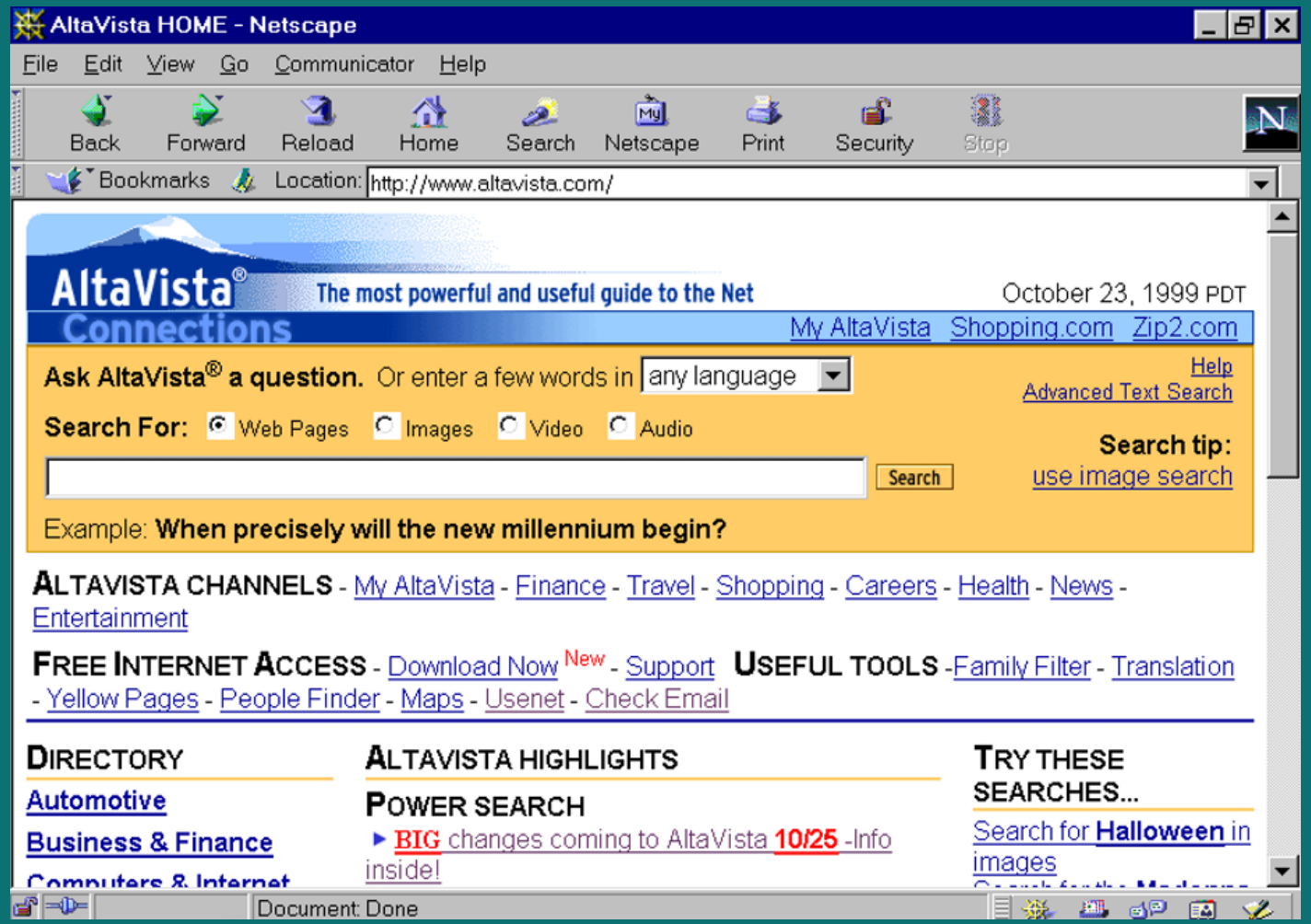
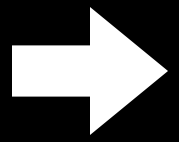
# MHFP: A Probabilistic Molecular Fingerprint for Big Data Settings



I stand at the seashore, alone,
and start to think.
There are the rushing waves
mountains of molecules
each stupidly minding its own
business
trillions apart
yet forming white surf in unison.

