INTERESTING APPLICATIONS OF (CHEMICAL) GRAPH EDIT DISTANCE

Roger Sayle and John Mayfield

NextMove Software, Cambridge, UK

ACS Spring 2023, CINF Division, Indianapolis, IN, USA, Sunday 26th March 2023

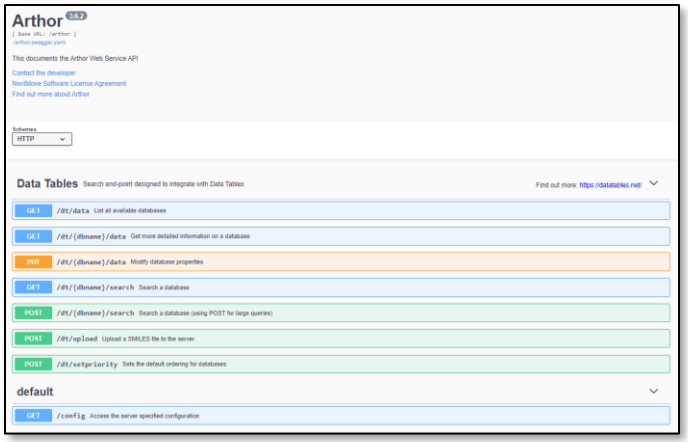
CHEMISTRY DATABASE APPLICATION PROGRAMMING INTERFACES (APIS)

Available Databases, Workflows and Use-Cases.

- Modern ultra-large chemical databases such as Enamine REAL, WuXi GalaXi, ZINC and SAVI are becoming so large, that for most researchers remote access via RESTful APIs are the only way to query them.
- Chemical similarity search has evolved to adapt to these scales, presenting new challenges and opportunities for APIs and their use-cases.

ARTHOR/SMALLWORD API DOCUMENTATION ON THE INTERNET

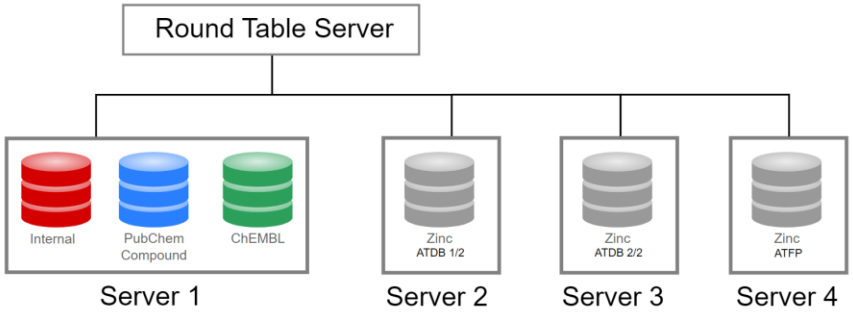
- Arthor's RESTful API, including arthor-swagger.yaml is hosted by the UCSF ZINC folks at:
- https://arthor.docking.org/api.html



- An (unofficial) python wrapper for the SmallWorld API written by Metteo Ferla at Oxford University, and accessing UCSF ZINC's sw.docking.org in on github
- https://github.com/matteoferla/Python_SmallWorld_API

EXAMPLE RESTFUL API USE-CASE

- **RoundTable** is a API that "federates" multiple **Arthor** chemical database (similarity/substructure) servers.
- Because RoundTable's API is the same as Arthor's, RoundTable can recursively federate instances.
  - Care needs to be taken to avoid cycles!

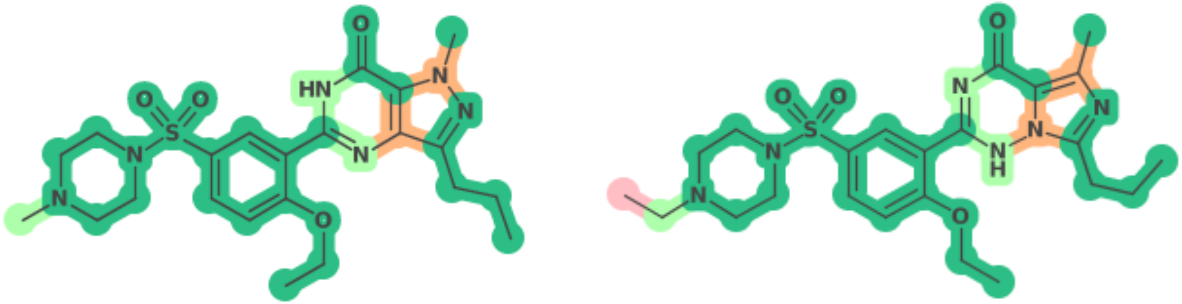


Example Heterogenous Configuration:
ZINC CartBlanche 2023 (35B mols)
Server 1: 1TB RAM (17B mols, 5DBs)
Server 2: 512GB RAM (10B mols, 3DBs)
Server 3: 768GB RAM (4B mols, 1DB)
Server 4: 1TB RAM (4B mols, 1DB)

