash for the path using bond types and each bond’s neighbor count
2. seed random-number generator with hash
3. generate `nBitsPerHash` random numbers between 0 and `fpSize` and set the corresponding bits
4. [optional]: “fold” fingerprint to target density of `tgtDensity` of bits set (minimum size `minSize` bits)

NOVARTIS

RDKit layered fingerprints

An alternate subgraph-hashing scheme

- Substructure fingerprint if appropriate layers are used
- Atom and bond types: contributions determined by which layers are included
- Layers:
  - 0x01: pure topology
  - 0x02: bond order
  - 0x04: atom types (atomic number)
  - 0x08: presence of rings
  - 0x10: ring sizes
  - 0x20: aromaticity
- Algorithm: same as RDKit fingerprint

RDKit layered fingerprint 2

An experimental substructure fingerprint

- Substructure fingerprint
- Use a set of pre-defined generic substructure patterns
- Algorithm:
  1. Find all mappings of each pattern onto the molecule
  2. Hash the subgraph defined by that mapping using atom numbers and set a bit
  3. Hash the subgraph defined by that mapping using bond types and set a bit

NOVARTIS

Avalon Fingerprints

highly optimized in-house fingerprint

P. Gedeck, B. Rohde, C. Bartels *J. Chem. Inf. Model*. **46:**1924-36 (2006)

- Substructure or similarity fingerprint (depending on flags)
- Handles query features well (when built from a CTAB)
- Part of the avalon toolkit
  - https://sourceforge.net/projects/avalontoolkit/
  - Optionally useable from within the RDKit

NOVARTIS

Avalon Fingerprints

terms

P. Gedeck, B. Rohde, C. Bartels *J. Chem. Inf. Model*. **46:**1924-36 (2006)

ATOM_COUNT	• Number of double, aromatic, and ring fusion bonds • Number of ring CH2, fusion CH, and spiro carbon atoms • Number of hetero atoms with attached hydrogen • Existence and number of rare hetero atoms
ATOM_SYMBOL_PATH	• single atom paths if non-C,H,N,O two-atom chain paths starting with atoms beyond Ne or ring-fusion atoms and ending with hetero atom • 3- or 4-atom paths starting at hetero atoms but ignoring atom type except for terminal atoms including paths ending in a ring closure (ring paths) • 5- to 7-atom paths and ring paths starting at atom beyond Ne and ending in ring or hetero atom • 2- to 6-atom chain paths starting at hetero atoms with three or more carbon substituent ignoring final atom type • 2- to 4- atom chain paths starting at quarternary atom ignoring terminal atom type • 2- to 4- atom chain and ring paths starting at spiro atom ignoring all but first atom type.
AUGMENTED_ATOM	• allyl or triple bond atoms • pairs of attachments • triples of attachments (setting additional bits for attachments with more than two hetero neighbours or two or more multiple bonds)
AUGMENTED_BOND	• attachment pairs sprouting of both ends of a bond. Note that both ends must have three or more non-hydrogen neighbours
HCOUNT_PAIR	• hydrogen attached to a non-CC-single bond • double/triple bonds with at least one hydrogen
HCOUNT_PATH	• atom type with attached number of hydrogen atoms • 2- to 5-atom chain paths starting at hydrogen bearing non-carbon atom ignoring non-terminal atom types • 3- to 6-atom paths starting at methyl, terminating in a hetero atom and ignoring non-terminal atom types • 1- to 5-atom chain paths starting at hetero atoms with more than one hydrogen ignoring non-terminal atom types
RING_PATH	• 3- to 8-atom paths along ring bonds starting at a ring fusion atom and terminating in a ring closure
BOND_PATH	• 4-atom paths along ring bonds ignoring atom type • 5-atom paths along ring bonds starting at atoms in non-6-membered ring • 4- to 6-atom paths of ring bonds that end in a ring closure • 5-atom paths along ring bonds starting with double or triple bonds
HCOUNT_CLASS_PATH	• 2- to 4-atom acyclic paths starting at carbon atoms with two or more hydrogen attachments and ending in a hetero atom, while mapping all hetero atoms to one class • 2- to 5-atom acyclic paths starting at a hydrogen bearing hetero atom and ending in a hetero atom, while mapping all hetero atoms to one class
ATOM_CLASS_PATH	• 3- to 4-atom acyclic paths along ring bonds, while mapping all hetero atoms to one class • 3- to 9-atom paths ending in a ring closure, while ignoring bond type, non-terminal atom type, and mapping all hetero atoms to one class • non-CC bonds while mapping all hetero atoms to one class • counts of QC and QQ bonds (Q being a heteroatom) ignoring bond type
RING_PATTERN	• 3-atom acyclic paths along ring bonds starting at non-carbon or atom in non-6-membered ring ignoring bond order and mapping hetero atoms to one class • 6- to 17-atom paths along ring bonds terminating in a ring closure and starting from a fusion atom or an atom with more than 2 neighbours
RING_SIZE_COUNTS	• counts bond membership for rings with sizes between 3 and 15+ (rings of size other that 5 or six receive more bits) • presence of bonds connecting rings of different size
DEGREE_PATHS	• 2- to 3-atom acyclic paths labelling atoms with their degree if path starts at atom with one neighbour, a fusion atom, or and atom with more than three neighbours • 2- to 5-atom paths labelling atoms with their degree if path starts at fusion atom bearing a hydrogen • 2-atom paths labelling atoms with their degree if path starts at hetero atom beyond Ne with additional bits set for paths along ring bonds
CLASS_SPIDERS	• tuples (c0, d1, d2, d3) of central atom class and graph distances to one CSP3 and two HETERO atoms or three HETERO atoms for distance up to 7 bonds. CSP3 atoms are carbon atoms with at least three single bonds to carbon
FEATURE_PAIRS	• distances of substituted ring atoms between 5 and 7 • distances of quarternary atoms and hetero atoms between 1 and 8 • distances between substituted ring atoms and querternary atoms between 1 and 6 • distance triples between one non-6-membered ring atom, one other ring atom and a hetero atom

