

# On modeling, selecting and using 'drug-like' chemical matter: Toward optimized fragment collections

---

**Jörg Degen**, Christof Wegscheid-Gerlach\*, Andrea Zaliani#, Matthias Rarey



**University of Hamburg**



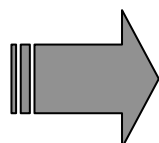
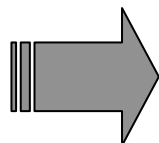
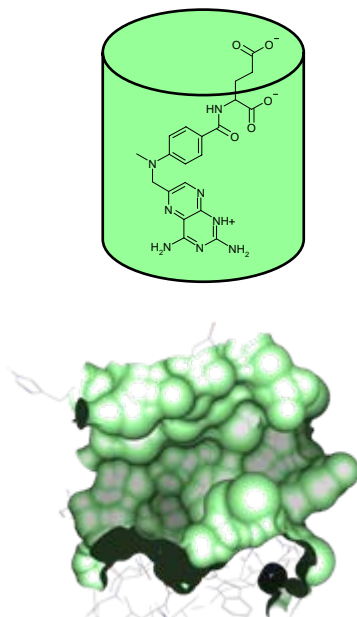
**Center for Bioinformatics**

Bundesstrasse 43

20146 Hamburg, Germany

[www.zbh.uni-hamburg.de](http://www.zbh.uni-hamburg.de)

# COMPUTER-AIDED MOLECULAR DESIGN



*Nature*, **432**, 823-865 (2004)

- ▶ The chemical space / universe is 'vastly, hugely, mind-bogglingly big'
- ▶ Too large for enumeration / screening
- ▶ Describe and explore the combinatorial nature of the problem



# AUTOMATIC FRAGMENTATION

---

## ▶ RECAP

- “Retrosynthetic Combinatorial Analysis Procedure”
- Chemical motifs easily accessible to combinatorial synthesis
- Retain ‘drug-like’ motifs/substituents

QuickTime™ and a decompressor are needed to see this picture.

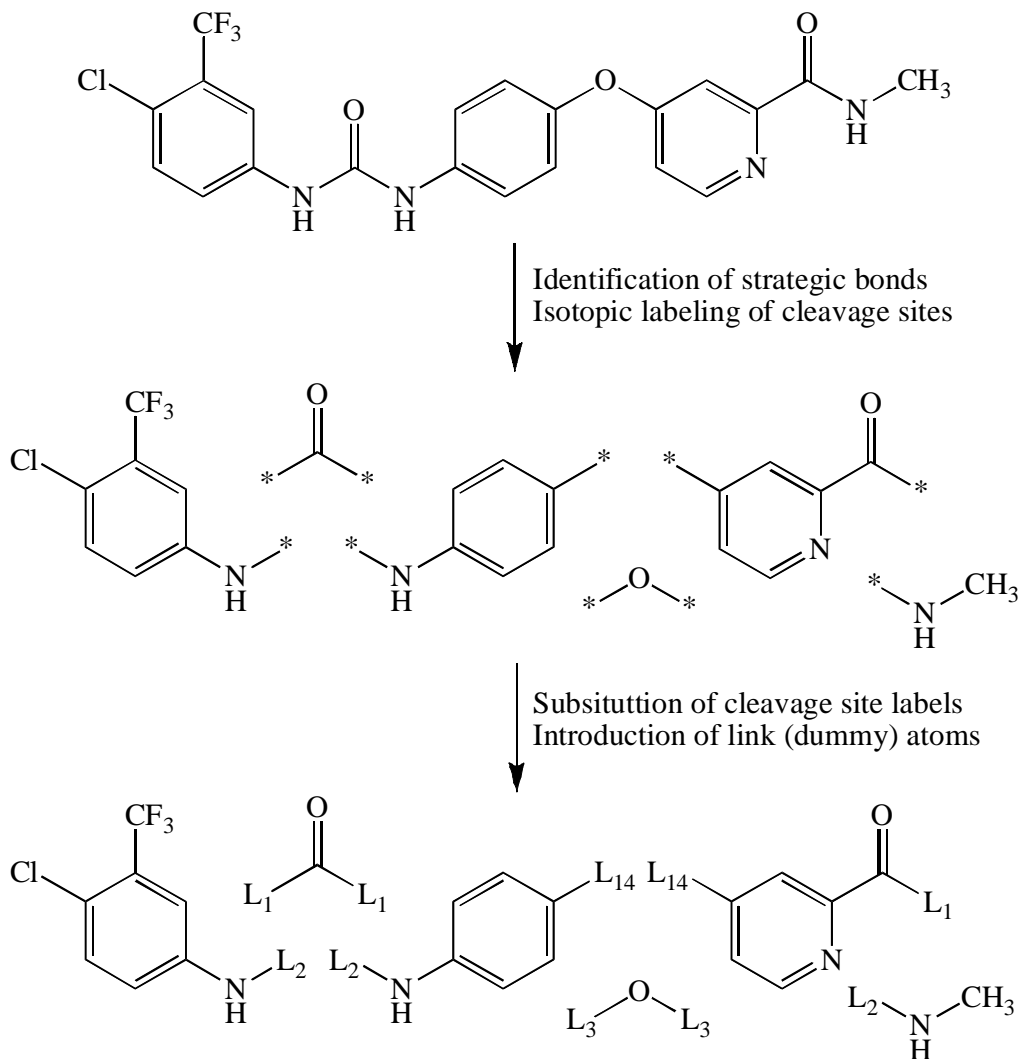
Lewell, X.Q. *et al.* (1998). *J Chem Inf Comput Sci*, 38, 511-522.