https://www.youtube.com/watch?v=brxp1x2NV6A

GRAPH EDIT DISTANCE (AND SMALLWORLD) IN 30 SECONDS

- Graph Edit Distance (GED) is a chemical similarity metric that counts the minimum number of edits required to transform one molecule into another; related to Maximum Common Substructure (MCS).

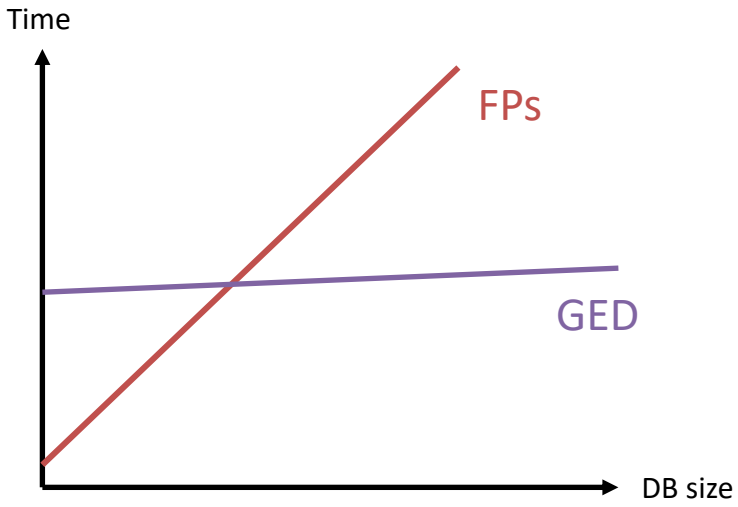


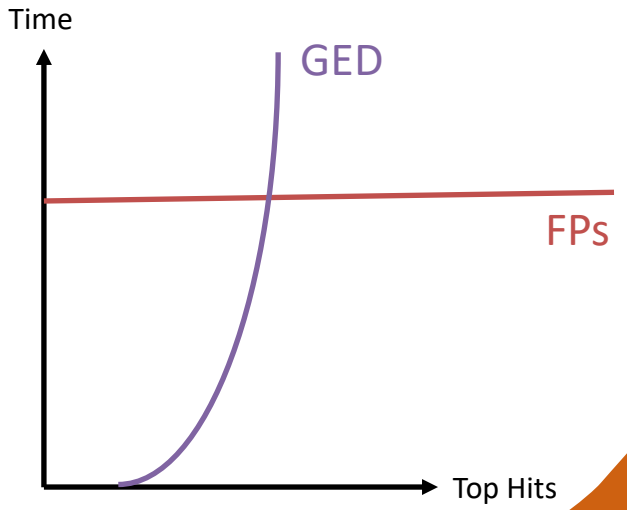
sildenafil to vardenafil = distance 5

- SmallWorld is an efficient implementation of GED/MCS requiring 44TB of pre-computed data.

THE GED VS FP TANIMOTO TRADE-OFF

- GED is almost constant cost ~O(1), where FPs scale almost linearly ~O(N).
- GED close neighbors are instantaneous, but distant neighbors take much longer, where FP results take same time.

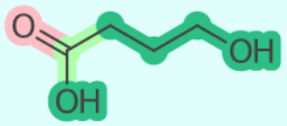




CHEMICAL SIMILARITY IN US COURTS

- In “U.S. v. Brown (2005)”, the Eleventh Circuit of the US Court of Appeals rejected “Tanimoto Similarity” for determining legal term “substantially similar”.

CHEMBL171623
SWIDX: B5R0.2
MW: 90.121
MF: $C_4H_{10}O_2$



CHEMBL1342
SWIDX: B6R0.4
MW: 104.105
MF: $C_4H_8O_3$

- butane-1,4-diol is substantially similar to controlled substance GBH (gamma-hydroxybutyric acid).
  - ECFP4 Tanimoto 0.365, Graph Edit Distance = 1.

415 F.3d 1257 (2005)

ASYNCHRONOUS APIS AND POLLING

- The fundamental change in mind-set is that “full” result sets are too large to download/return and that similarity searches never finish.
- SmallWorld’s API is stateless, polling, robust and speculative.