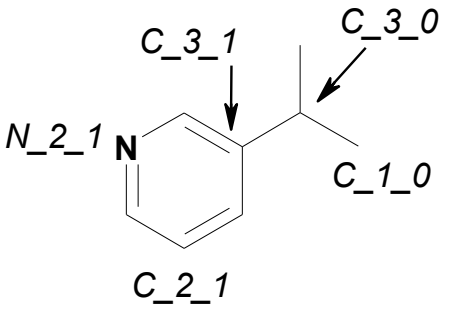
Atom-Pair and Topological-Torsion Fingerprints

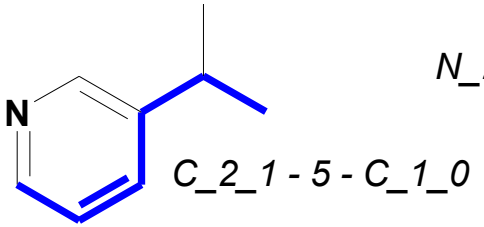
related descriptors from the distant past

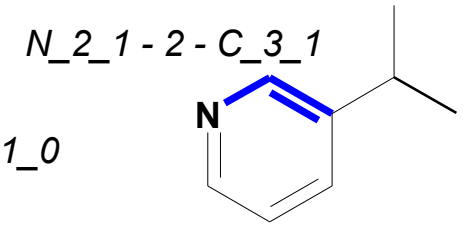
- Similarity Fingerprint
- Atom-type:

  (Element, #heavy neighbors, #pi electrons)
- Atom Pair[1]:

  Atom-type – topological distance – Atom-type

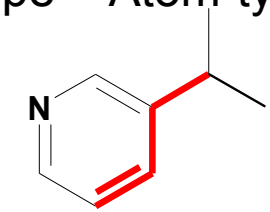


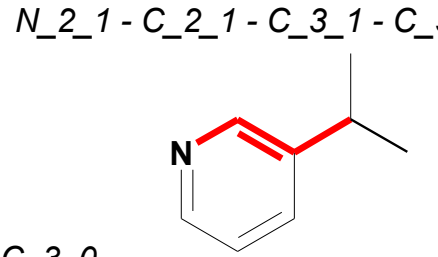




- Topological Torsion[2]:

  Atom-type – Atom-type – Atom-type – Atom-type





1 R.E. Carhart, D.H. Smith, R. Venkataraghavan *JCICS* **25:**64-73 (1985)

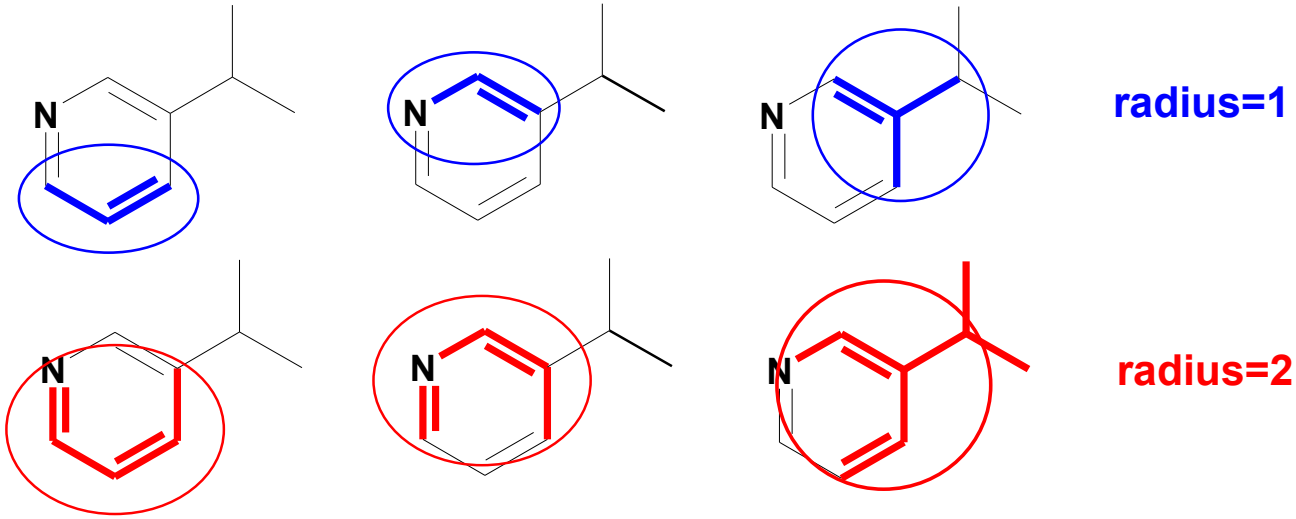
2 R. Nilakantan, N. Bauman, J. S. Dixon, R. Venkataraghavan; *JCICS* **27:**82-5 (1987).

NOVARTIS

Morgan/Circular Fingerprints

new and popular

- Similarity fingerprint
- Atom types :
  - Connectivity: (Element, #heavy neighbors, #Hs, charge, isotope, inRing)
  - Chemical features: Donor, Acceptor, Aromatic, Halogen, Basic, Acidic
- Fingerprint takes into account the neighborhood of each atom:



- Typical radii: 0-3 bonds

Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742-754 (2010).

NOVARTIS

Morgan/Circular Fingerprints

Chemical Feature definitions

- Adapted from A. Gobbi, D. Poppinger *Biotech and Bioeng* **61**:47-54 (1998)

