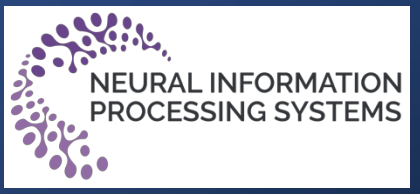


Is Function Similarity Over Engineered? Building a Benchmark

Rebecca Saul[1,2], Chang Liu[3], Noah Fleischmann[1,2], Richard Zak[1,2,4]
Kristopher Micinski[3], Edward Raff[1,2,4], James Holt[1]

1 Laboratory for Physical Sciences 2 Booz Allen Hamilton 3 Syracuse University 4 U.M.B.C.

Binary Function Similarity Detection (BFSD)

The problem of determining whether two binary functions are similar in the absence of source code.

- A core component of critical security tasks including reverse engineering, malware analysis, and vulnerability detection.
- Many ML approaches, using a variety of features (raw bytes, disassembly, control flow graphs (CFGs), dynamic analysis) and architectures (CNNs, RNNs, GNNs, Transformers).

No Meaningful BFSD Benchmarks

Existing datasets are:

- Small – the median model is trained on less than 4k binaries
- Unrepresentative – composed of Linux binaries, even though real-world BFSD primarily interacts with Windows binaries
- Underspecified – missing details about deduplication and labeling

REFuSe-Bench: 6 Principles for a New Benchmark

Any binary/project application should have all of its functions in only one of the train/test sets, not both.

You must check for the same function across binaries.

Designers need to be specific on labeling details with code.

Allow standard compiler optimizations.

Use larger datasets with Windows executables.

Do not restrict the search space using information not available at deployment time.

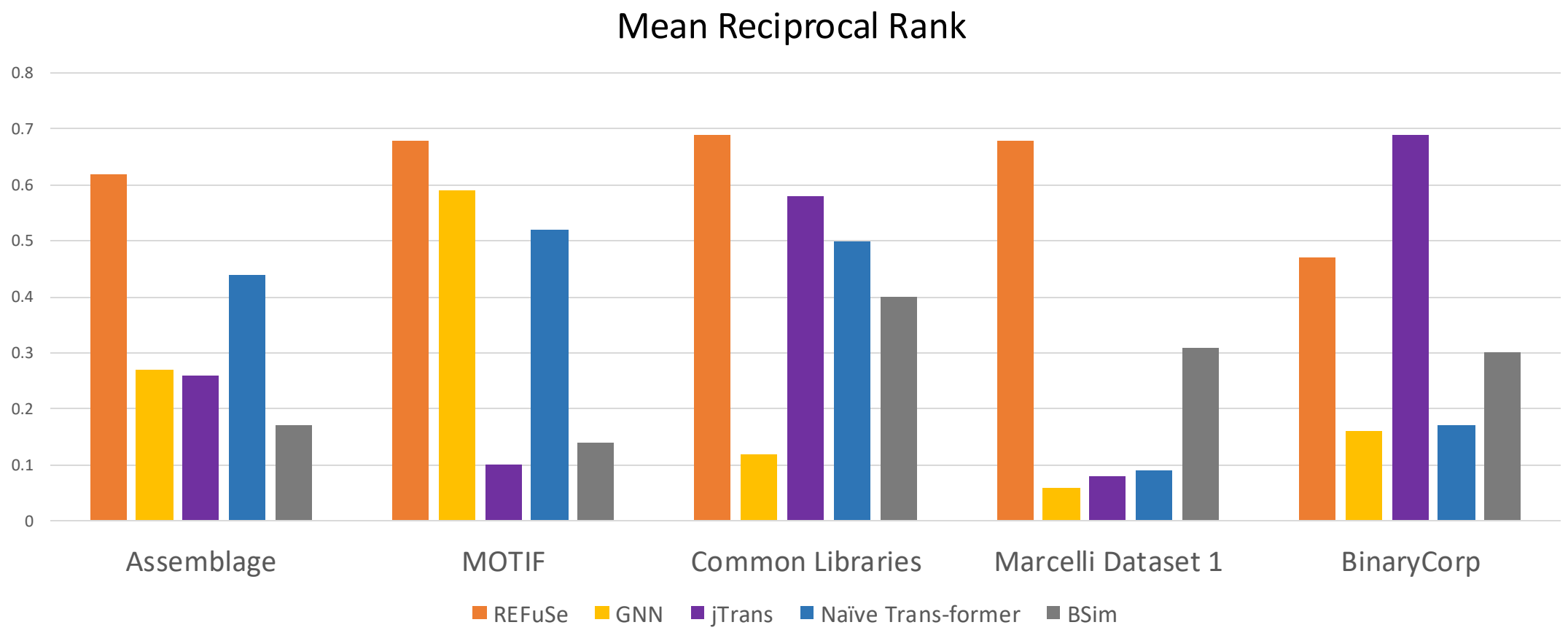
The REFuSe-Bench Datasets

Dataset	OS	No. Binaries	No. Functions	Composition
Assemblage	Windows	135,975	24,545,694	C/C++ GitHub Projects
MOTIF	Windows	3095	2,442,164	Malware (454 Families)
Common Libraries	Windows	40	106,545	abseil, cjson, glfw3, libxml2, openssl, sdl1, zlib
Marcelli Dataset 1	Linux	919	668,400	nmap, z3
BinaryCorp	Linux	9,675	4,791,673	ArchLinux, Arch User Repository

The REFuSe-Bench Models

Model	Function Representation	Architecture	Training Data
REFuSe	Raw bytes	CNN	Assemblage
GNN (Li et. al., 2019)	CFGs	GNN	Marcelli Dataset-1
jTrans	Disassembly	Transformer-Encoder	BinaryCorp
Naïve Transformer	Raw bytes	Transformer Encoder	Assemblage
BSim (Ghidra)	P-code	Hand-crafted feature vectors	UNKNOWN

Results


Mean Reciprocal Rank
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Assemblage
MOTIF
Common Libraries
Marcelli Dataset 1
BinaryCorp
REFuSe
GNN
jTrans
Naïve Trans-former
BSim

Connect with us!

- Corresponding Author: Rebecca Saul
  - saul_rebecca@bah.com

- Paper: https://arxiv.org/pdf/2410.22677

- GitHub: https://github.com/FutureComputing4AI/Reverse-Engineering-Function-Search