Invocation

```
curl -d 'smi=c1ccccc1C(=O)O' -d db=chembl_31 'http://nextmove/smallworld/search/view'
```

First result

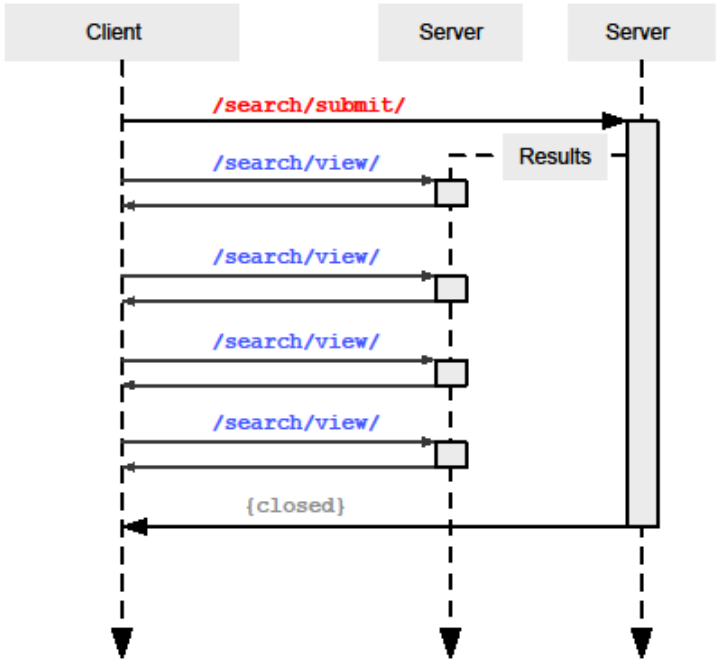
```
{ "recordsTotal" : 0, "status" : { "dist" : 0, "state" : "RUNNING",… },
  "data" : [ ]}
```

Second result

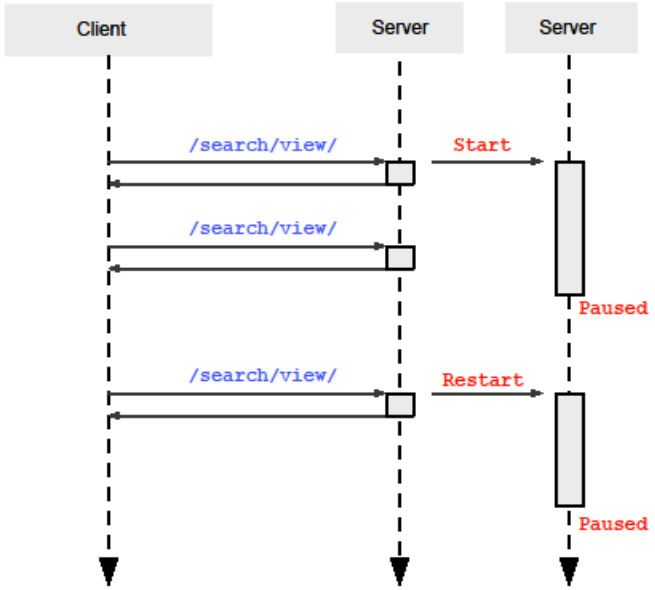
```
{ "recordsTotal" : 24676, "status" : { "dist" : 5, "state" : "WAITING", …},
  "data" : [ [ "C=C(C)c1ccccc1 CHEMBL1344773", 0 ],
             [ "CB(O)c1ccccc1 CHEMBL78562", 0 ],
             [ "CC(=O)N1CCCCC1 CHEMBL12196", 0 ],
             [ "CC(=O)N1CCNCC1 CHEMBL2448844", 0 ], … ]}
```

SESSIONS (SERVER SOCKETS) VS POLLING

Old-style API



New-style API



FEATURES OF THE ASYNCHRONOUS API

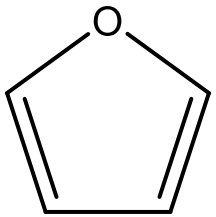
- As the results are deterministic/stable, the API can avoid stateful “session/context” management.
- Instead the server maintains a cache of recent queries, creating a new session if required.
- The server performs work for after each query/poll for a specified timeout.