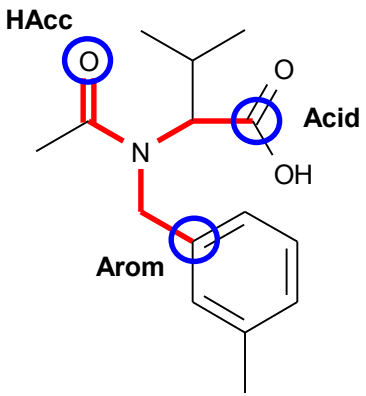
Donor	[$([N;!H0;v3,v4&+1]),\ $([O,S;H1;+0]),\ n&H1&+0]
Acceptor	[$([O,S;H1;v2;!$(*-*=[O,N,P,S])]),\ $([O,S;H0;v2]),\ $([O,S;-]),\ $([N;v3;!$(N-*=[O,N,P,S])]),\ n&H0&+0,\ $([o,s;+0;!$([o,s]:n);!$([o,s]:c:n)])]
Aromatic	[a]
Halogen	[F,Cl,Br,I]
Basic	[#7;+,\ $([N;H2&+0][$([C,a]);!$([C,a](=O))]),\ $([N;H1&+0]([$([C,a]);!$([C,a](=O))])[$([C,a]);!$([C,a](=O))]),\ $([N;H0&+0]([C;!$(C(=O))])([C;!$(C(=O))])[C;!$(C(=O))])]
Acidic	[$([C,S](=[O,S,P])-[O;H1,-1])]

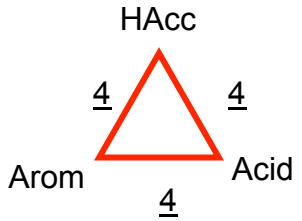
https://sourceforge.net/p/rdkit/code/2174/tree/trunk/Code/GraphMol/Fingerprints/MorganFingerprints.cpp

2D Pharmacophore Fingerprints

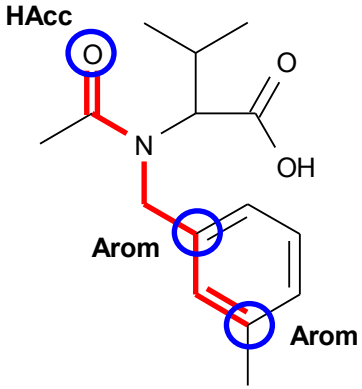
another "blast from the past"

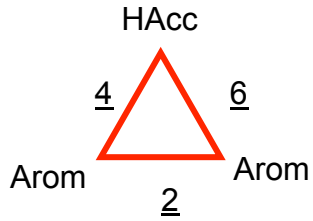
- Identify feature points in a molecule
- Calculate inter-feature topological distances
- Assign bit id to feature – distance combination





- Can be stored as counts or bits
- Feature definitions and distance bins are user-definable





One set of useful feature definitions:

A. Gobbi, D. Poppinger *Biotech and Bioeng* **61**:47-54 (1998)

NOVARTIS

2D Pharmacophore Fingerprints

Supplemental : features + distances -> bit ids

Example: Signature from:

2 Patterns

2 - 3 point pharmacophores

2 distance bins (1,3),(3,8)

Total Signature Size: 38 bits

2 point pharmacophores:

Combos: AA, AB, BB

2 bits/pharmacophore (1 distance with 2 bins)

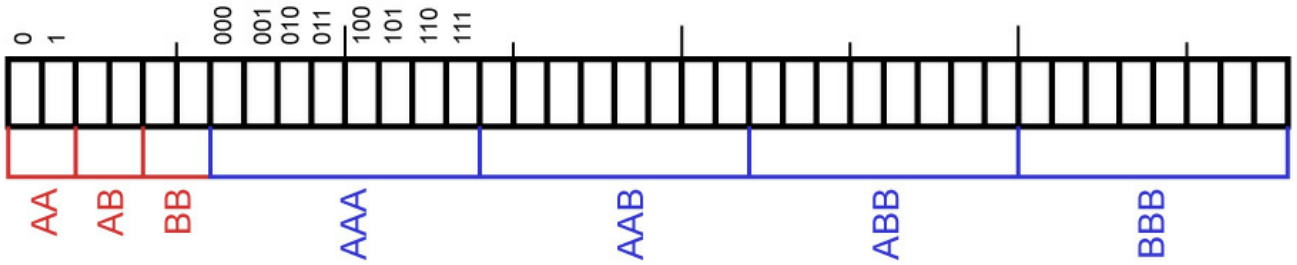
Total: 6 bits

3 point pharmacophores:

Combos: AAA, AAB, ABB, BBB

8 bits/pharmacophore (3 distances with 2 bins)

Total: 32 bits



NOVARTIS

2D Pharmacophores

- Gobbi2d [A. Gobbi, D. Poppinger *Biotech and Bioeng* **61**:47-54 (1998)]
  - Features (from paper):
    - Hydrophobic: `[$([C;H2,H1](!=*)[C;H2,H1][C;H2,H1][$([C;H1,H2,H3]);!$(C=*)]),$(C([C;H2,H3])([C;H2,H3])[C;H2,H3])]`
    - Donor: `[$([N;!H0;v3]),$([N;!H0;+1;v4]),$([O,S;H1;+0]),$([n;H1;+0])]`
    - Acceptor: `[$([O,S;H1;v2]-[!$(*=[O,N,P,S])]),$([O,S;H0;v2]),$([O,S;-]),$([N&v3;H1,H2]-[!$(*=[O,N,P,S])]),$([N;v3;H0]),$([n,o,s;+0]),F]`
    - AromaticAttachment: `[$([a;D3](@*)(@*)*)]`
    - AliphaticAttachment: `[$([A;D3](@*)(@*)*)]`
    - UnusualAtom: `[!#1;!#6;!#7;!#8;!#9;!#16;!#17;!#35;!#53]`
    - BasicGroup: `[$([N;H2&+0][$([C,a]);!$([C,a](=O))]),$([N;H1&+0]([$([C,a]);!$([C,a](=O))])[$([C,a]);!$([C,a](=O))]),$([N;H0&+0]([C;!$(C(=O))])([C;!$(C(=O))])[C;!$(C(=O))]),$([N,n;X2;+0])]`
    - AcidicGroup: `[$([C,S](=[O,S,P])-[O;H1])]`
  - Distance bins (GL): `[(2,3),(3,4),(4,5),(5,6),(6,7),(7,8),(8,100)]`
  - "fuzzing" from original paper not done
- Pharm2d:
  - Feature definitions from BaseFeatures.fdef :

    Donor, Acceptor, NegIonizable, PosIonizable, ZnBinder, Aromatic, Hydrophobe, LumpedHydrophobe
  - Distance bins: `[(2,3),(3,4),(4,5),(5,6),(6,7),(7,8),(8,100)]`

