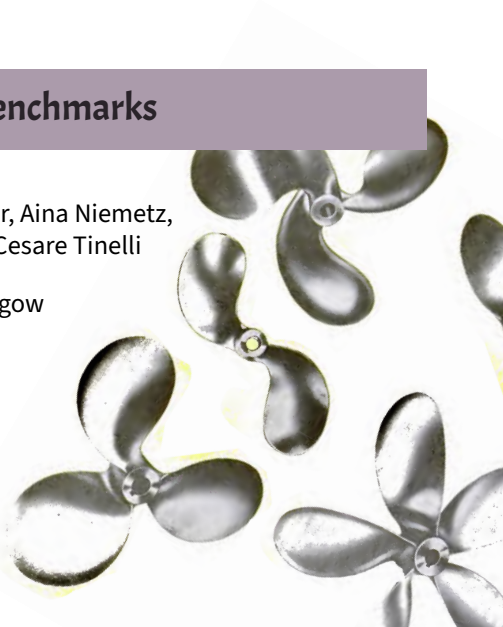
A Catalog of SMT-LIB Benchmarks

Hans-Jörg Schurr, Mathias Preiner, Aina Niemetz,
Clark Barrett, Pascal Fontaine, Cesare Tinelli

SMT Workshop — Glasgow
August 11, 2025



Are you an SMT-LIB user, and you ...

- are overwhelmed by the $34\,614\,311$ queries in the $495\,177$ SMT-LIB benchmarks?

Are you an SMT-LIB user, and you ...

- are overwhelmed by the $34\,614\,311$ queries in the $495\,177$ SMT-LIB benchmarks?
- want to know which benchmarks were solved by your nemesis, but not by you?

Are you an SMT-LIB user, and you ...

- are overwhelmed by the $34\,614\,311$ queries in the $495\,177$ SMT-LIB benchmarks?
- want to know which benchmarks were solved by your nemesis, but not by you?
  - in 2012?

Are you an SMT-LIB user, and you ...

- are overwhelmed by the $34\,614\,311$ queries in the $495\,177$ SMT-LIB benchmarks?
- want to know which benchmarks were solved by your nemesis, but not by you?
  - in 2012?
- want to evaluate your new trick to handle `str.suffixof`?

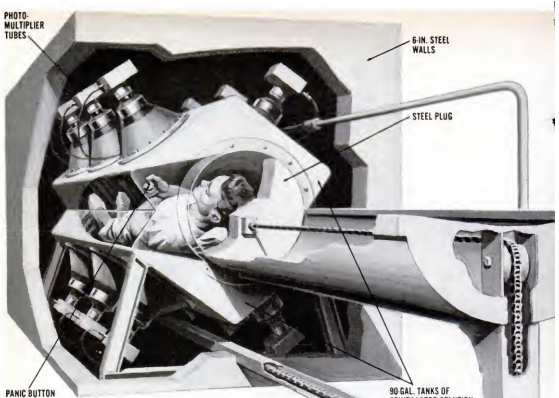
We have the solution for you!

One nifty **database** with all you need to answer these questions!

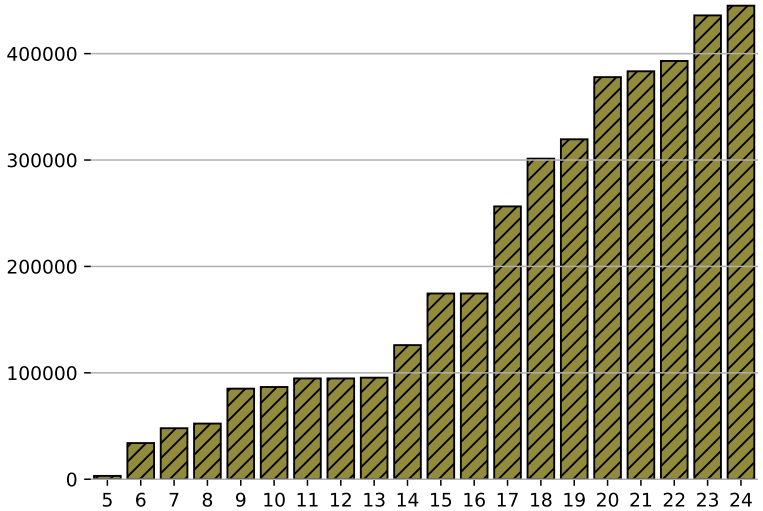
- Download a single SQLite file and
  - use your favorite language to perform queries,
  - or just use raw SQL.
- Around 5.5 GiB of of juicy data
- Benchmark and query metadata!
- Results from 20 years of SMT-COMP!