USE CASE: PROPERTY PREDICTION

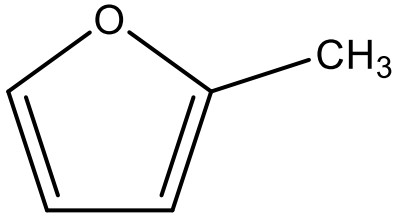
- One method for estimating the physical properties of a compound, is to identify the closest analogue for which the experimental value has been measured and interpolate from there.
- This is much like a 1NN or kNN classifier in machine learning.
- The better/more relevant the distance measure, the more accurate/predictive the model.
- GED outperforms ECFP Tanimoto on top-1 Briem and Lessel bioactivity classification (normally top-10).

USE CASE: PROPERTY PREDICTION

- Knowledge of the edit operations to the nearest experimental neighbour allows more than averaging.



furan
Expt logP: 1.34
Pred logP: 1.36



2-methylfuran
Expt logP: 1.85
Pred logP: 1.87

- For logP, we can determine corrections for Leo and Hansch π values.

USE CASE: VIRTUAL SCREENING

- For each database hit, SmallWorld can return an atom-atom mapping to the query molecule.

Invocation

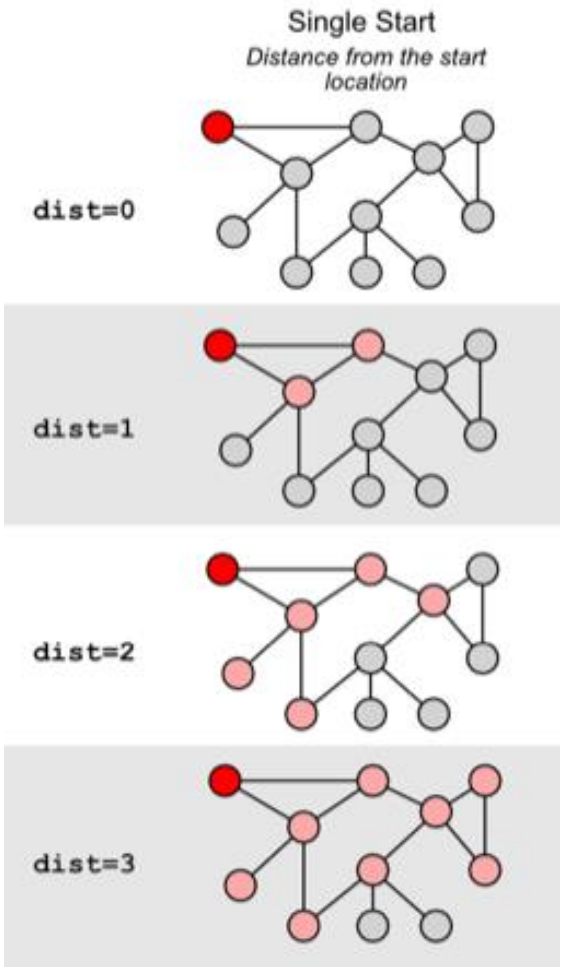
```
curl -d 'smi=C=1C=C(C(=C(C1)Cl)Cl)N2CCN(CC2)CCCCOC=3C=CC4=C(C3)NC(=O)CC4' \
     -d db=chembl_31 -d scores=AtomAlignment
```

