Fast and Slow Enigmas and Parental Guidance

Zarathustra Goertzel, Karel Chvalovský, Jan Jakubův, Miroslav Olšák, and Josef Urban

Czech Technical University in Prague
University of Innsbruck, Austria

FROCOS 2021

E Prover (a Saturation-based ATP)

- Goal: Prove conjecture from premises.
- E has two sets of clauses:
  - *Processed* clauses P (initially empty)
  - *Unprocessed* clauses U (Negated Conjecture and Premises)
- Given Clause Loop:
  - Select ‘*given clause*’ g to add to P
  - Apply *inference rules* to g and all clauses in P
  - Process new clauses. Add non-trivial and non-redundant ones to U.
- Proof search succeeds when empty clause is inferred.
- Proof consists of some of the given clauses.

Given Clause Loop in E

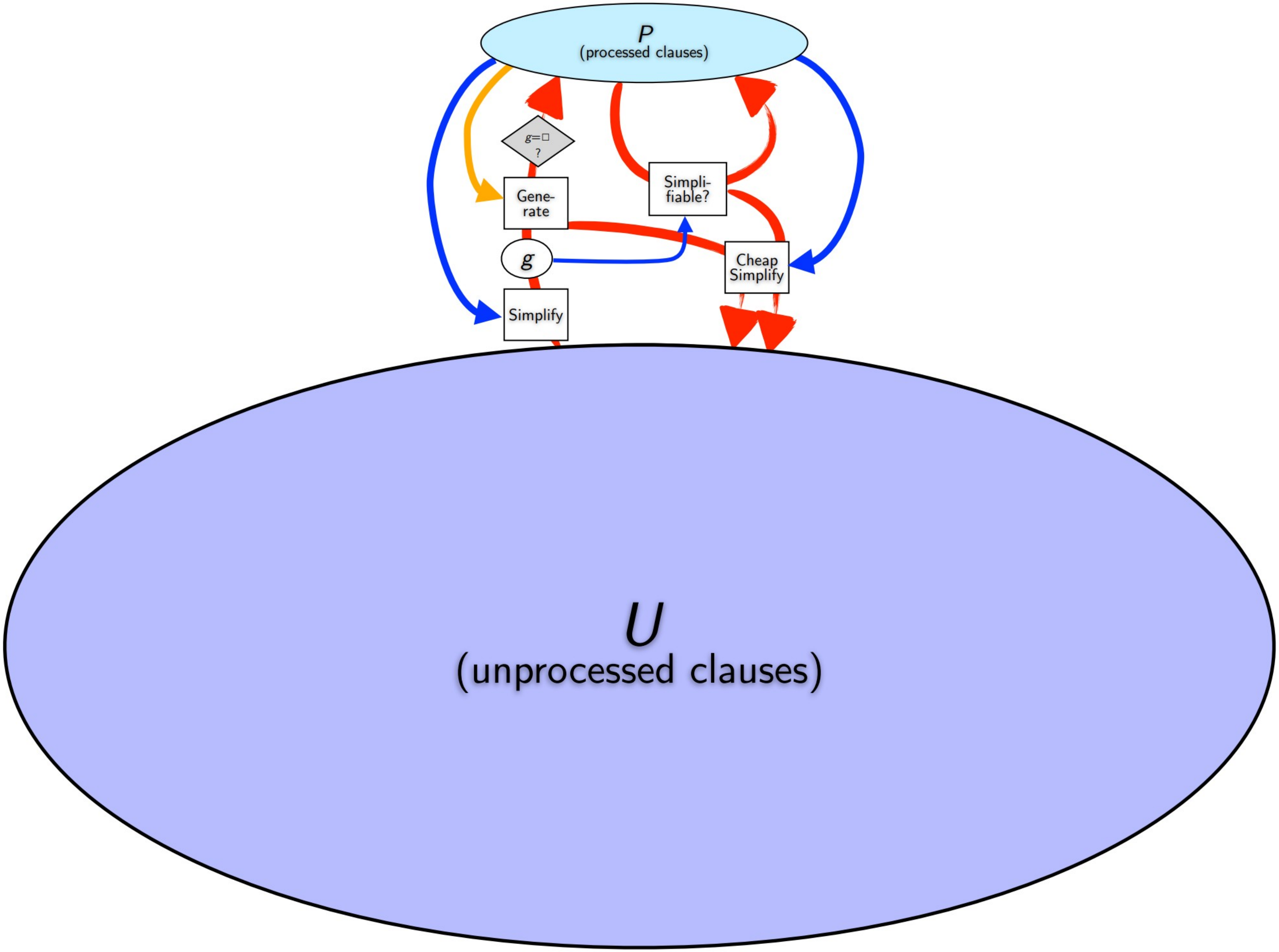


Image thanks to Stephan Schulz's presentation on E

E Strategies

- *Clause Evaluation Functions (CEFs)* consist of:
  - *Priority functions*: partition clauses into priority queues.
    - e.g., *ConstPrio, PreferUnit*
  - **Weight functions**: order clauses in queues based on a score.
    - e.g.: **Clauseweight**, **FIFOWeight**
- Weighted by frequency of use, for example:

-H'(5*Clauseweight(ConstPrio,1,1,1), 1*FIFOWeight(ConstPrio))'

Given Clause Loop in E + ML Guidance

<u>Clause Selection Models</u>:

Fast – GBDT
Slow – GNN

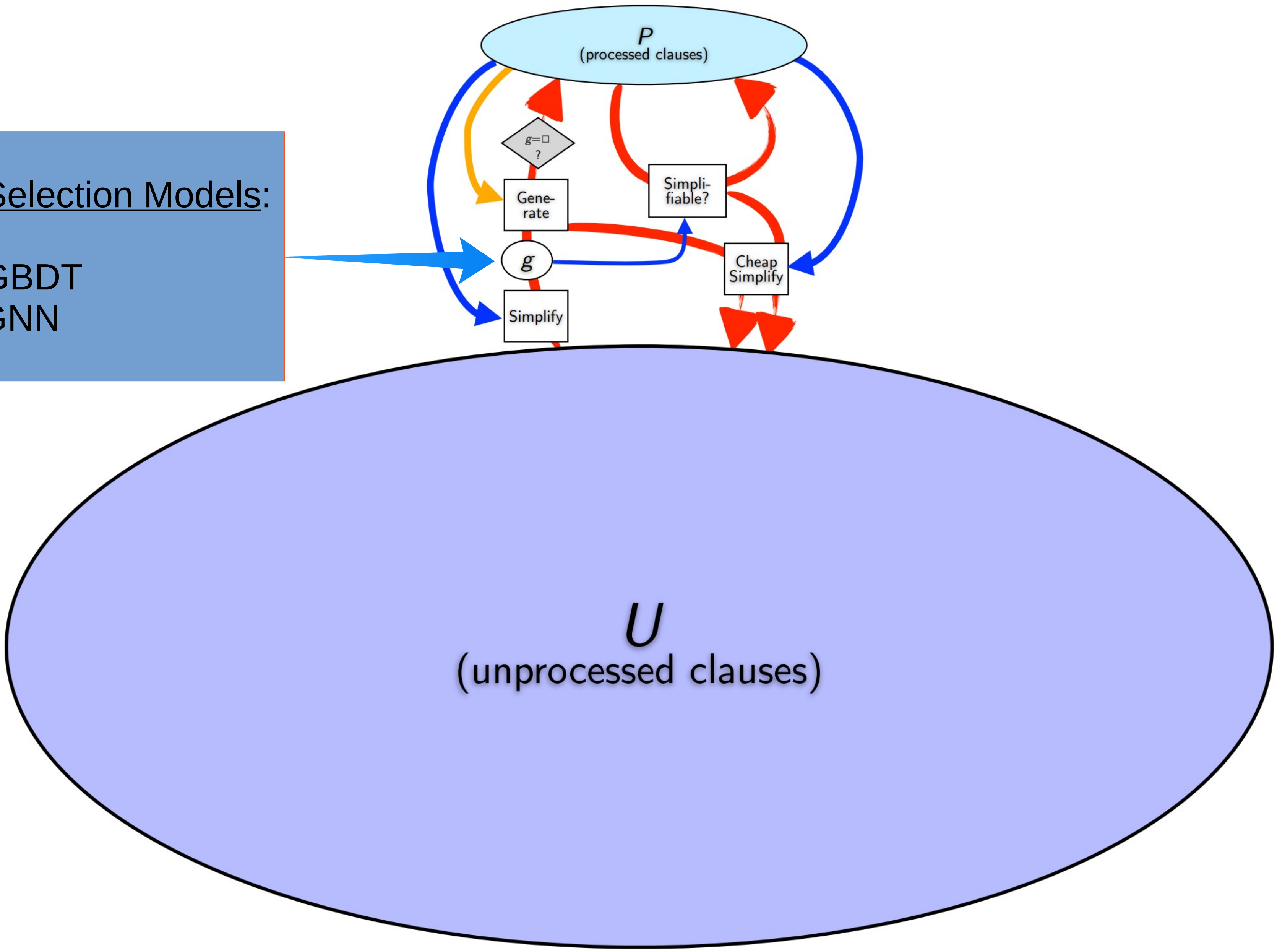