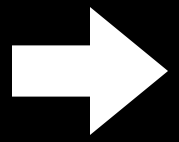


…

| $h_1$ | $h_n$ | $h_{min}$ |
|---|---|---|
| 3453409 | 475834 | 475834 |
| 9426754 | 967903 | 967903 |
| 487468 | 7974346 | 487468 |
| 536845 | 3465 | 3465 |
| 343457 | 786978 | 343457 |

$$P(h_{min}(A) = h_{min}(B)) = J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

I stand at the seashore, alone,
and start to think.
There are the rushing waves
mountains of molecules
each stupidly minding its own
business
trillions apart
yet forming white surf in unison.

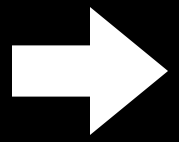


...

| $h_1$ | $h_n$ | $h_{min}$ |
|---|---|---|
| 3453409 | 475834 | 475834 |
| 9426754 | 967903 | 967903 |
| 487468 | 7974346 | 487468 |
| 536845 | 3465 | 3465 |
| 343457 | 786978 | 343457 |

$$P(h_{min}(A) = h_{min}(B)) = J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

I stand at the seashore, alone,
and start to think.
There are the rushing waves
mountains of molecules
each stupidly minding its own
business
trillions apart
yet forming white surf in unison.

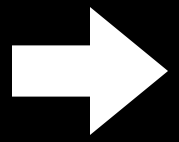


...

| $h_1$ | $h_n$ | $h_{min}$ |
|---|---|---|
| 3453409 | 475834 | 475834 |
| 9426754 | 967903 | 967903 |
| 487468 | 7974346 | 487468 |
| 536845 | 3465 | 3465 |
| 343457 | 786978 | 343457 |

$$P(h_{min}(A) = h_{min}(B)) = J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

I stand at the seashore, alone,
and start to think.
There are the rushing waves
mountains of molecules
each stupidly minding its own
business
trillions apart
yet forming white surf in unison.

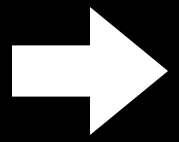


...

| $h_1$ | $h_n$ | $h_{min}$ |
|---|---|---|
| 3453409 | 475834 | 475834 |
| 9426754 | 967903 | 967903 |
| 487468 | 7974346 | 487468 |
| 536845 | 3465 | 3465 |
| 343457 | 786978 | 343457 |

$$P(h_{min}(A) = h_{min}(B)) = J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

I stand at the seashore, alone,
and start to think.
There are the rushing waves
mountains of molecules
each stupidly minding its own
business
trillions apart
yet forming white surf in unison.

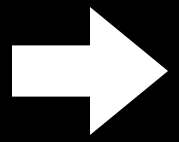


...

| $h_1$ | $h_n$ | $h_{min}$ |
|---|---|---|
| 3453409 | 475834 | 475834 |
| 9426754 | 967903 | 967903 |
| 487468 | 7974346 | 487468 |
| 536845 | 3465 | 3465 |
| 343457 | 786978 | 343457 |

$$P(h_{min}(A) = h_{min}(B)) = J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