# FRAGMENT GENERATION - SHREDDING

## ▶ Nexavar (Sorafenib)

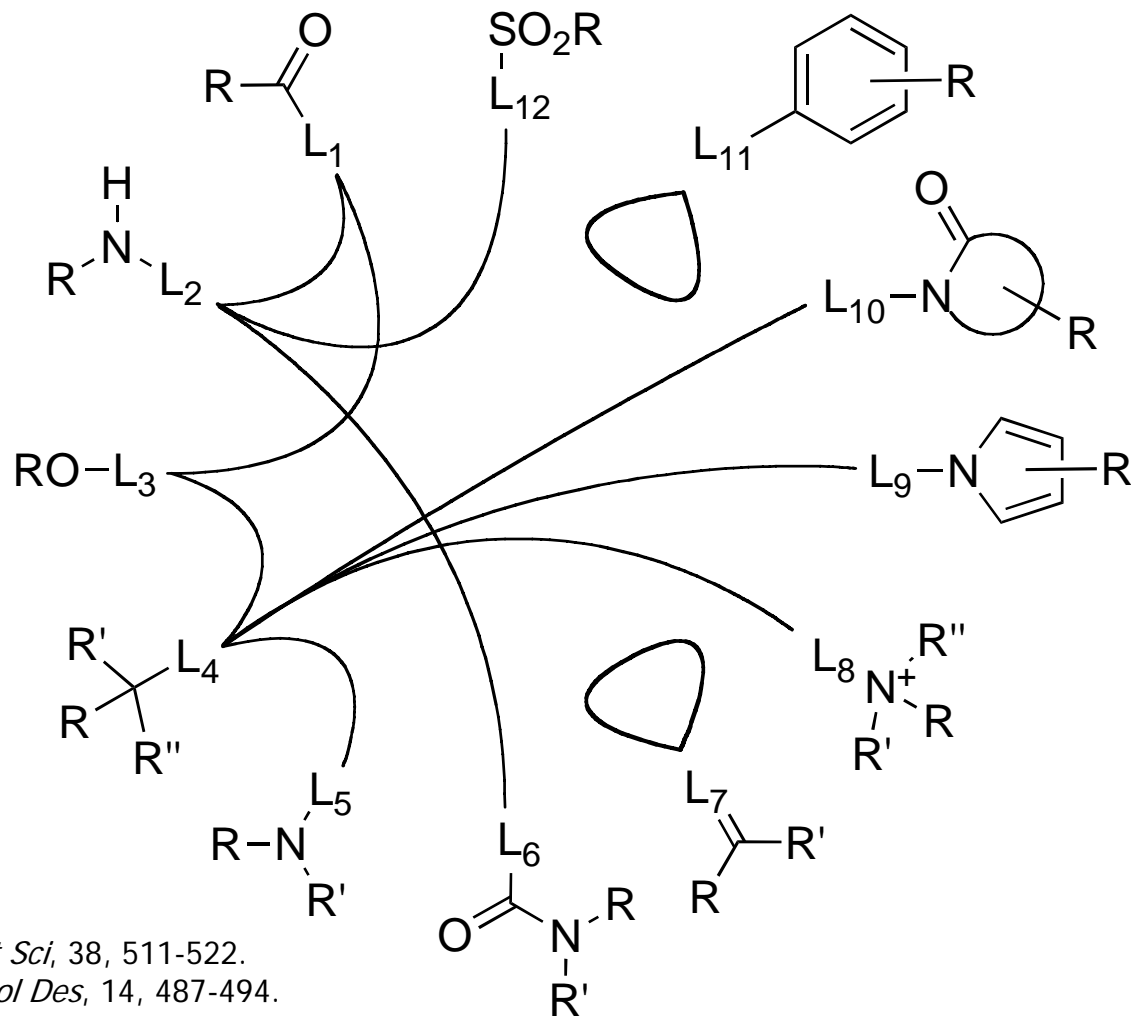
- Simultaneous shredding
- Avoid the generation of overlapping fragments
- Modeling of chemical motifs via link atoms



# FROM FRAGMENTS TO FRAGMENT SPACES

## ► Fragment space

- Based on the original RECAP set of rules
- Unification of chemical environments
- Many 'in-house' adaptations of RECAP



[1] Lewell, X.Q. *et al.* (1998)., *J Chem Inf Comput Sci*, 38, 511-522.

[2] Schneider, G. *et al.* (2000). *J Comput-Aided Mol Des*, 14, 487-494.

[3] Rarey, M. & Stahl, M. (2001). *J Comput-Aided Mol Des*, 15, 497-520.



# WHY GENERATE A NEW SPACE?

---

## ► Motivation / Goal

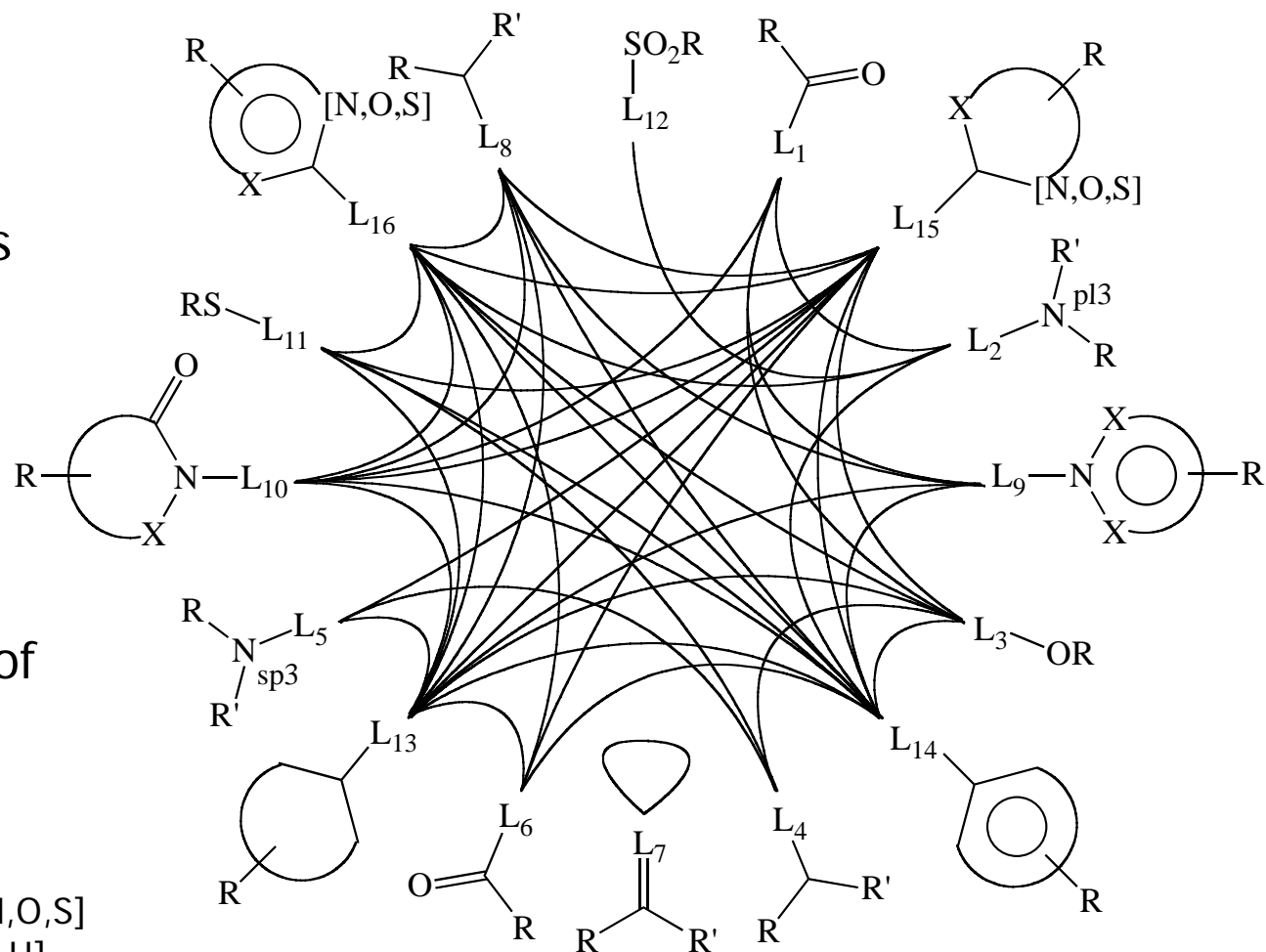
- Optimize the spaces for conformation-dependent applications  
→ Treat geometries of link environments consistently
- Model the chemical motifs that *result* from fragment connection  
→ Synthetically relevant functionalization of motifs is *not* the focus
- Use more elaborate shredding and fragment connection rules  
→ Introduce additional link types and formulate existing ones more specific
- Generate high-quality and high-performance fragment spaces  
→ Publicly available fragment sets, retain 'drug-like' chemical motifs