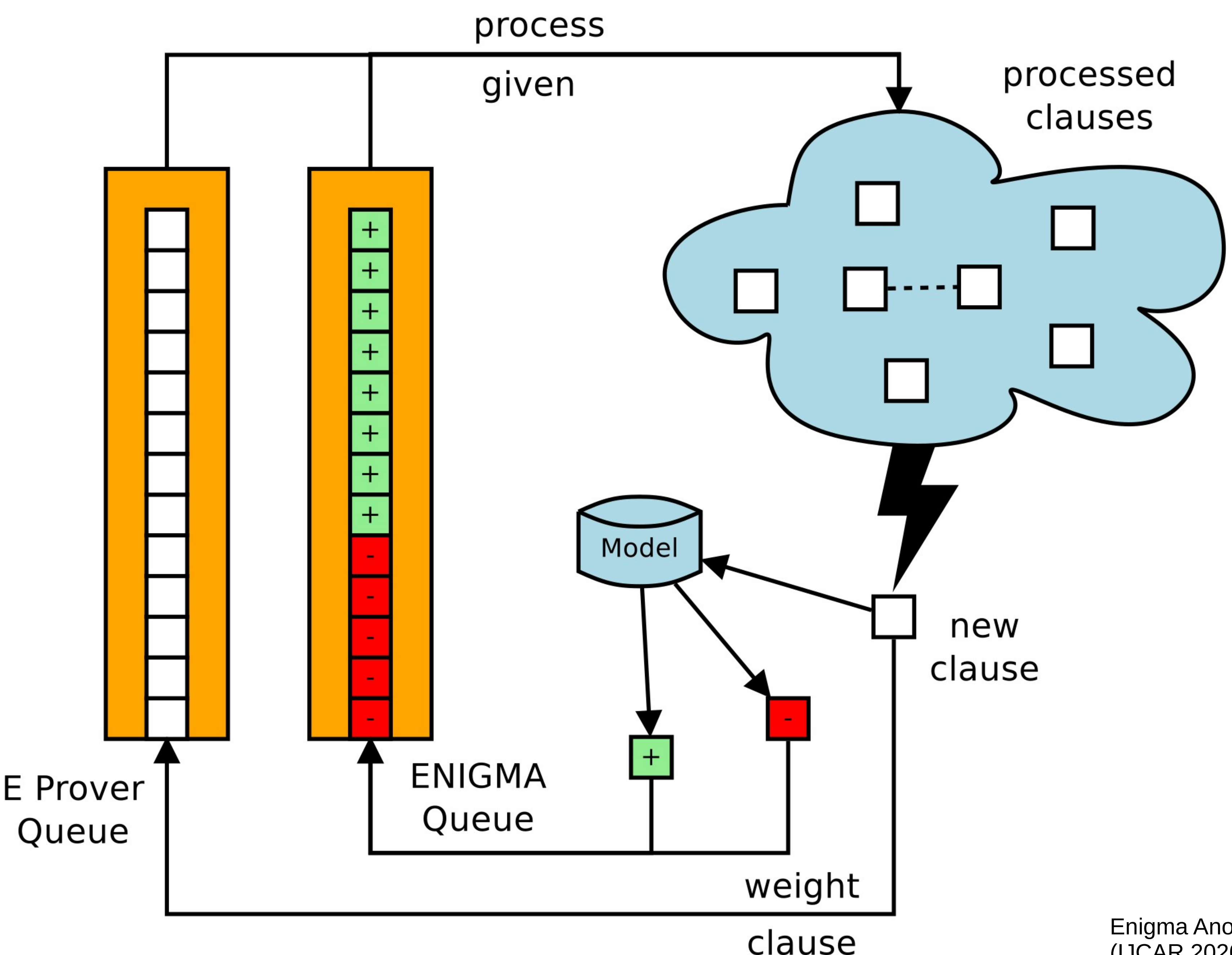