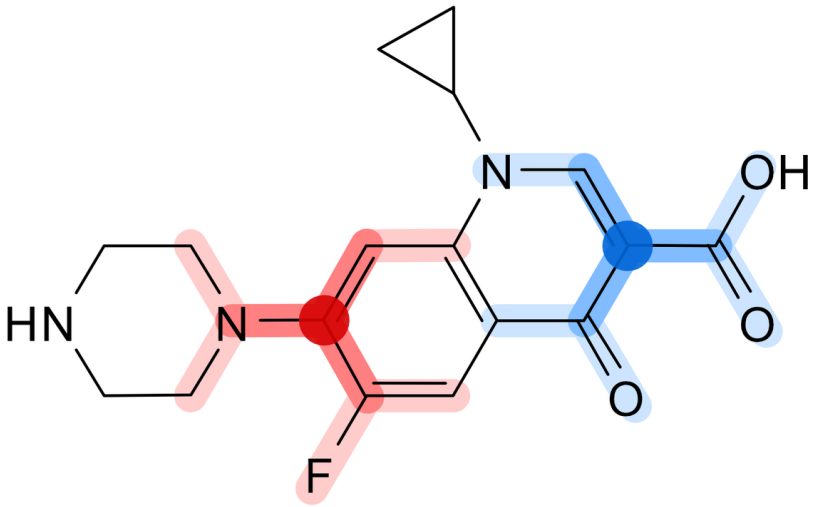
I stand at the seashore, alone,
and start to think.
There are the rushing waves
mountains of molecules
each stupidly minding its own
business
trillions apart
yet forming white surf in unison.

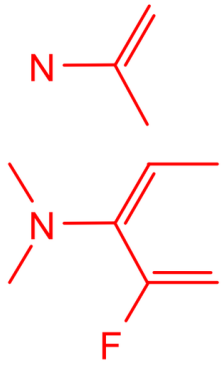


...

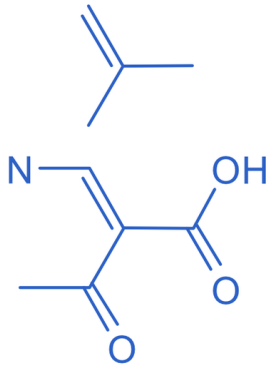
| $h_1$ | $h_n$ | $h_{min}$ |
|---|---|---|
| 3453409 | 475834 | 475834 |
| 9426754 | 967903 | 967903 |
| 487468 | 7974346 | 487468 |
| 536845 | 3465 | 3465 |
| 343457 | 786978 | 343457 |

$$P(h_{min}(A) = h_{min}(B)) = J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