Result Data

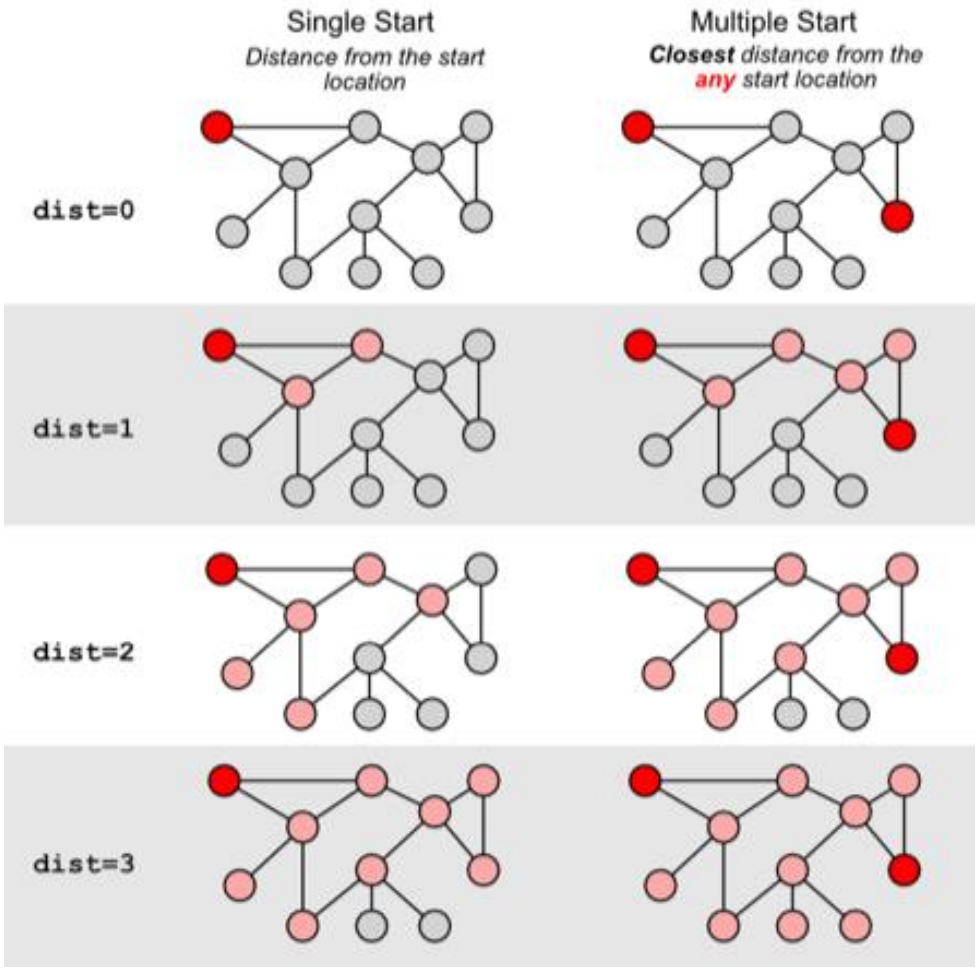

```
{    "hitSmiles" : "O=C1CCc2ccc(cc2N1)OCCCCN3CCN(CC3)c4c(cccc4Cl)Cl CHEMBL158929",
     "qrySmiles" : "C=1C=C(C(=C(C1)Cl)Cl)N2CCN(CC2)CCCCOC=3C=CC4=C(C3)NC(=O)CC4",
     "qryMappedSmiles" :
"[CH:1]=1[CH:2]=[C:3]([C:4](=[C:5]([CH:6]1)Cl)[Cl:8])[N:9]2[CH2:10][CH2:11][N:12]([CH2:
13][CH2:14]2)[CH2:15][CH2:16][CH2:17][CH2:18][O:19][C:20]=3[CH:21]=[CH:22][C:23]4=[C:24
]([CH:25]3)[NH:26][C:27](=[O:28])[CH2:29][CH2:30]4",
     "hitMappedSmiles" :
"[O:28]=[C:27]1[CH2:29][CH2:30][C:23]2=[CH:22][CH:21]=[C:20]([CH:25]=[C:24]2[NH:26]1)[O
:19][CH2:18][CH2:17][CH2:16][CH2:15][N:12]3[CH2:11][CH2:10][N:9]([CH2:14][CH2:13]3)[C:3
]4=[C:2]([CH:1]=[CH:6][CH:5]=[C:4]4[Cl:8])Cl",
     … },
```


- This can be used to perform 3D RMS superposition.