Part I

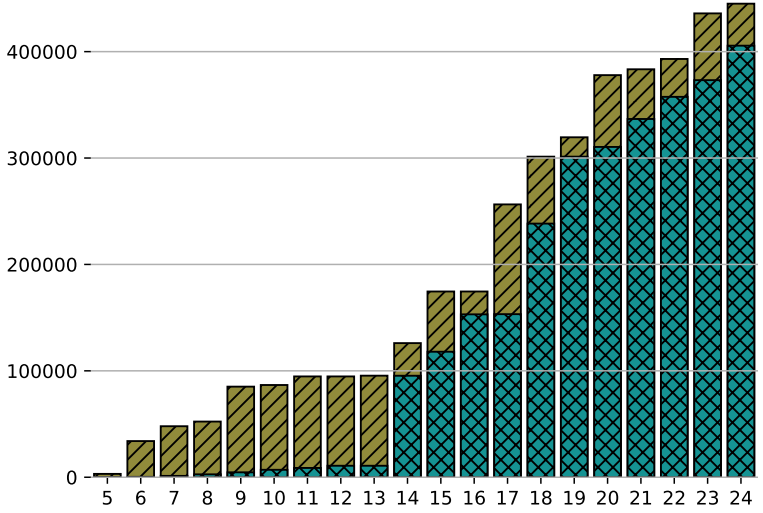
Case Study



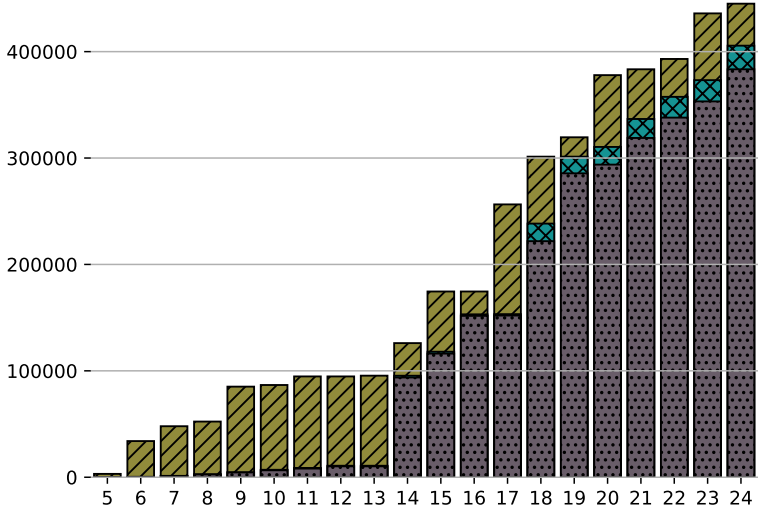
New Non-Incremental Benchmarks



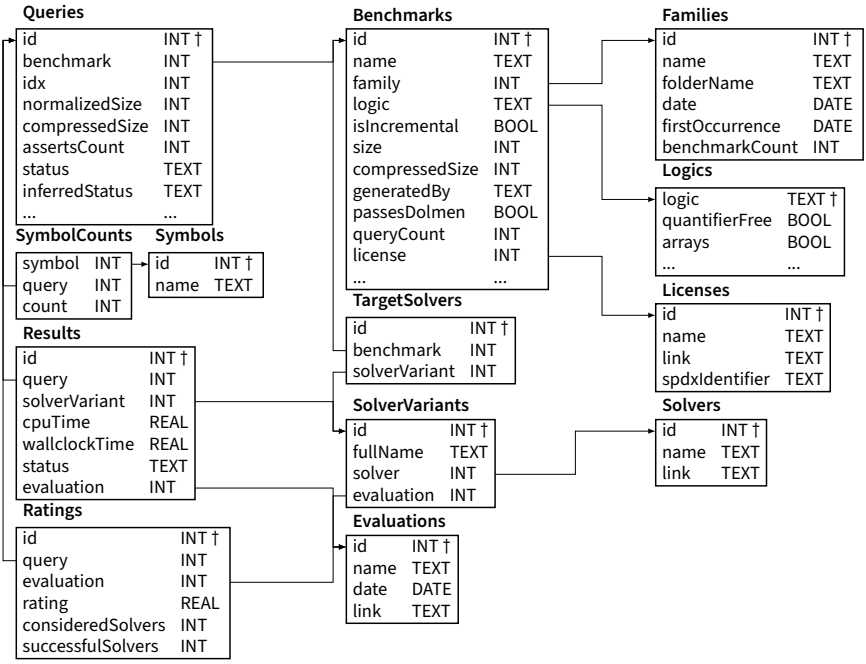
Used Benchmarks

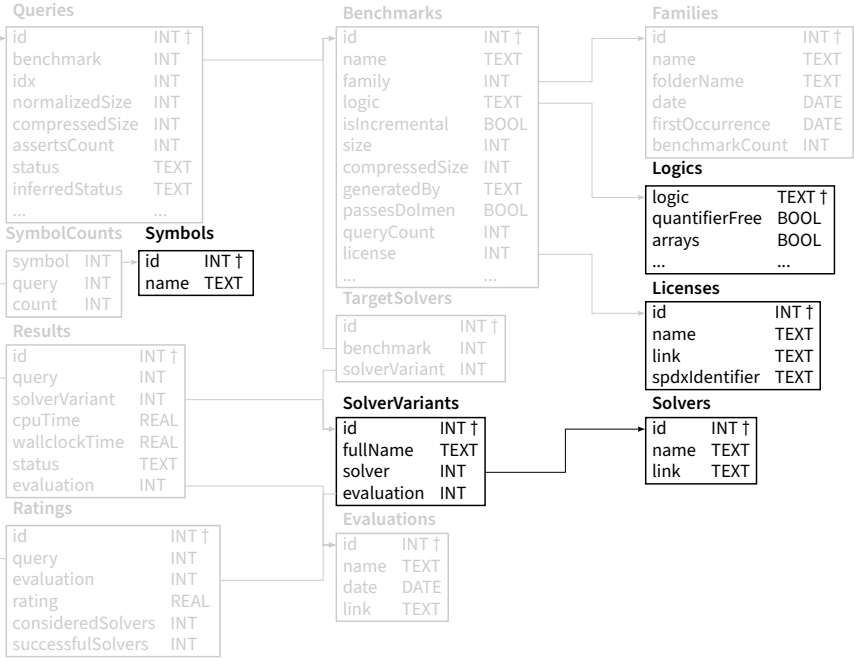


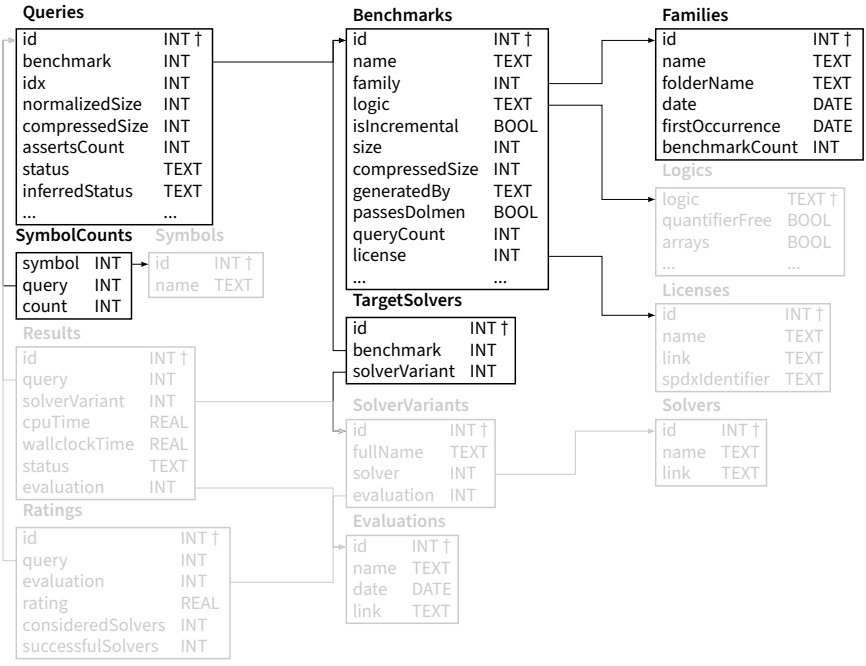
Solved Benchmarks

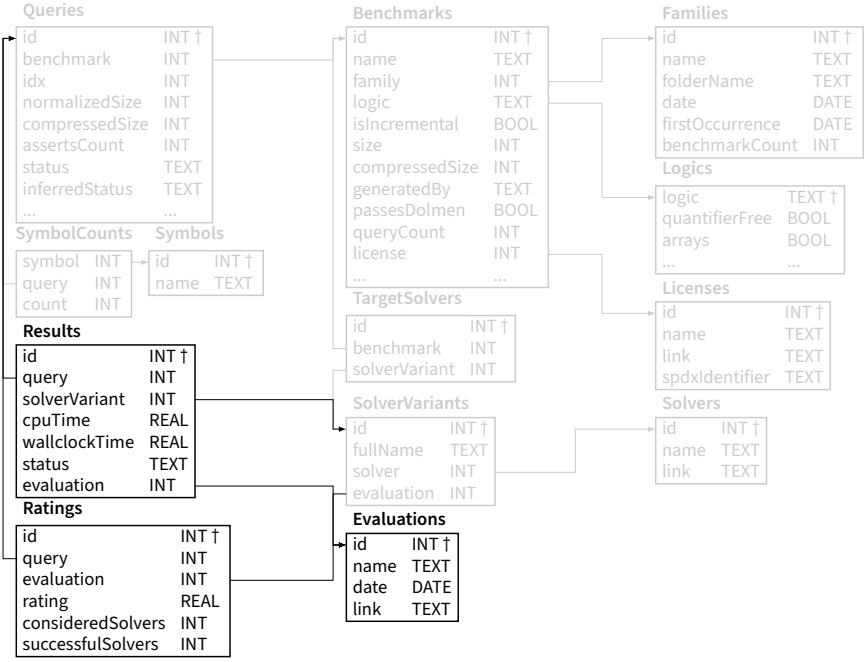


Part II
Database Structure


Queries
id INT †
benchmark INT
idx INT
normalizedSize INT
compressedSize INT
assertsCount INT
status TEXT
inferredStatus TEXT
... ...
SymbolCounts
symbol INT
query INT
count INT
Symbols
id INT †
name TEXT
Results
id INT †
query INT
solverVariant INT
cpuTime REAL
wallclockTime REAL
status TEXT
evaluation INT
Ratings
id INT †
query INT
evaluation INT
rating REAL
consideredSolvers INT
successfulSolvers INT
Benchmarks
id INT †