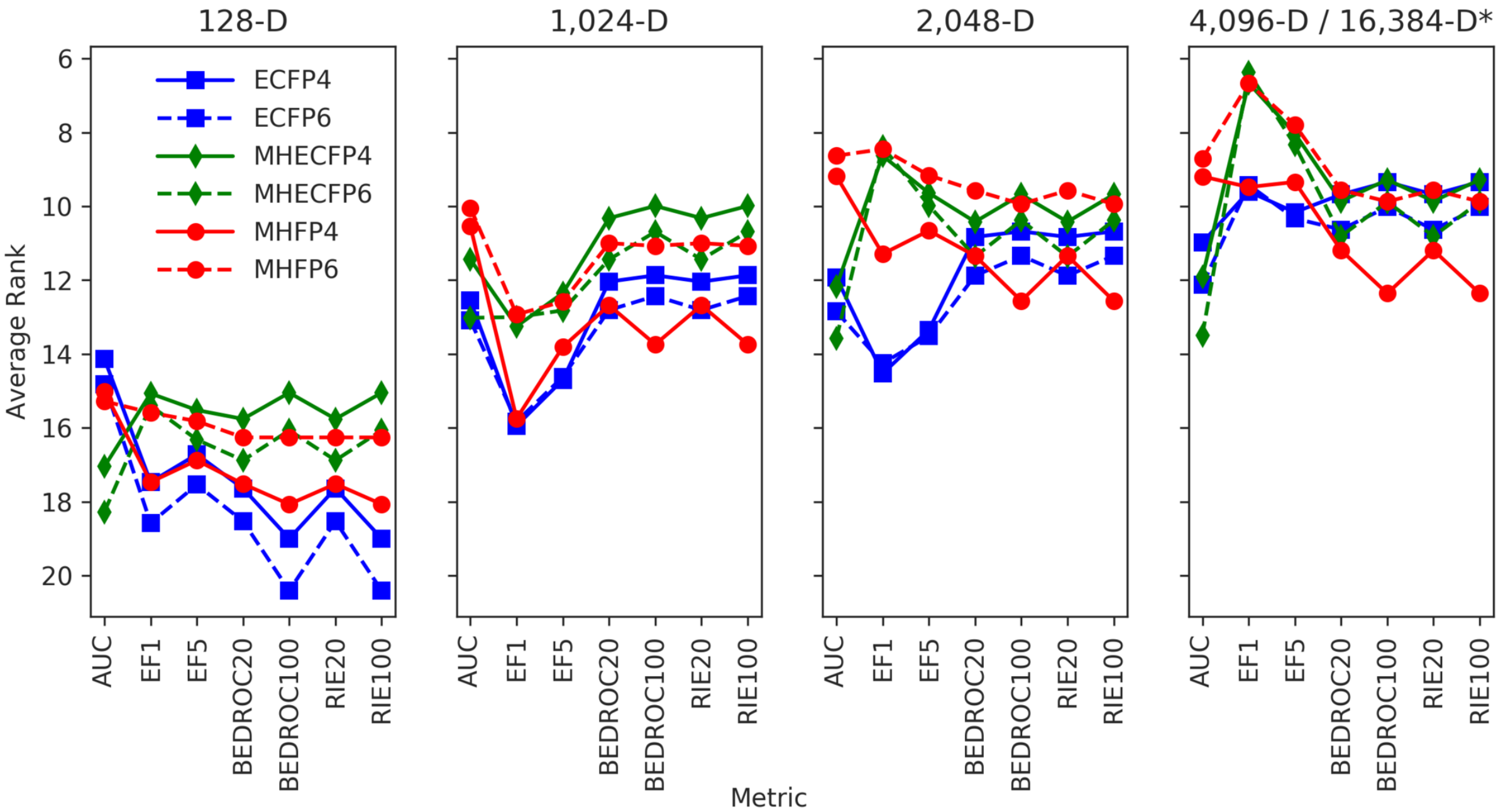



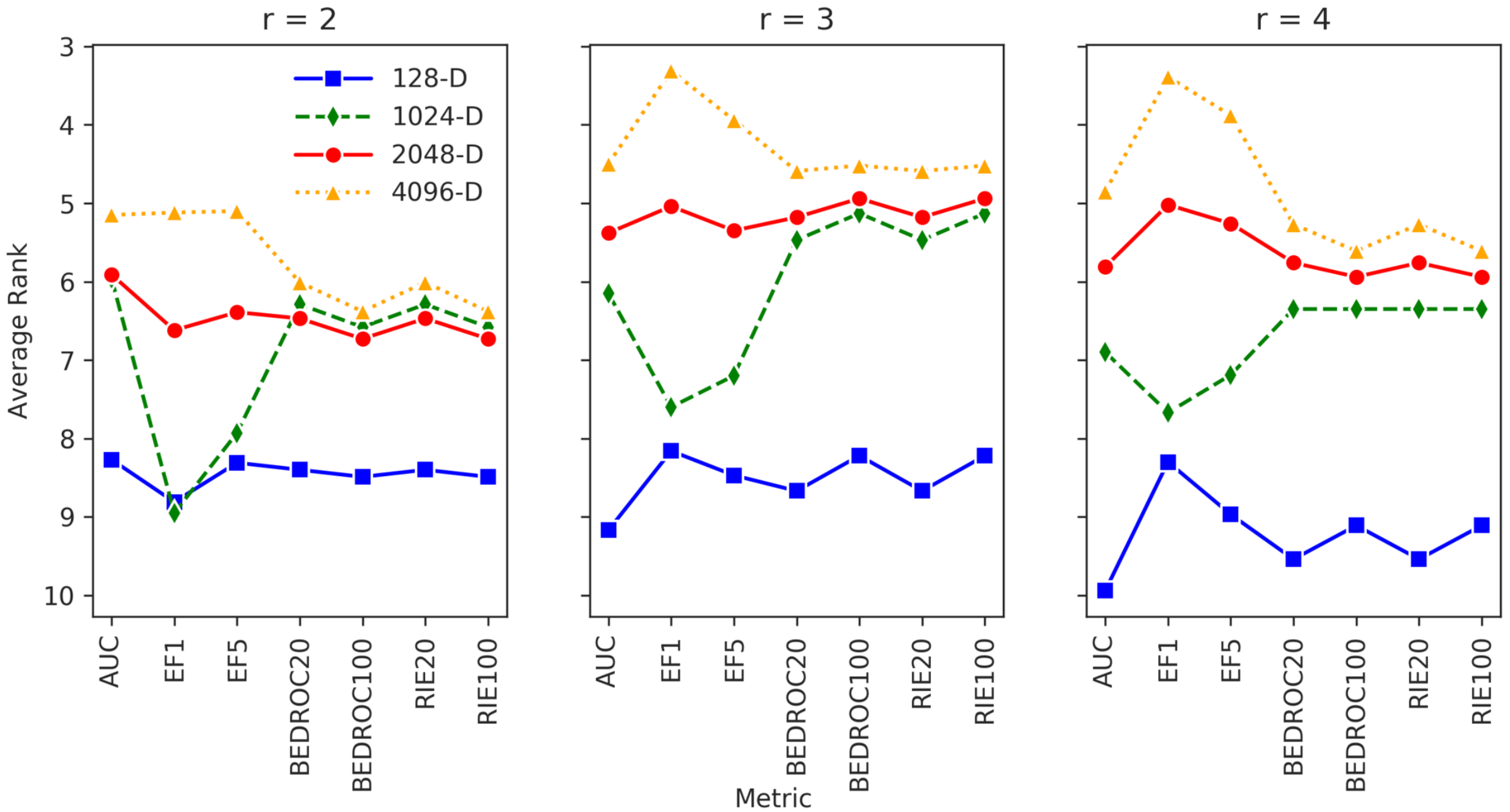

C
C(=C)(C)N
C(=CC)(C(=C)F)N(C)C



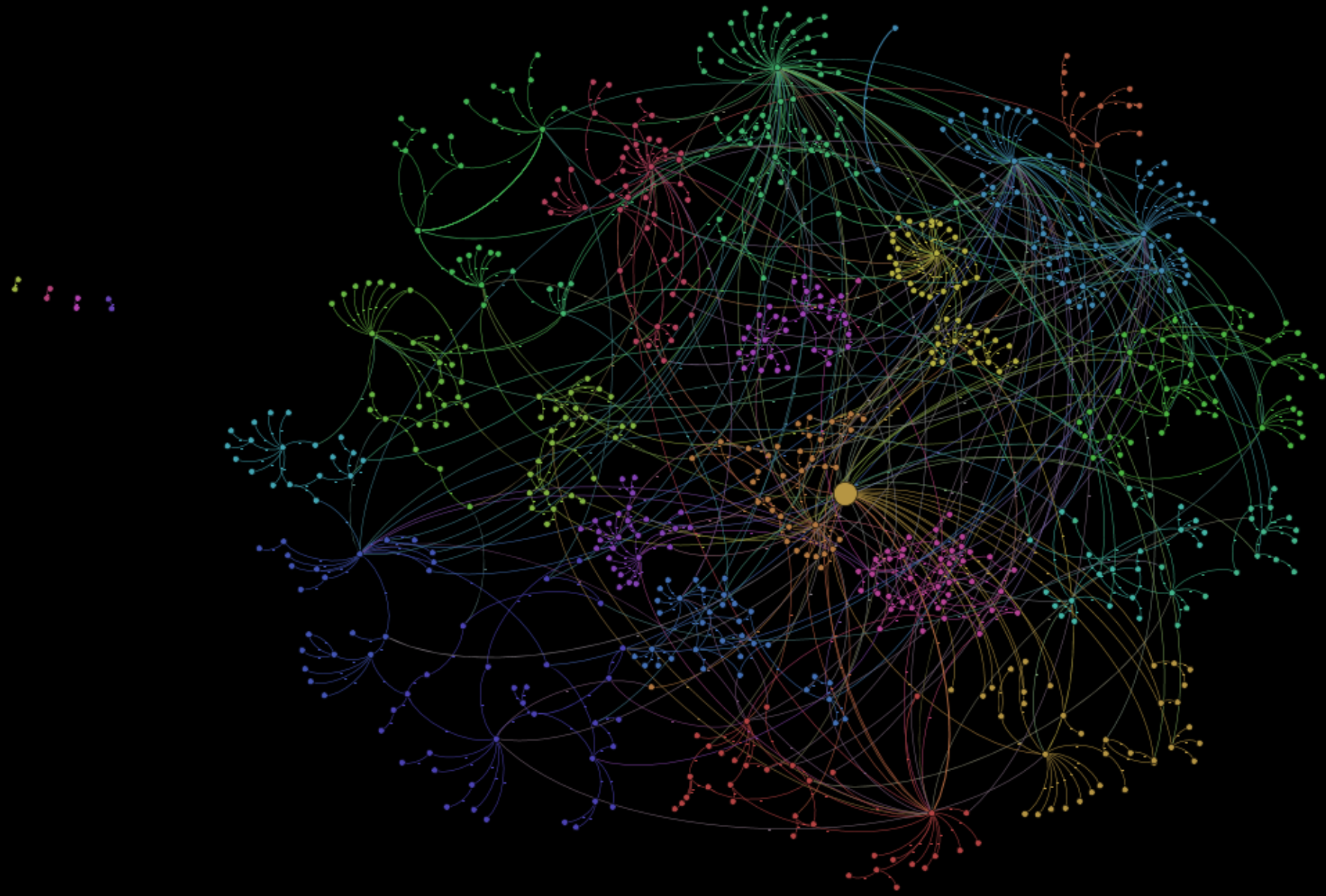

C
C(=C)(C)C
C(=CN)(C(=O)O)C(=O)C

| Method | AUC | EF 1% | EF 5% | BEDROC20 | EDROC100 | RIE20 | RIE100 |
|---|---|---|---|---|---|---|---|
| Shingling (r=2) | 0.773 | 31.590 | 9.718 | 0.496 | 0.492 | 8.520 | 25.368 |
| Shingling (r=3) | **0.776** | 32.698 | **9.897** | **0.506** | **0.505** | **8.691** | **26.118** |
| ECFP (r=2) | 0.755 | **32.702** | 9.792 | 0.500 | 0.503 | 8.590 | 26.027 |
| ECFP (r=3) | 0.754 | 32.563 | 9.780 | 0.499 | 0.502 | 8.571 | 25.956 |