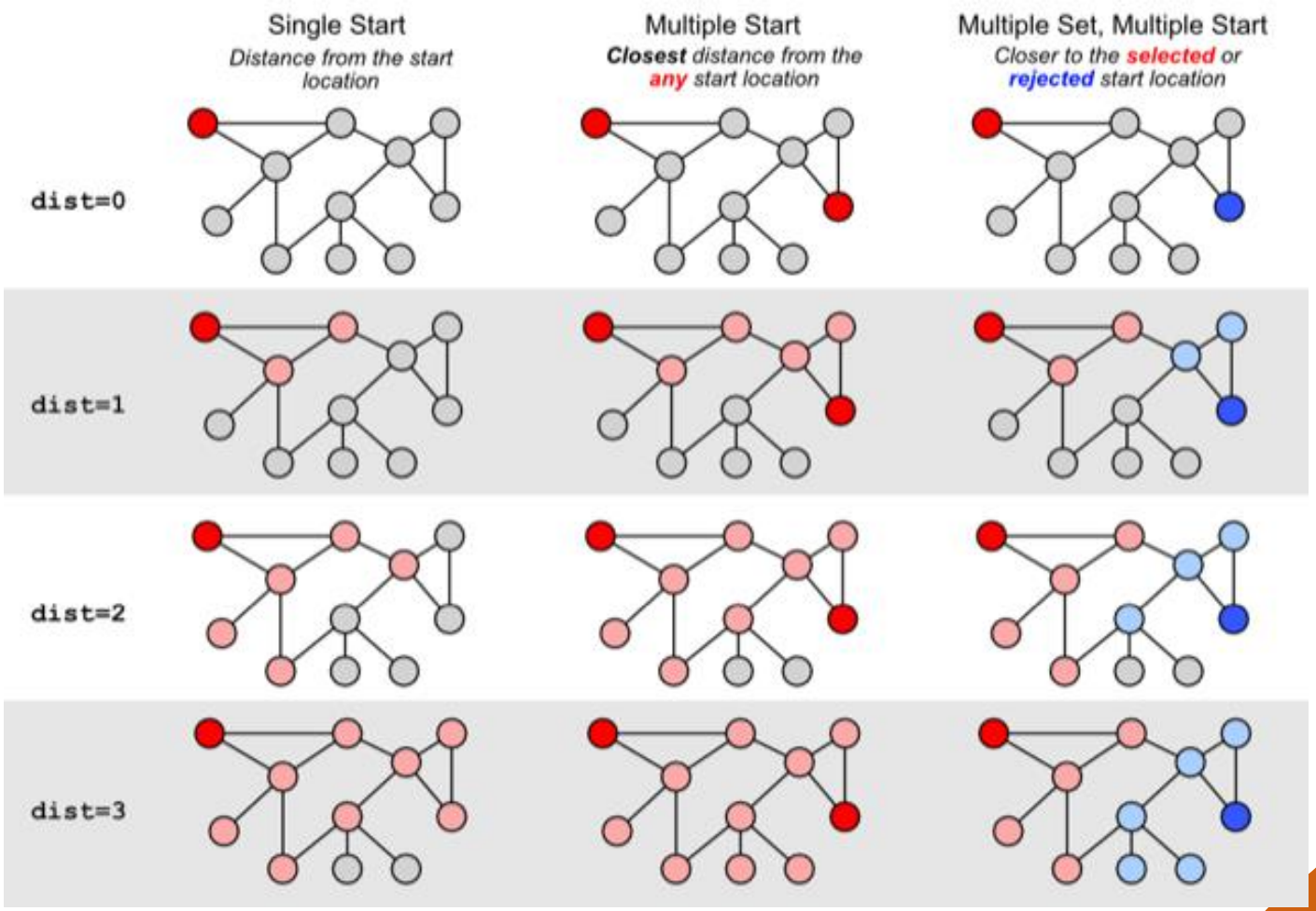
MULTI-SOURCE SEARCH AND INACTIVES



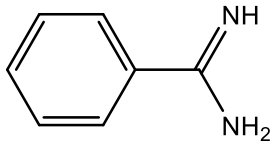
MULTI-SOURCE SEARCH AND INACTIVES



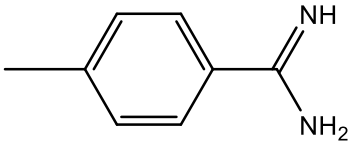
MULTI-SOURCE SEARCH AND INACTIVES



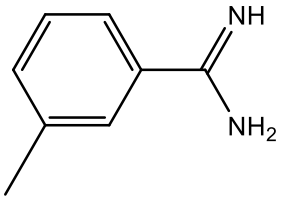
GED-DEFINED ACTIVITY CLIFFS



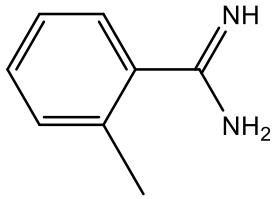
Active Lead Compound



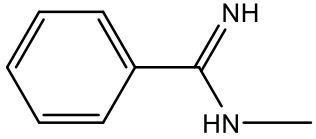
A1: Active



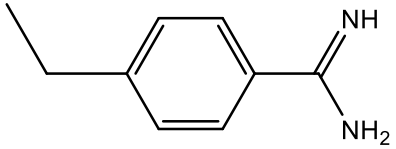
A2: Active



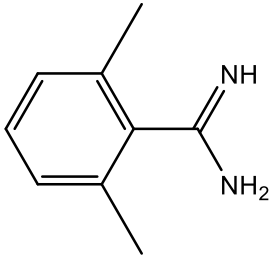
A3: Inactive



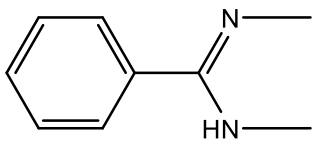
A4: Inactive



Predicted: Active



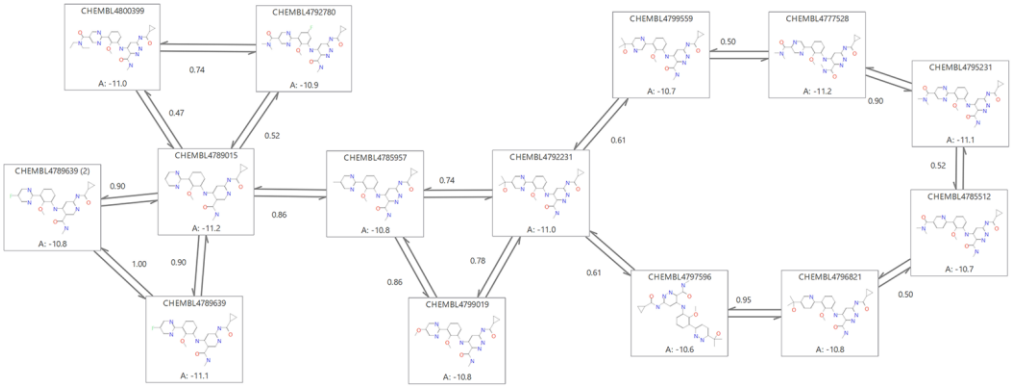
Predicted: Inactive



Predicted: Inactive

FEP AND ITERATIVE LOMAP

- Alchemical Free Energy Perturbation (FEP) constructs calculation networks with thermodynamic cycles.
- LOMAP uses distance matrices and N^3 algorithms.