name TEXT
family INT
logic TEXT
isIncremental BOOL
size INT
compressedSize INT
generatedBy TEXT
passesDolmen BOOL
queryCount INT
license INT
... ...
TargetSolvers
id INT †
benchmark INT
solverVariant INT
SolverVariants
id INT †
fullName TEXT
solver INT
evaluation INT
Evaluations
id INT †
name TEXT
date DATE
link TEXT
Families
id INT †
name TEXT
folderName TEXT
date DATE
firstOccurrence DATE
benchmarkCount INT
Logics
logic TEXT †
quantifierFree BOOL
arrays BOOL
... ...
Licenses
id INT †
name TEXT
link TEXT
spdxIdentifier TEXT
Solvers
id INT †
name TEXT
link TEXT

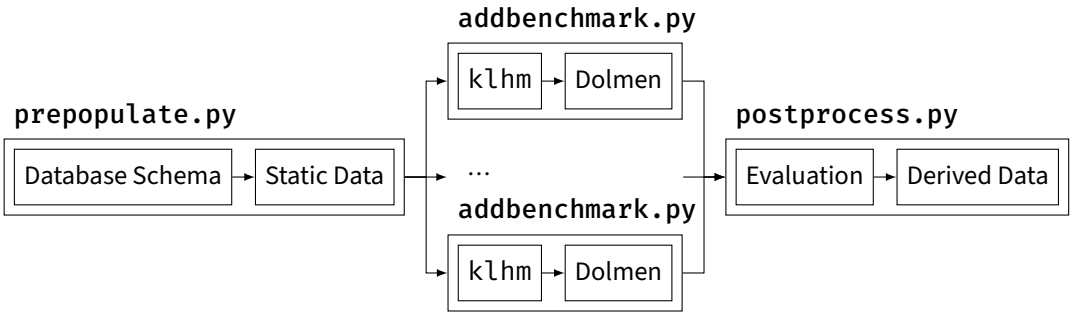

Queries
id INT †
benchmark INT
idx INT
normalizedSize INT
compressedSize INT
assertsCount INT
status TEXT
inferredStatus TEXT
... ...
SymbolCounts
symbol INT
query INT
count INT
Symbols
id INT †
name TEXT
Results
id INT †
query INT
solverVariant INT
cpuTime REAL
wallclockTime REAL
status TEXT
evaluation INT
Ratings
id INT †
query INT
evaluation INT
rating REAL
consideredSolvers INT
successfulSolvers INT
Benchmarks
id INT †
name TEXT
family INT
logic TEXT
isIncremental BOOL
size INT
compressedSize INT
generatedBy TEXT
passesDolmen BOOL
queryCount INT
license INT
... ...
TargetSolvers
id INT †
benchmark INT
solverVariant INT
SolverVariants
id INT †
fullName TEXT
solver INT
evaluation INT
Evaluations
id INT †
name TEXT
date DATE
link TEXT
Families
id INT †
name TEXT
folderName TEXT
date DATE
firstOccurrence DATE
benchmarkCount INT
Logics
logic TEXT †
quantifierFree BOOL
arrays BOOL
... ...
Licenses
id INT †
name TEXT
link TEXT
spdxIdentifier TEXT
Solvers
id INT †
name TEXT
link TEXT