NOVARTIS

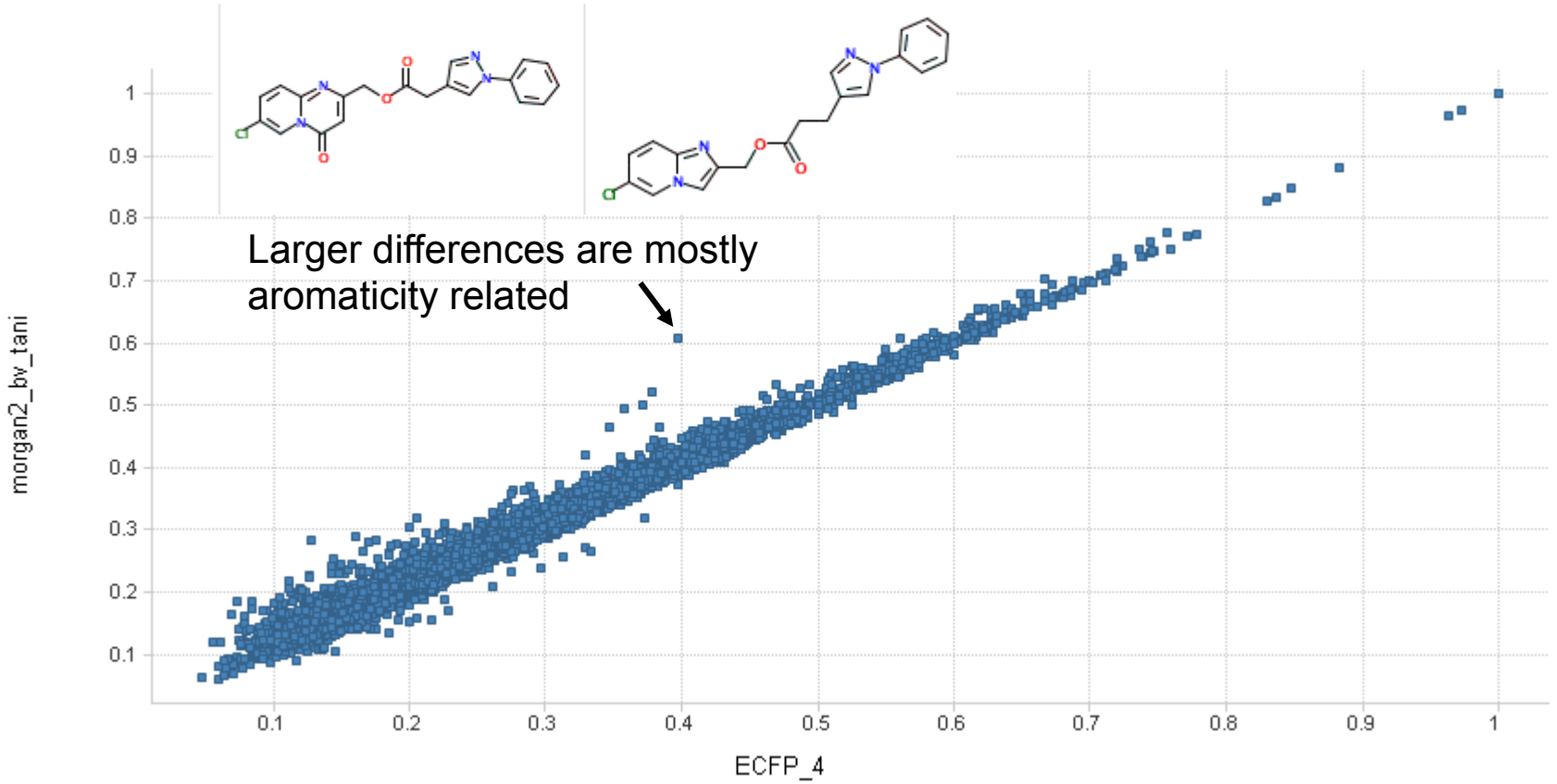
Comparing fingerprints

- Pick 10K random pairs of vendor compounds that have at least some topological similarity to each other (Avalon similarity >= 0.5)
- Compare similarities calculated with Pipeline Pilot and the RDKit