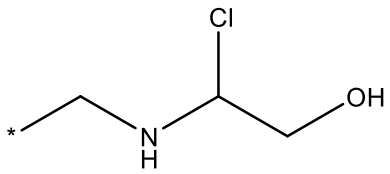


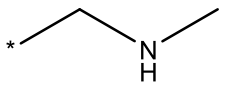
- Incremental neighbour methods (using GED) allows exploration of ultra-large libraries.

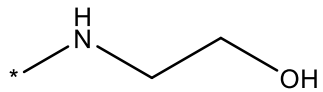
Image courtesy of Cresset

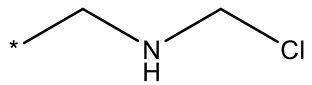
R-GROUP EDIT DISTANCE

- GED can be extended to anchored graphs.

A= 

B= 

C= 

D= 

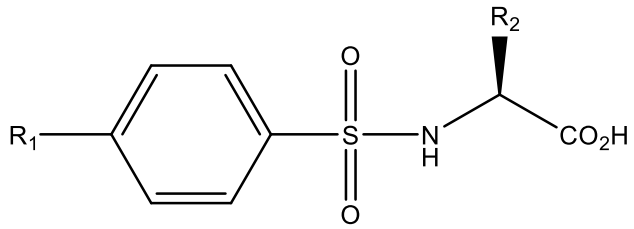
Distance Matrix

	A	B	C	D
A	0	3	2	2
B	3	0	3	1
C	2	3	0	3
D	2	1	3	0

Min Cost Order
B-D-A-C (5)

ADDITIVITY AND RETROSYNTHESIS

- A significant application of R-group edit distance is in retrosynthesis and virtual library search, thanks to the additivity properties of edit distances.
- GED(A-scaffold-B,A'-scaffold-B) = RGrpED(A,A') + RGrpED(B,B')



R1	R2
A1	B1
A2	B2
A3	B3
A4	B4
A5	B5
…	…

Hits
A1-scaffold-B1
A1-scaffold-B2
A2-scaffold-B1
A2-scaffold-B2
A1-scaffold-B3
…