Image thanks to Stephan Schulz’s presentation on E

Given Clause Loop in E + ML Guidance



Mizar Experiment Setting

- Mizar Mathematical Library (MML) – 57880 problems

  1148 articles

Given Clause Loop in E + ML Guidance

Prior work

<u>Clause Selection Models</u>:

Fast – E CEFs: 15k/58k
Fast – GBDT: 24.3k/58k
Slow – GNN: 23.3k/58k
Joint training*: 38k/58k

Enigma Anonymous
(IJCAR 2020)

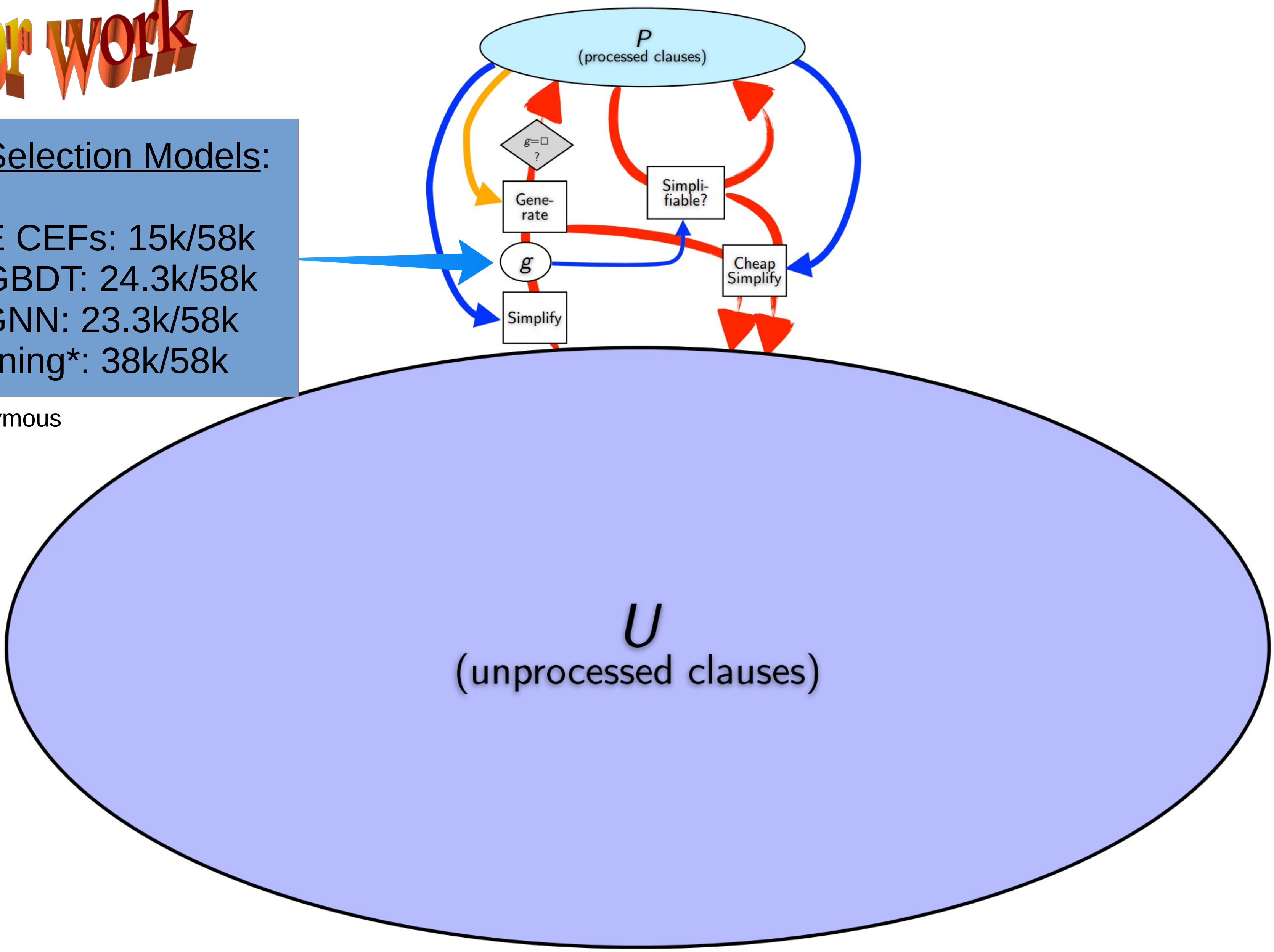


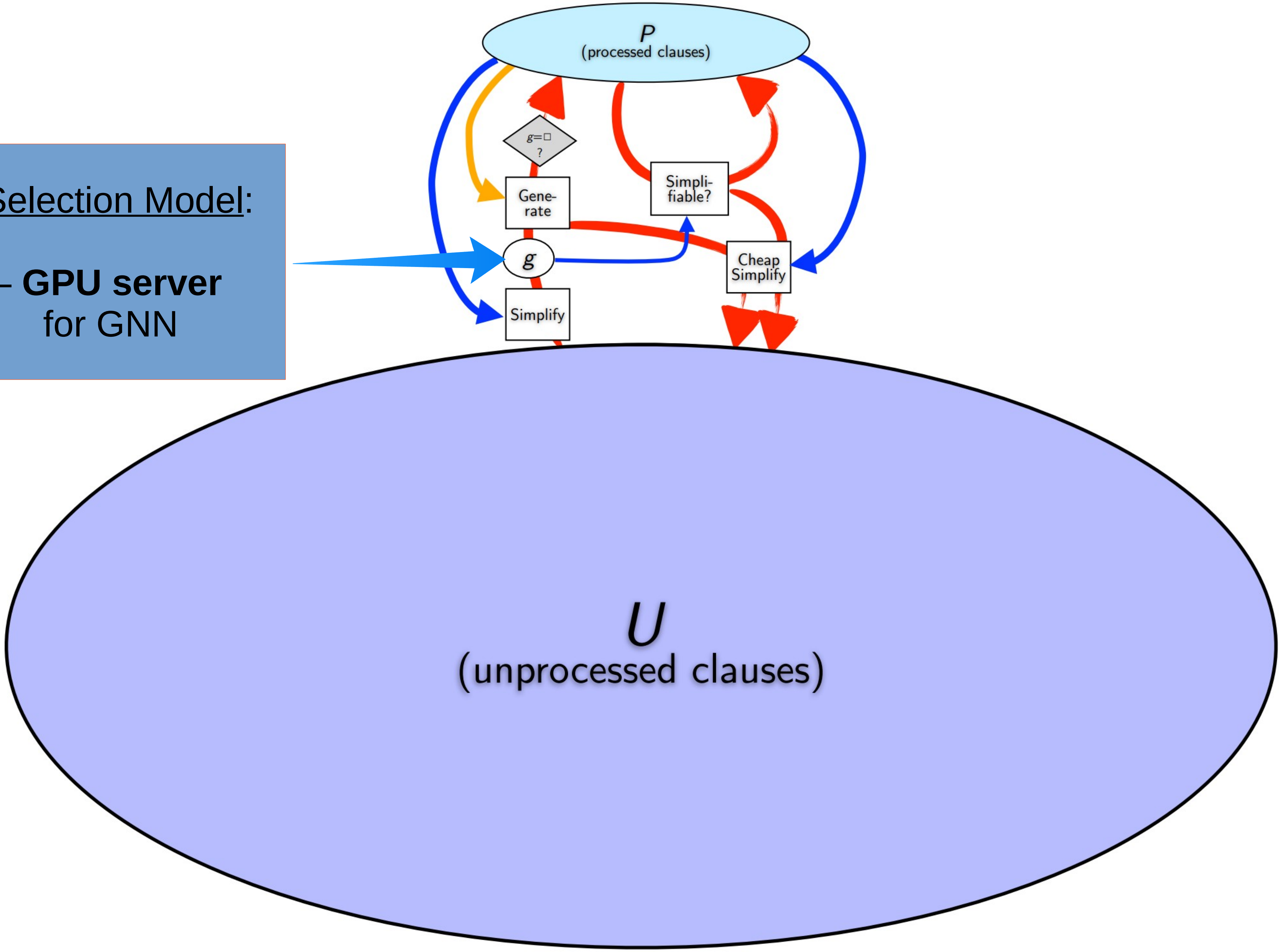
Image thanks to Stephan Schulz’s presentation on E