NOVARTIS

Comparing fingerprints

- RDKit Morgan2 vs PP ECFP4

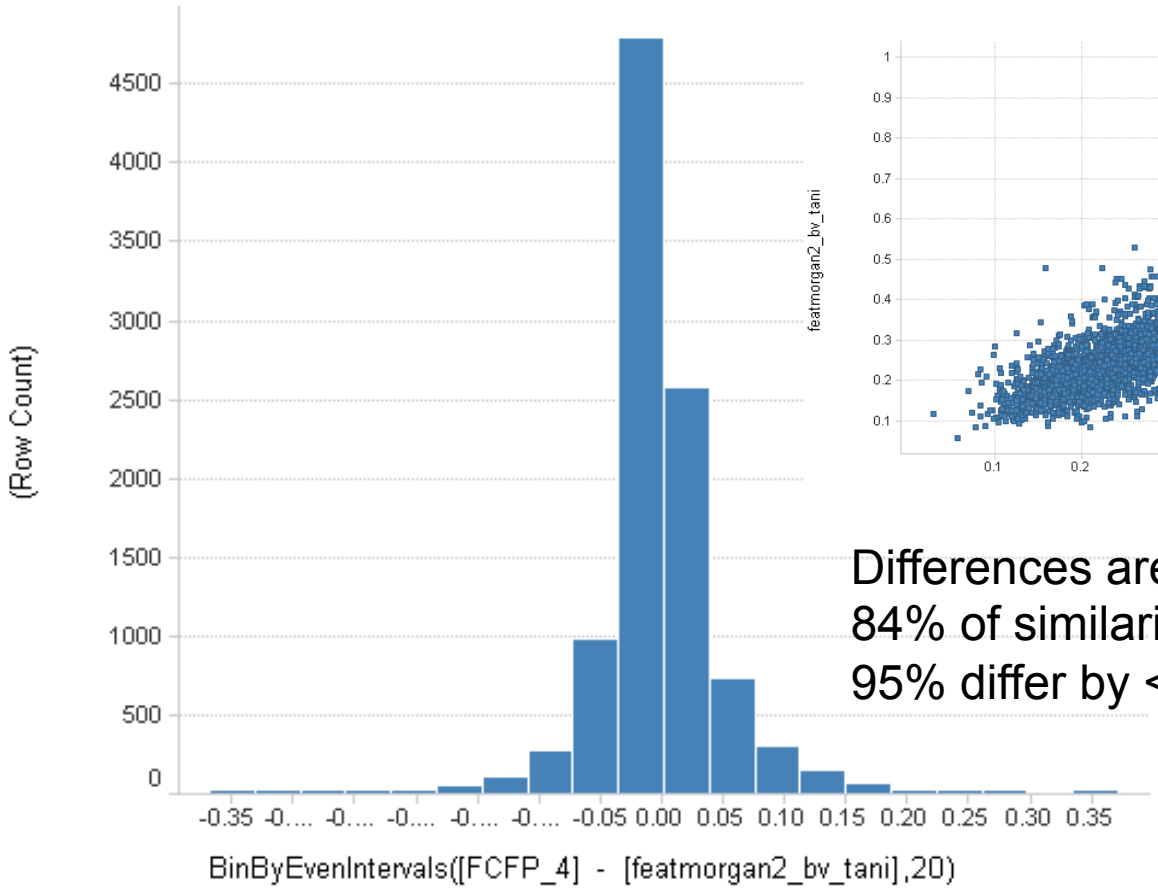


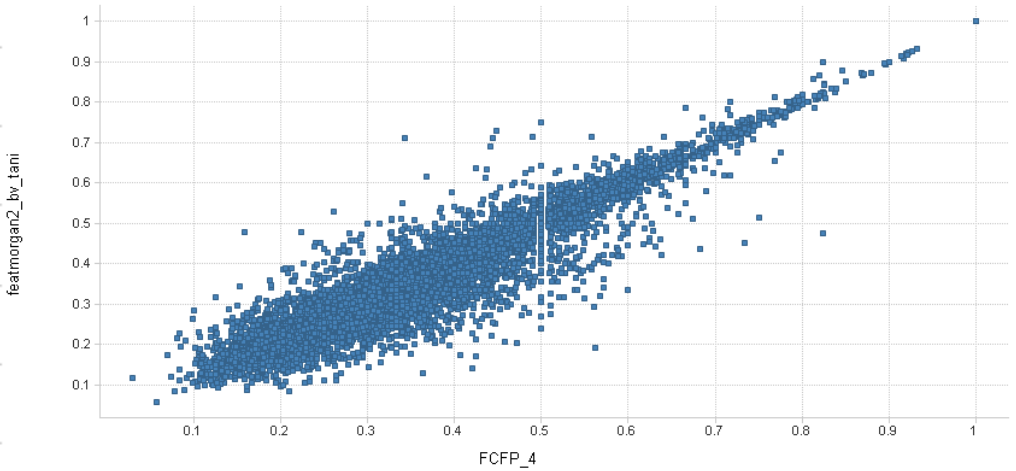
- RDKit Morgan3 vs PP ECFP6 is similar

NOVARTIS

Comparing fingerprints

- RDKit FeatMorgan2 vs PP FCFP4





Differences are due to feature definitions
84% of similarities differ by <0.05
95% differ by <0.1