MADWORLD PEPTIDE UNIVERSE RESULTS

```
46 AA, MaxDist=421, Hits=70368744177664000000000000000000000000000000000000000000000000
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN   0 (100.00%) 0
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIVPGATCPGDYAN   1 ( 97.83%) 1
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIVIPGATCPGDYAN   1 ( 97.83%) 2
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCVIIPGATCPGDYAN   1 ( 97.83%) 3
TTCCPSIVARSNFNVCRLPGTPEAVCATYTGCIIIPGATCPGDYAN   1 ( 97.83%) 4
TTCCPSIVARSNFNICRLPGTPEAICATYTGCIIIPGATCPGDYAN   1 ( 97.83%) 5
TTCCPSIIARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN   1 ( 97.83%) 6
TTCCPSVVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN   1 ( 97.83%) 7
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIILPGATCPGDYAN   2 ( 97.83%) 8
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIVVPGATCPGDYAN   2 ( 95.65%) 9
TTCCPSIVARSNFNVCRLPGTPEAICATYTGCILIPGATCPGDYAN   2 ( 97.83%) 10
TTCCPSIIARSNFNVCRLPGTPEAVCATYTGCVIIPGATCPGDYAN   3 ( 93.48%) 100
TTCCPSIVARSNFNICRMPGTPEAICATYTGCIILPGATCPGDYAN   5 ( 93.48%) 1000
TTCCPSIVARSNFNICRLPGTPEAICATYTGCIVVPGATCPGDYVN   7 ( 91.30%) 10000
TTCCPSIVARSNFNICKLPGTPEAICATYTGCIIIPGATCPGDYRN   9 ( 93.48%) 100000
TTCCPSIVAKSNFNICRLPGTPDTICATYTGCIIIPGATCPGDYAN  11 ( 91.30%) 1000000
TTCCPSIYAQSNFNVCRVPGTPEAICATYTGCVIIPGATCPGDYAN  13 ( 91.30%) 10000000
TTCCPDIVARSNFNVCRLPGTPEALCATYTGCIIIPGKTCPGDFAN  15 ( 91.30%) 100000000
TTCCPSIVAHSTFNVCRLPGTPQAICATYTGCIIIPGATCPGEYAN  18 ( 91.30%) 1000000000
TTCCPALMARENFNVCRLPGTPEAICATYTGCCVLPGATCPGDYAN  20 ( 84.78%) 10000000000
TTCCPSIIARNDFNICRLPGTPEALCGTYTGCVIVPGANCPGDYAN  23 ( 80.43%) 100000000000
STCCPAIVARSNINICRMPGTPEAICAGYTGCVIVPGATCPGDYAN  25 ( 82.61%) 1000000000000
TTCCPTLTARSNFNVCRLPGTPDAICATFEGCVTIPGATCPGDYAN  28 ( 82.61%) 10000000000000
TTCCPSAMARSNFNVCKMPGTPEAICVTYTGCIVIPGATCPQDYAH  31 ( 82.61%) 100000000000000
TTCCPSLIANKNFNVCALPGTPQAYCACYTGCVVIPGATCPGDYAN  34 ( 78.26%) 1000000000000000
TTCCPSMIARSNFNNCRIPGTPEAQCATYTGCVYLPGVTCPGDYIN  37 ( 78.26%) 10000000000000000
TTCCPNIMCRSNSNVCRLPVHPEAMCATYTGCLIVPGATCPGDYAN  40 ( 80.43%) 100000000000000000
TTCCPPVIAKLDFNVCRIPGPPEAYCATYNGCVVLPGATCPGDYAN  43 ( 71.74%) 1000000000000000000
ATCCPDIIIRSEFNVCRLPGNPRAICETYTGCIILPGSTCPGDYQN  46 ( 76.09%) 10000000000000000000
GGEEFWGRWCWWPWRECDFIGFCWGEWGDGIEGGGFIWGEFILDWW 421 (  0.00%) -259200000000
YYEEWWGWWVWWPWWEVGWLYWCWGEWYPYLEGGGWLWYEWLWPWW 421 (  0.00%) -1
```

CONCLUSIONS

- The scale of some scientific challenges requires a centralization of computational resources, such that next generation applications will be client-server.
- The use of chemical similarity based on graph edit distance is one example, is one example, allowing better scaling to modern ultra-large databases.
- These RESTful APIs enable novel use cases and workflows that were not possible (or overlooked) with previous technologies.

ACKNOWLEDGEMENTS

Andrew Grant, AstraZeneca
John Irwin, ZINC group, UCSF
Yurii Moroz, Enamine and ChemSpace
Darren Green, GSK
Pat Walters, Relay Therapeutics
Evan Bolton, PubChem Group, NCBI
Marc Nicklaus, NCI
Gergely Zahoranszky-Kohalmi, NCATS
Noel O'Boyle, Heptares
Andrew Dalke, Dalke Scientific Software
Christos Nicolaou, Recursion (ex. Eli Lilly)
Oliver Korb, Roche
Jose Batista, OpenEye Scientific Software
Jameed Hussain, Dotmatics

NextMove Software

MICROSERVICES AS APP COMPONENTS

- Modern web-based graphical interfaces are often composed of RESTful microservices.
- The text input box in the Arthor and SmallWorld apps can call out to an identifier resolver (IDR) service, than can provide both identifier lookup and name-to-structure (NCI/CADD resolver API compatible).
- The structure depictions in the Arthor and SmallWorld apps (by default) use a depiction service based on CDK. Drag & Drop (to JSME) also uses this service.