# THE BRICS FRAGMENT SPACE

## ► BRICS

- Models some explicit isosteres
- Treats link geometries consistently
- Retains history of some chemical environments



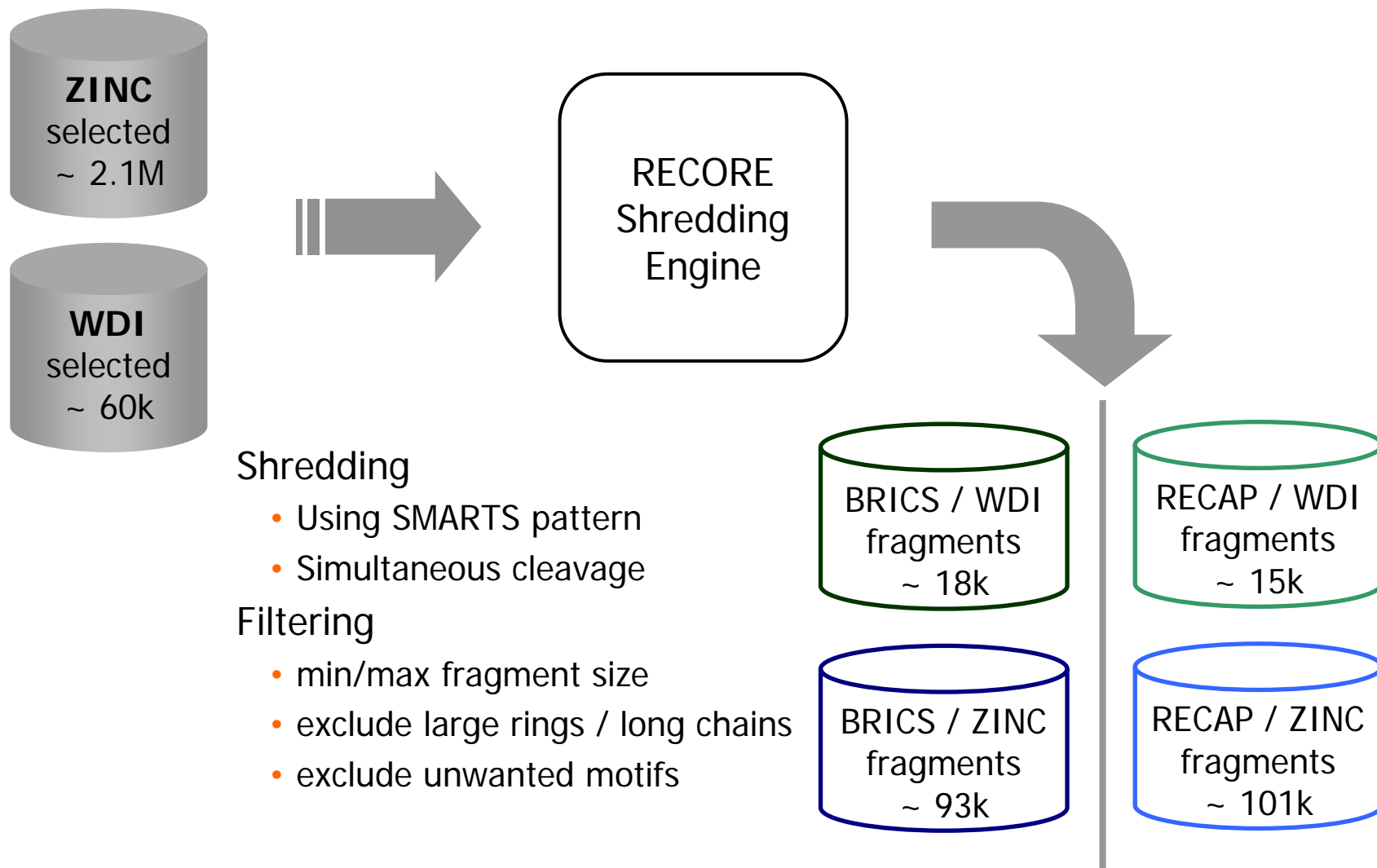
X ~ [C,N,O,S]

R' ~ [R'', H]

R may contain link(s)