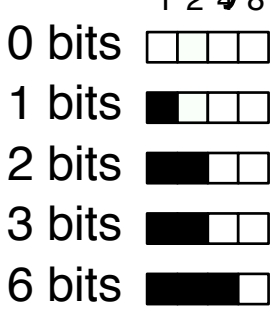
NOVARTIS

Simulating count-based fingerprints

- Sometimes it makes sense to count the number of times a feature appears instead of simply that it appears -> count-based fingerprints
- “Dice” similarity for count-based fingerprints:

$$Sim(V_i, V_j) = \frac{2.0 * \sum_b \min(V_{ib}, V_{jb})}{\sum_b V_{ib} + \sum_b V_{jb}}$$

- Problem: count vectors take up more disk space/memory than bit vectors and similarity calculations using count vectors are slower.
- Partial solution: simulate counts by including multiple bits per feature:



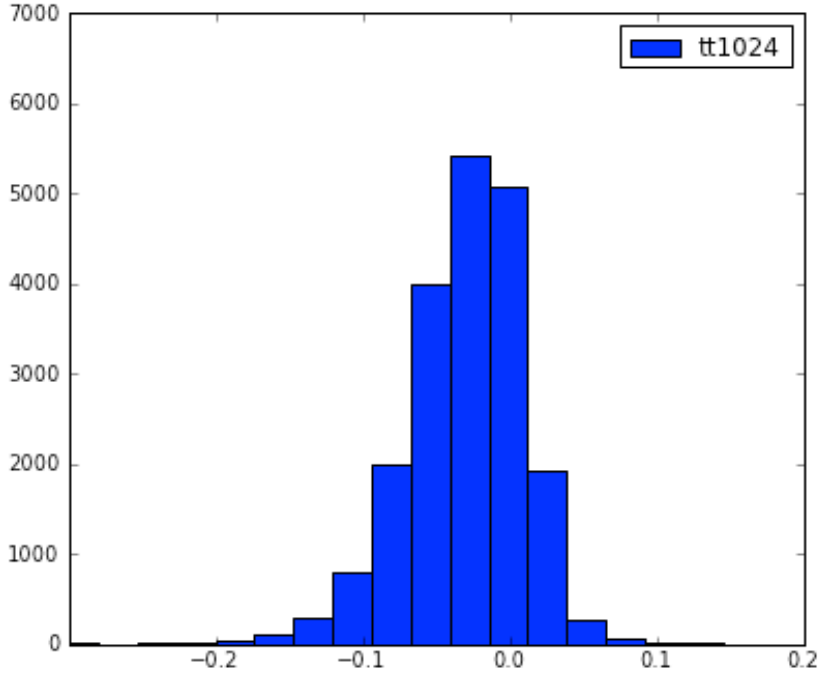
NOVARTIS

Simulating count-based fingerprints

How well does it work?

- Dataset: 20K pairs of “drug like” commercial compounds selected to have a minimum topological similarity (based on Avalon fingerprint)
- Compare unhashed count-based topological-torsion fingerprint to a hashed bit vector version

#bits	99%	95%	90%	80%
4096	0.13	0.09	0.07	0.05
2048	0.13	0.09	0.07	0.05
1024	0.14	0.10	0.08	0.06
512	0.16	0.12	0.10	0.07



NOVARTIS

SSS Screening performance of fingerprints

- Database: 20K diverse ZINC drug-like molecules
- "Fragments" queries: 500 diverse fragment-like molecules from ZINC
- "Leads" queries: 500 diverse lead-like molecules from ZINC
- “Pieces” queries: 823 pieces constructed by doing a BRICS fragmentation of a set of molecules from the pubchem screening set. Size range from 1->64 atoms

- Metric: what fraction of the fingerprint matches actually are substructure matches

fp	zinc_frags	zinc_leads	pubchem_pieces
avalon	0.13	0.22	0.41
layered	0.08	0.10	0.35
layered2	0.78	0.26	0.64

Expt date: 07.07.2012

NOVARTIS