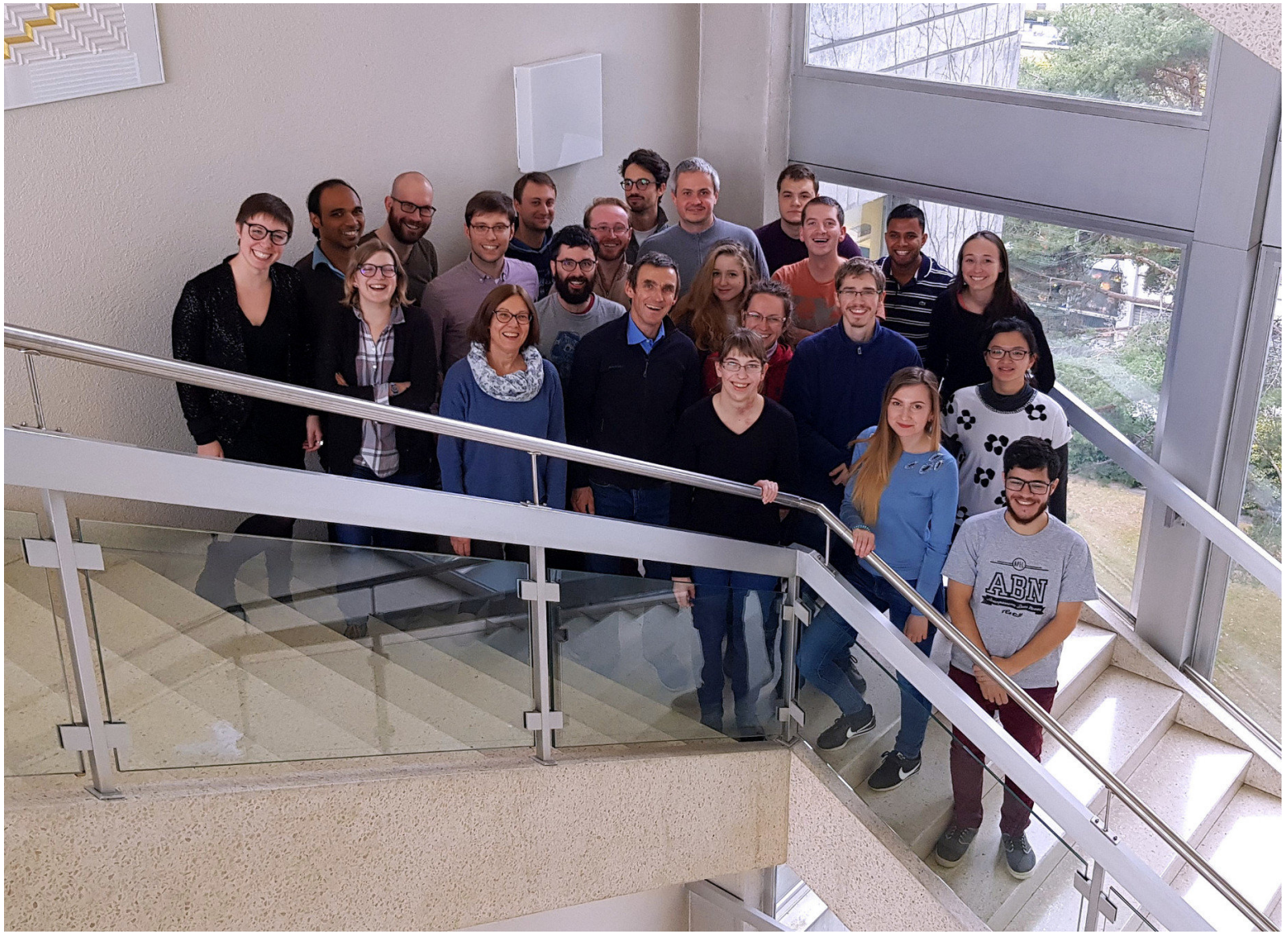

128-D
1,024-D
2,048-D
4,096-D / 16,384-D*
ECFP4
ECFP6
MHECFP4
MHECFP6
MHFP4
MHFP6
Average Rank
6
8
10
12
14
16
18
20
AUC
EF1
EF5
BEDROC20
BEDROC100
RIE20
RIE100
Metric


r = 2
r = 3
r = 4
128-D
1024-D
2048-D
4096-D
Average Rank
3
4
5
6
7
8
9
10
AUC
EF1
EF5
BEDROC20
BEDROC100
RIE20
RIE100
Metric

# Outlook



http://github.com/reymond-group




http://gdb.tools

http://viz.gdb.tools