# FRAGMENT GENERATION



# SHREDDING RESULTS

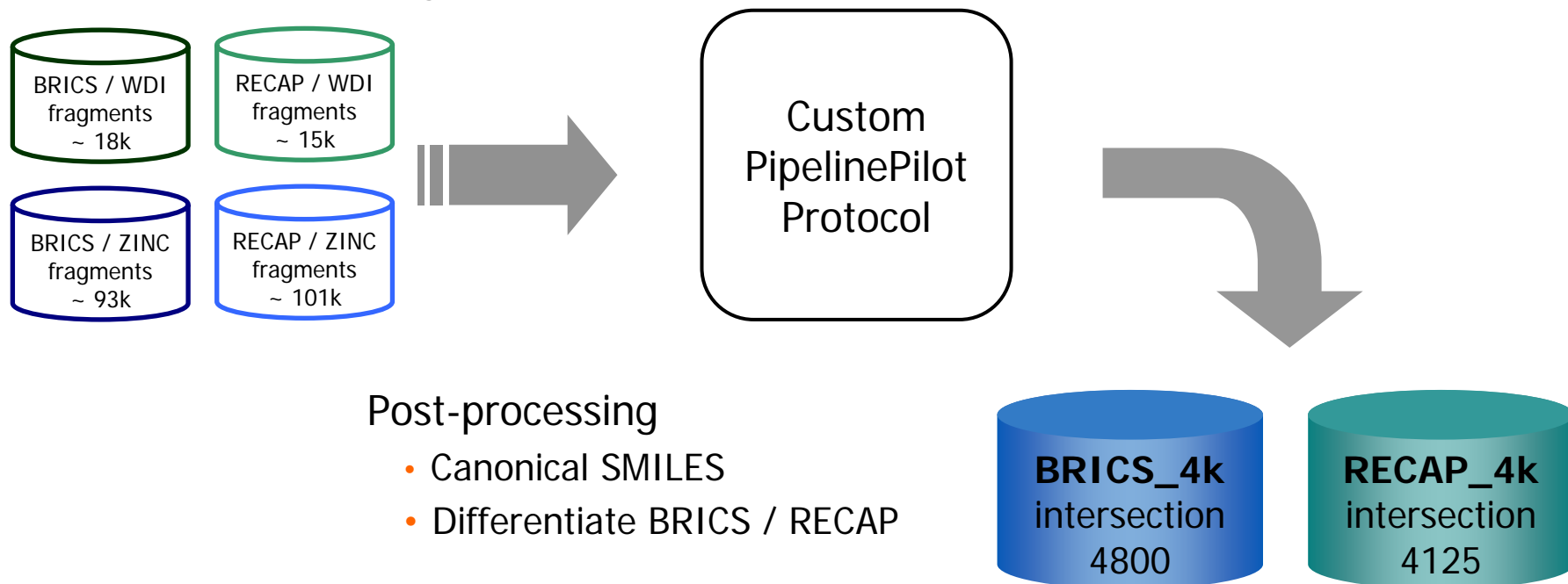
## ► Fragmentation statistics

	WDI		ZINC	
	BRICS	RECAP	BRICS	RECAP
Uncleaved compounds	20715 (35%)	28362 (47%)	214793 (11%)	415762 (21%)
Cleaved compounds	39163 (65%)	31516 (53%)	1806262 (89%)	1605293 (79%)
Unique fragments	18291 (100%)	15650 (100%)	93309 (100%)	101889 (100%)
1-connection fragments	13256 (72%)	12236 (78%)	76196 (81%)	90023 (88%)
2-connection fragments	4344 (24%)	3079 (19%)	15964 (17%)	11395 (11%)
3-connection fragments	591 (03%)	290 (02%)	1118 (01%)	453 (<1%)
4-connection fragments	83 (<1%)	36 (<1%)	30 (<1%)	17(<1%)
5-connection fragments	12 (<1%)	7 (<1%)	1 (<1%)	1(<1%)
6-connection fragments	5 (<1%)	1 (<1%)	-	-
7-connection fragments	-	1 (<1%)	-	-



# COMPILATION OF THE NEW FRAGMENT SETS

BRICS / RECAP fragments from WDI / ZINC



Compile 'intersections'

- Determine fragment identity
- Generate overlapping sets



# PERFORMANCE EVALUATION

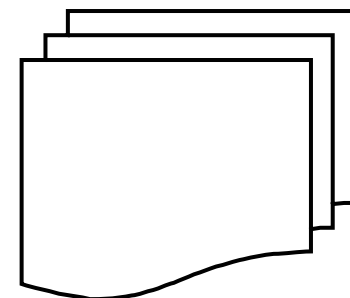
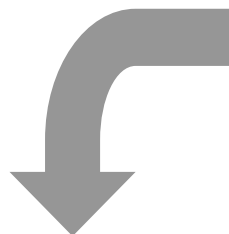
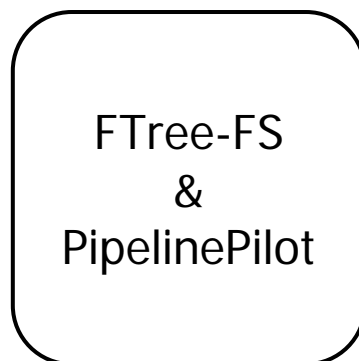
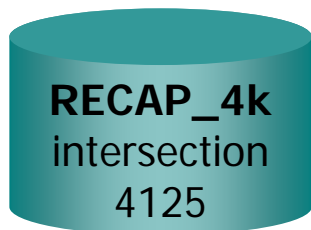
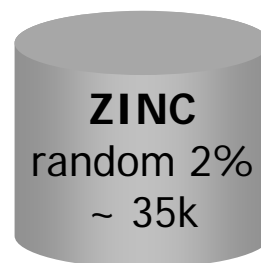
## Multiple Ftree-FS searches

- For RECAP/BRICS queries/spaces
- 25 globally most similar results

## Re-ranking with Fingerprint

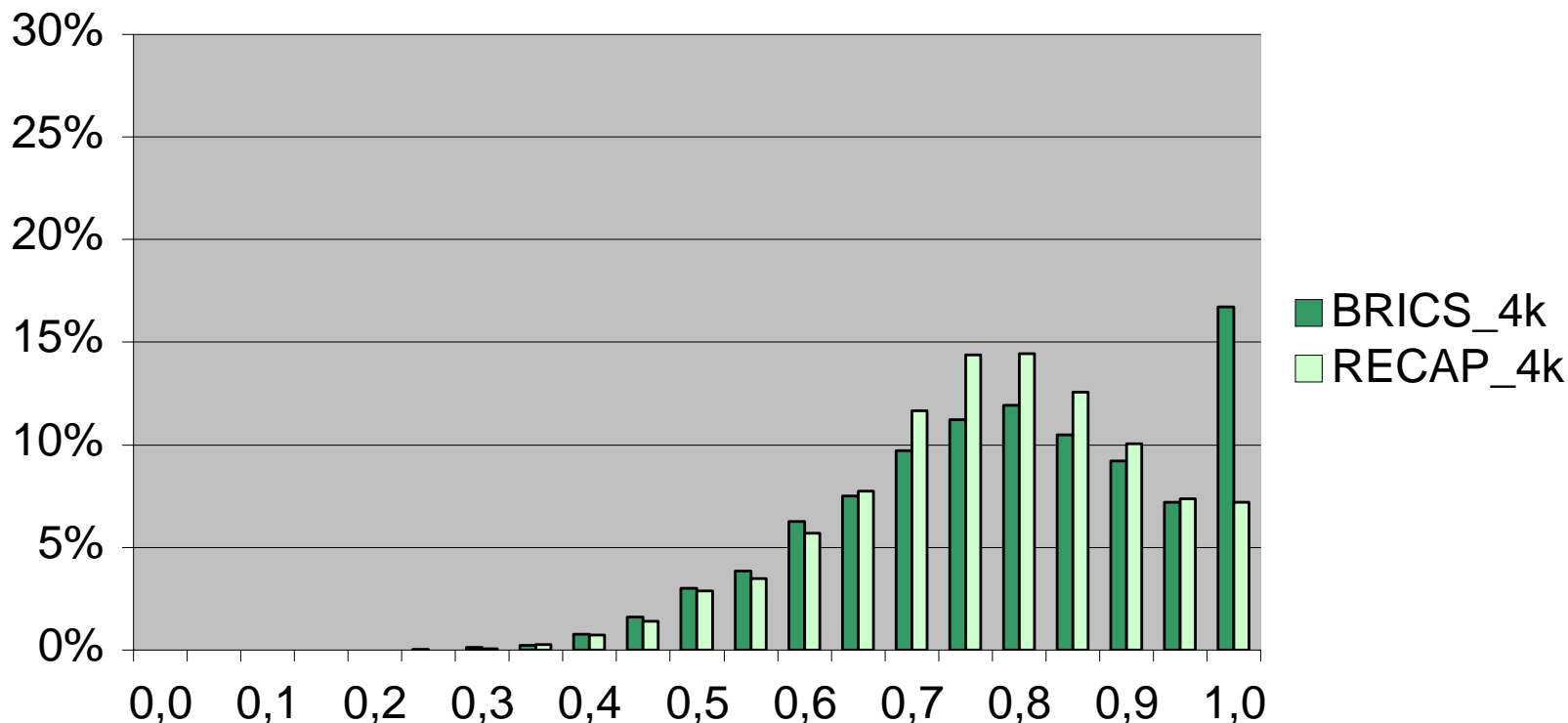
- MDL public keys for results/queries
- Pair-wise Tanimoto distances
- Keep only the best solution