Queries
id INT †
benchmark INT
idx INT
normalizedSize INT
compressedSize INT
assertsCount INT
status TEXT
inferredStatus TEXT
... ...
Benchmarks
id INT †
name TEXT
family INT
logic TEXT
isIncremental BOOL
size INT
compressedSize INT
generatedBy TEXT
passesDolmen BOOL
queryCount INT
license INT
... ...
Families
id INT †
name TEXT
folderName TEXT
date DATE
firstOccurrence DATE
benchmarkCount INT
Logics
logic TEXT †
quantifierFree BOOL
arrays BOOL
... ...
Licenses
id INT †
name TEXT
link TEXT
spdxIdentifier TEXT
SymbolCounts
symbol INT
query INT
count INT
Symbols
id INT †
name TEXT
TargetSolvers
id INT †
benchmark INT
solverVariant INT
Results
id INT †
query INT
solverVariant INT
cpuTime REAL
wallclockTime REAL
status TEXT
evaluation INT
SolverVariants
id INT †
fullName TEXT
solver INT
evaluation INT
Solvers
id INT †
name TEXT
link TEXT
Ratings
id INT †
query INT
evaluation INT
rating REAL
consideredSolvers INT
successfulSolvers INT
Evaluations
id INT †
name TEXT
date DATE
link TEXT


Queries
id INT †
benchmark INT
idx INT
normalizedSize INT
compressedSize INT
assertsCount INT
status TEXT
inferredStatus TEXT
... ...
SymbolCounts
symbol INT
query INT
count INT
Symbols
id INT †
name TEXT
Benchmarks
id INT †
name TEXT
family INT
logic TEXT
isIncremental BOOL
size INT
compressedSize INT
generatedBy TEXT
passesDolmen BOOL
queryCount INT
license INT
... ...
Families
id INT †
name TEXT
folderName TEXT
date DATE
firstOccurrence DATE
benchmarkCount INT
Logics
logic TEXT †
quantifierFree BOOL
arrays BOOL
... ...
Licenses
id INT †
name TEXT
link TEXT
spdxIdentifier TEXT
TargetSolvers
id INT †
benchmark INT
solverVariant INT
Results
id INT †
query INT
solverVariant INT
cpuTime REAL
wallclockTime REAL
status TEXT
evaluation INT
SolverVariants
id INT †
fullName TEXT
solver INT
evaluation INT
Solvers
id INT †
name TEXT
link TEXT
Ratings
id INT †
query INT
evaluation INT
rating REAL
consideredSolvers INT
successfulSolvers INT
Evaluations
id INT †
name TEXT
date DATE
link TEXT

How We Build The Catalog

Pipeline implemented as a set of Python scripts. Hand-curated data is pure Python.



Extracting Benchmark Data

Klammerhammer

- Standalone tool
- Produces JSON
  - one object per query
- Handles `push`/`pop`

Implementation

- Written in Zig
- Hardcoded symbol lists
- One linear scan
- Compression using **zstd** C API (quadratic!)

Competition Data

We collect results from **all** SMT-COMPs for the **current** benchmarks.

2005, 2006
primeval competitions – raw HTML webpages

2007–2012
SMTexec. Results were lost, we found the MySQL database.

2013
SMT-EVAL. Excellent documentation. CSV.

2014–2023
StarExec. CSV until 2016, then JSON. Some larger changes to the benchmark set.

2024–
BenchCloud. same JSON file. Incremental tracks!

Difficulty Rating

$$1 - \frac{\#\text{ solvers with a successful variant}}{\#\text{ solvers with a variant that attempted the query}}$$

Caveats

- Differences between competitions
  - Hardware
  - Time limit
  - Memory limit
- Small number of solvers

Differences to TPTP

TPTP omits superseded solvers

- Intention: ignore experimental variants
- We track variants explicitly!
- Every solver represents serious work.

Conclusion

Releases

- Yearly release with benchmarks
  - Schema might change
  - Will include the prior years SMT-COMP results
- We want you to play around with it
- and welcome feature requests!

`https://zenodo.org/records/16290040`

Fun Facts

- Most queries: $2\,630\,828$
- Largest query: 1.9 GB (48.20 MB compressed)
- Deepest term: $3\,515\,188$
- Best solver name: **Tiffany de Wintermonte**