## Query molecules



# COMPARISON OF RECAP AND BRICS

## ► Performance of the '4k intersections' for the WDI queries

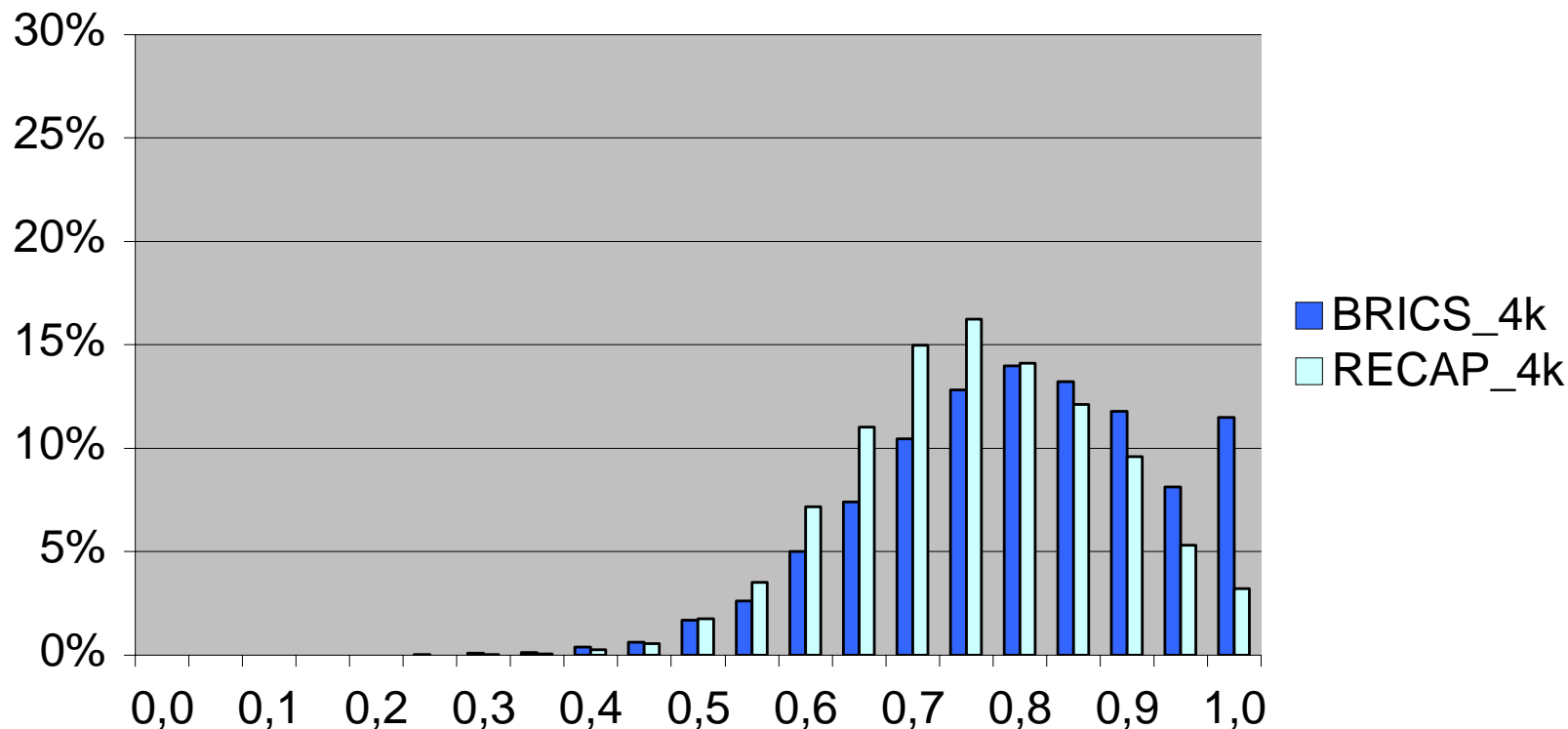


Similarity of reranked solutions using MDL public keys



# COMPARISON OF RECAP AND BRICS

## ► Performance of the '4k intersections' for the ZINC queries



Similarity of reranked solutions using MDL public keys



# CAN WE STILL DO BETTER?

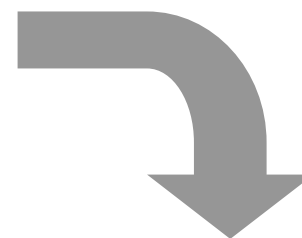
BRICS fragments WDI / ZINC

BRICS / WDI  
fragments  
~ 18k

BRICS / ZINC  
fragments  
~ 93k



Custom  
PipelinePilot  
Protocol



Post-processing

- Calculate MDL public keys
- Use binned MW ranges (30 Da)

Compile enriched intersections

- Determine fragment similarity
- Tanimoto distance (MDL keys)
- Generate set with  $\text{sim} > 0.9$
- Generate set with  $\text{sim} > 0.8$

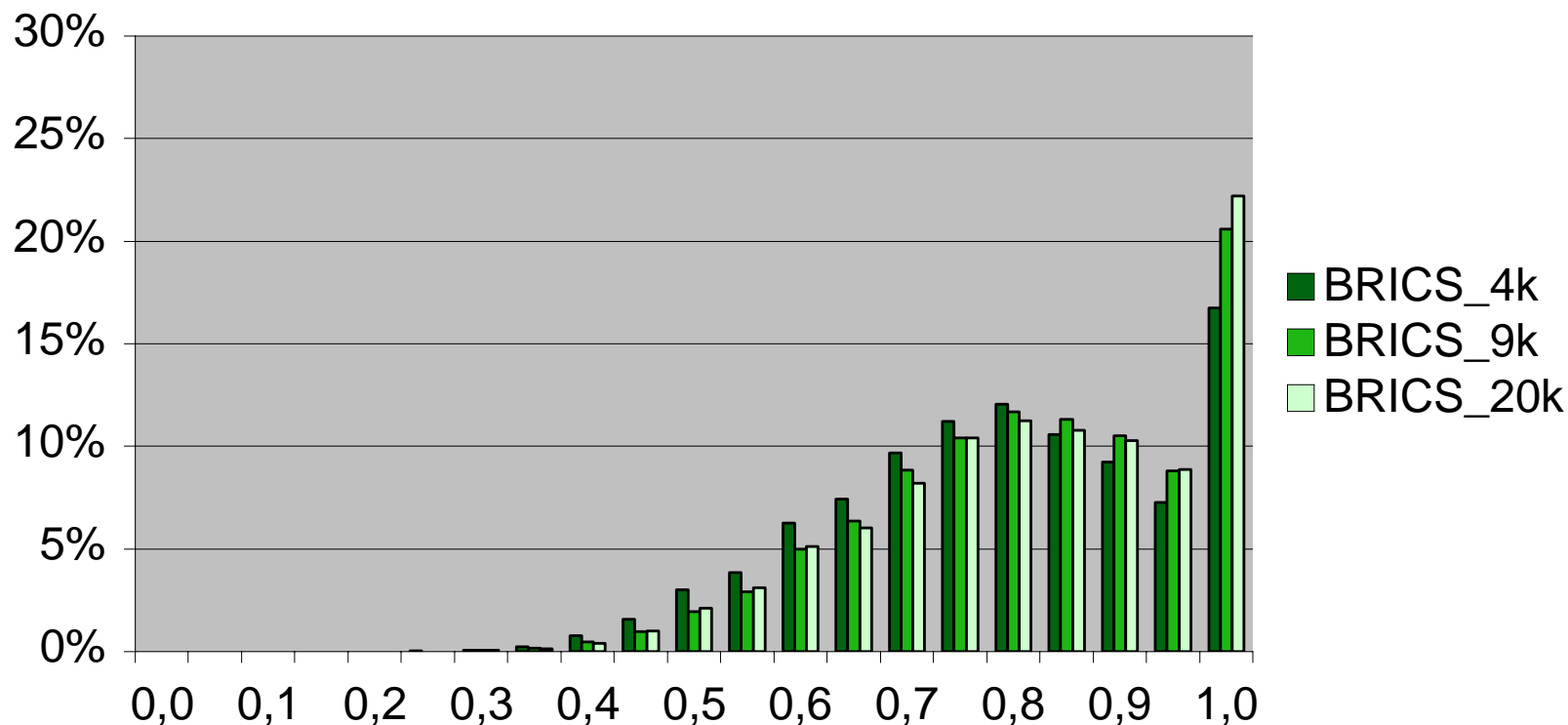
**BRICS\_9k**  
intersection 0.9  
9344

**BRICS\_20k**  
intersection 0.8  
22343



# ASSESSMENT OF THE PERFORMANCE

## ► BRICS fragment spaces for WDI queries

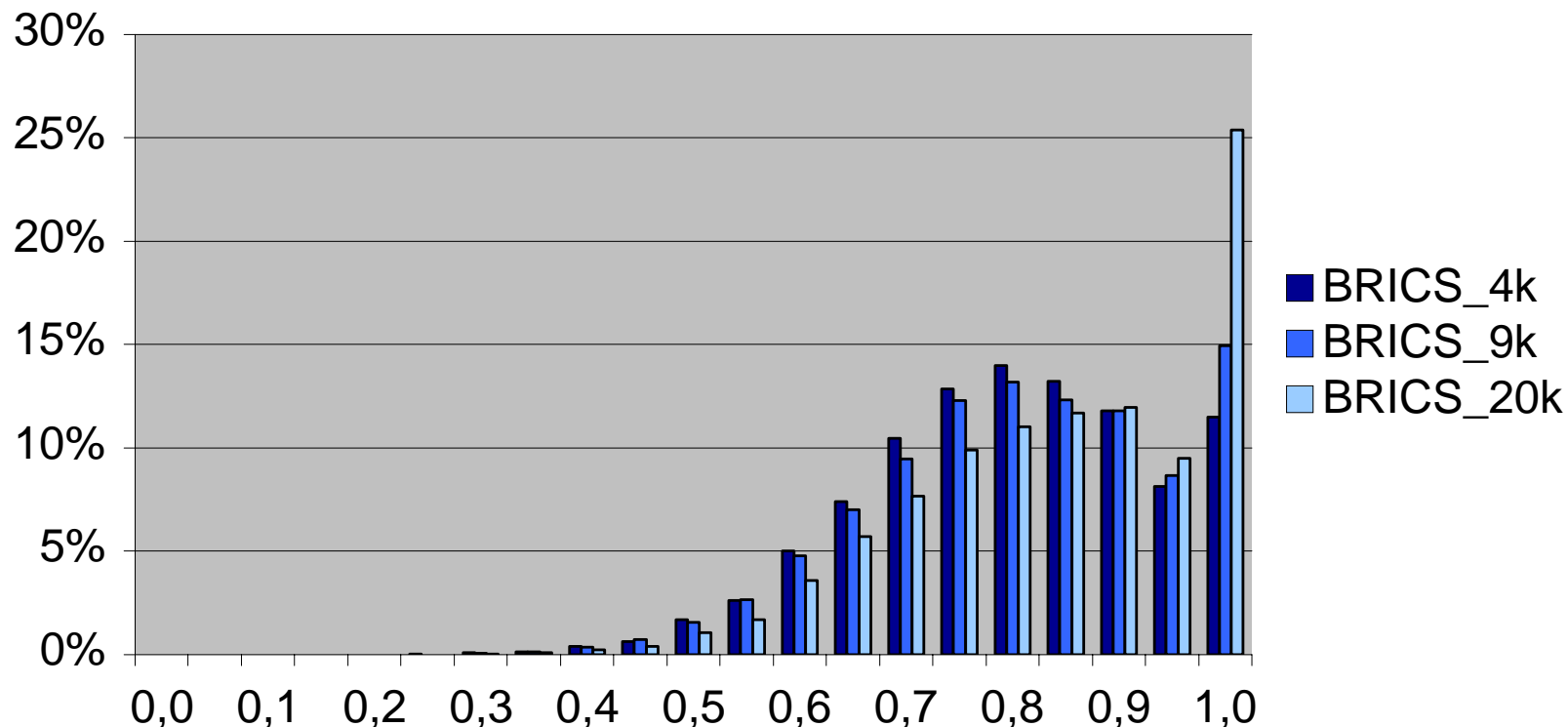


Similarity of reranked solutions using MDL public keys



# ASSESSMENT OF THE PERFORMANCE

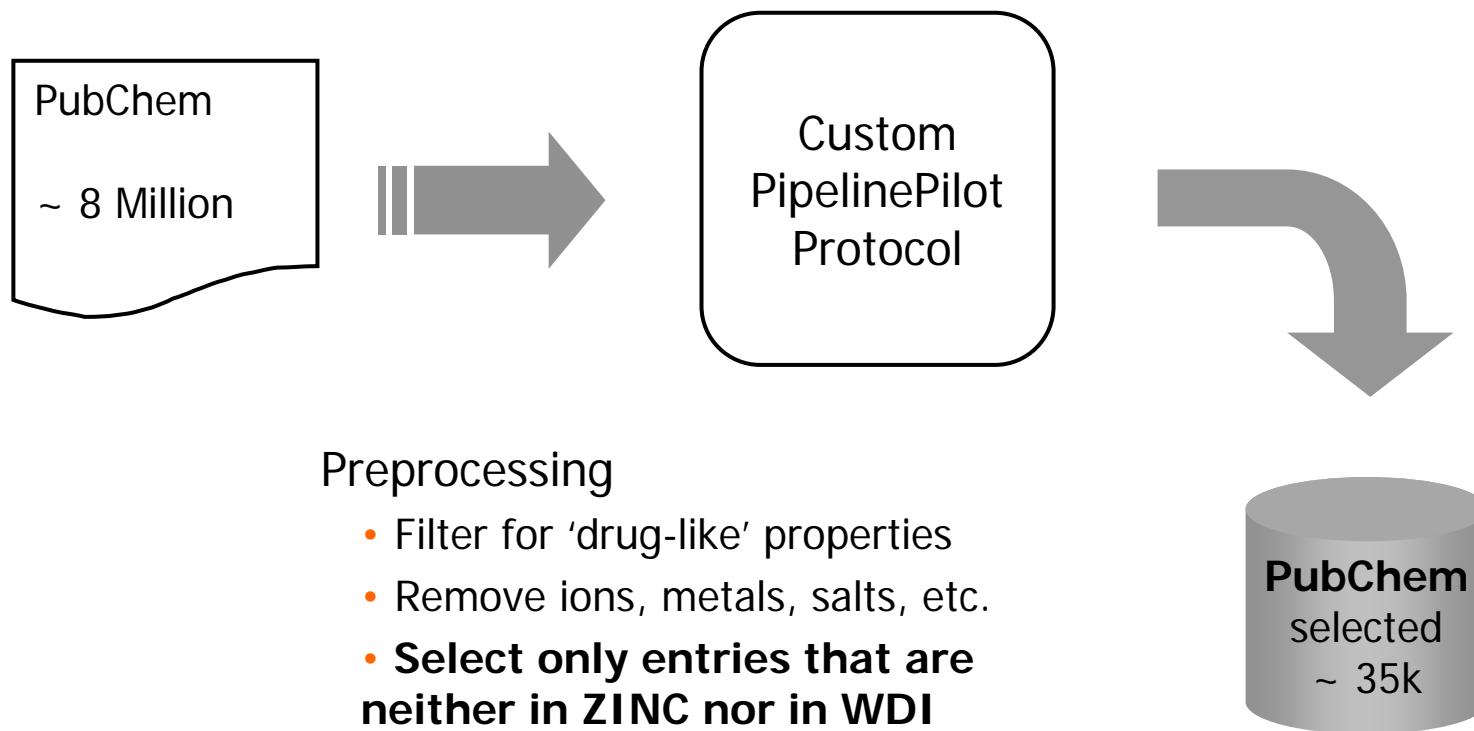
## ► BRICS fragment spaces for ZINC queries



Similarity of reranked solutions using MDL public keys



# COMPILING A 'REAL' TEST

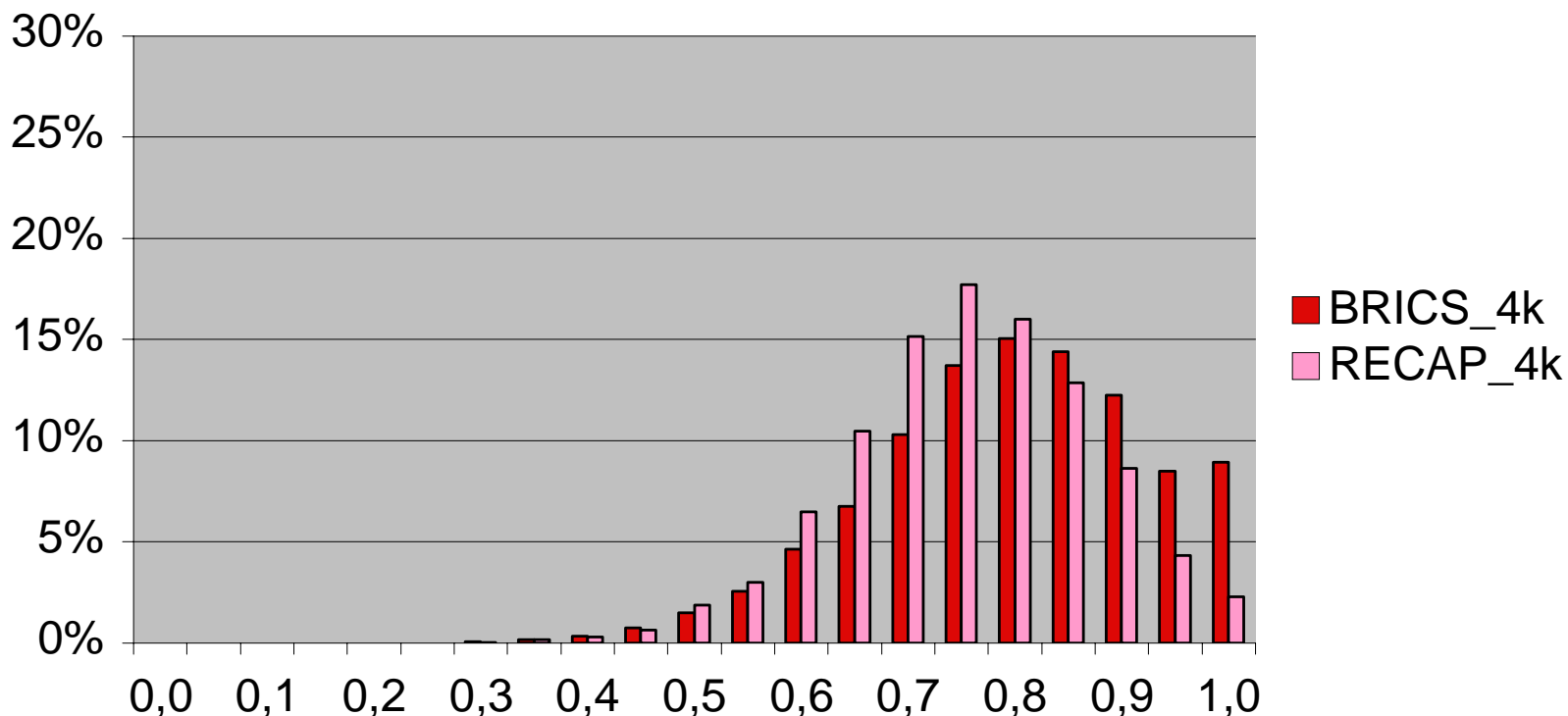


## Preprocessing

- Filter for 'drug-like' properties
- Remove ions, metals, salts, etc.
- **Select only entries that are neither in ZINC nor in WDI**
- k-means clustering (10000 clust.)
- Select random 0.5% per cluster

# COMPARISON OF RECAP AND BRICS

## ► Performance for the PubChem queries

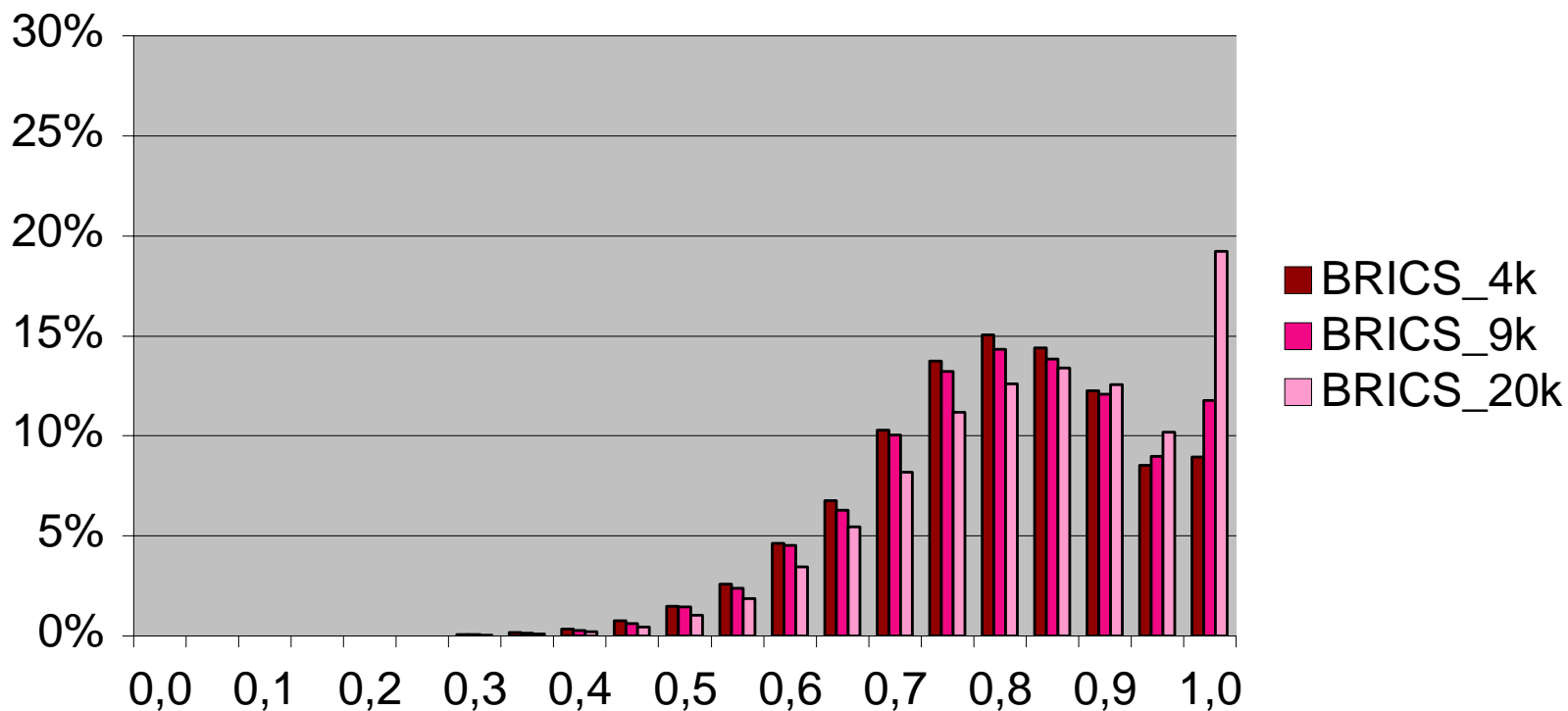


Similarity of reranked solutions using MDL public keys



# ASSESSMENT OF THE PERFORMANCE

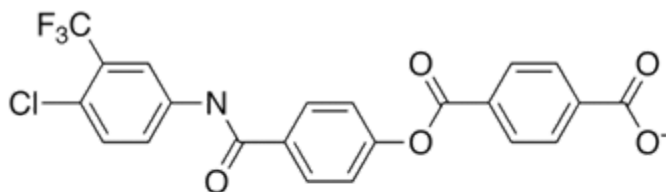
## ► BRICS fragment spaces for PubChem queries



Similarity of reranked solutions using MDL public keys

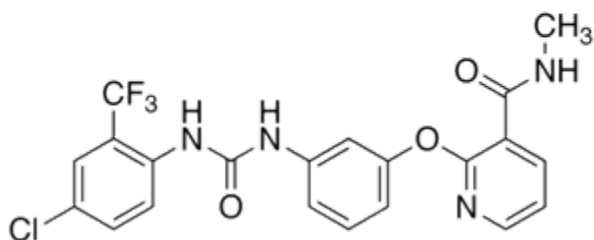


# EXAMPLE CASE STUDY



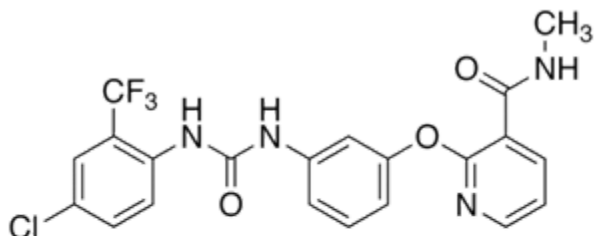
RECAP\_4k

0,945 Feature Tree  
0,704 MDL Fingerprint



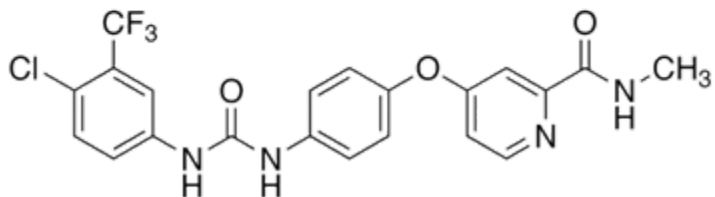
BRICS\_4k

0,999 Feature Tree  
0,926 MDL Fingerprint



BRICS\_9k

0,999 Feature Tree  
0,926 MDL Fingerprint



BRICS\_20k

0,999 Feature Tree  
1,000 MDL Fingerprint

# SUMMARY

---

- ▶ Improved approach for modeling fragment spaces
  - Three fragment spaces BRICS\_4k, BRICS\_9k, BRICS\_20k
  - ZINC-derived fragments with decreasing similarity to WDI-fragments
- ▶ Assess performance with three query test sets
  - ~35k molecules each from WDI, ZINC and PubChem
  - Compare BRICS and RECAP as well as the enriched BRICS sets
- ▶ Bottom line
  - Exactly rebuilt (FTree & MDL) 19% - 26% of all queries molecules
  - Close analogues (sim>0.9) for another >20% of all queries
  - Useful for modeling a large variety of different molecules



# ACKNOWLEDGEMENTS

---



**Matthias Rarey**

Patrick Maaß

Juri Pärn

Ingo Reulecke

[...]



Bayer HealthCare  
Bayer Schering Pharma

Christof Wegscheid-Gerlach



Andrea Zaliani



Christian Lemmen

NovoBench project

4SC AG

ALTANA Pharma AG

BioSolveIT GmbH

CCC Erlangen

Lilly Forschung GmbH

Molecular Networks GmbH

ZBH Hamburg

partially funded by the BMBF under grant 313324A.

<http://www.zbh.uni-hamburg.de/BRICS>

Degen, J., Wegscheid-Gerlach, C., Zaliani, A. & Rarey, M. (2008). *ChemMedChem*, in print.