Given Clause Loop in E + ML Guidance

Contribution 1

<u>Clause Selection Model</u>:

Slow++ – **GPU server** for GNN

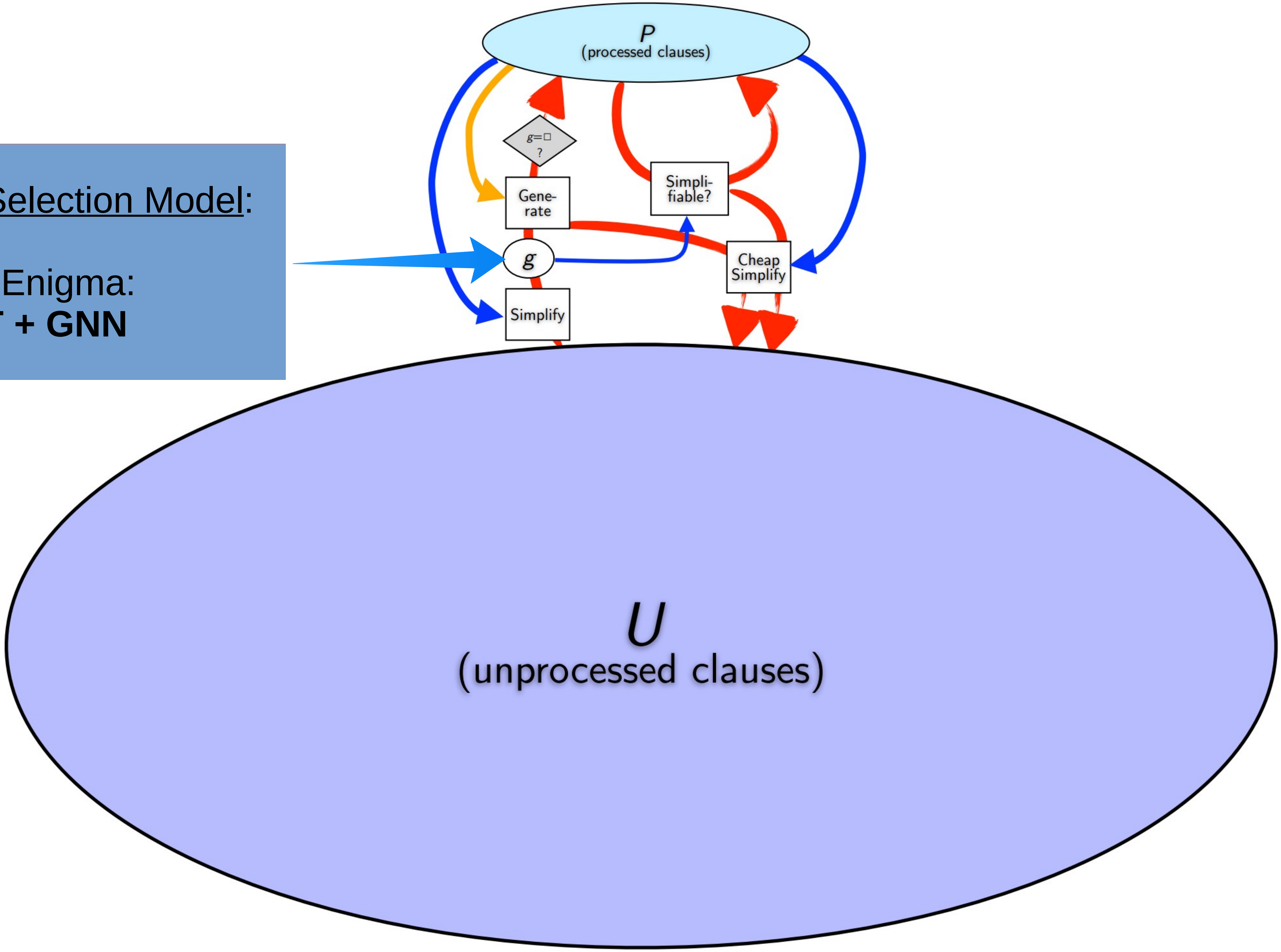


Given Clause Loop in E + ML Guidance

Contribution 2

<u>Clause Selection Model</u>:

2-phase Enigma:
GBDT + GNN



Given Clause Loop in E + ML Guidance

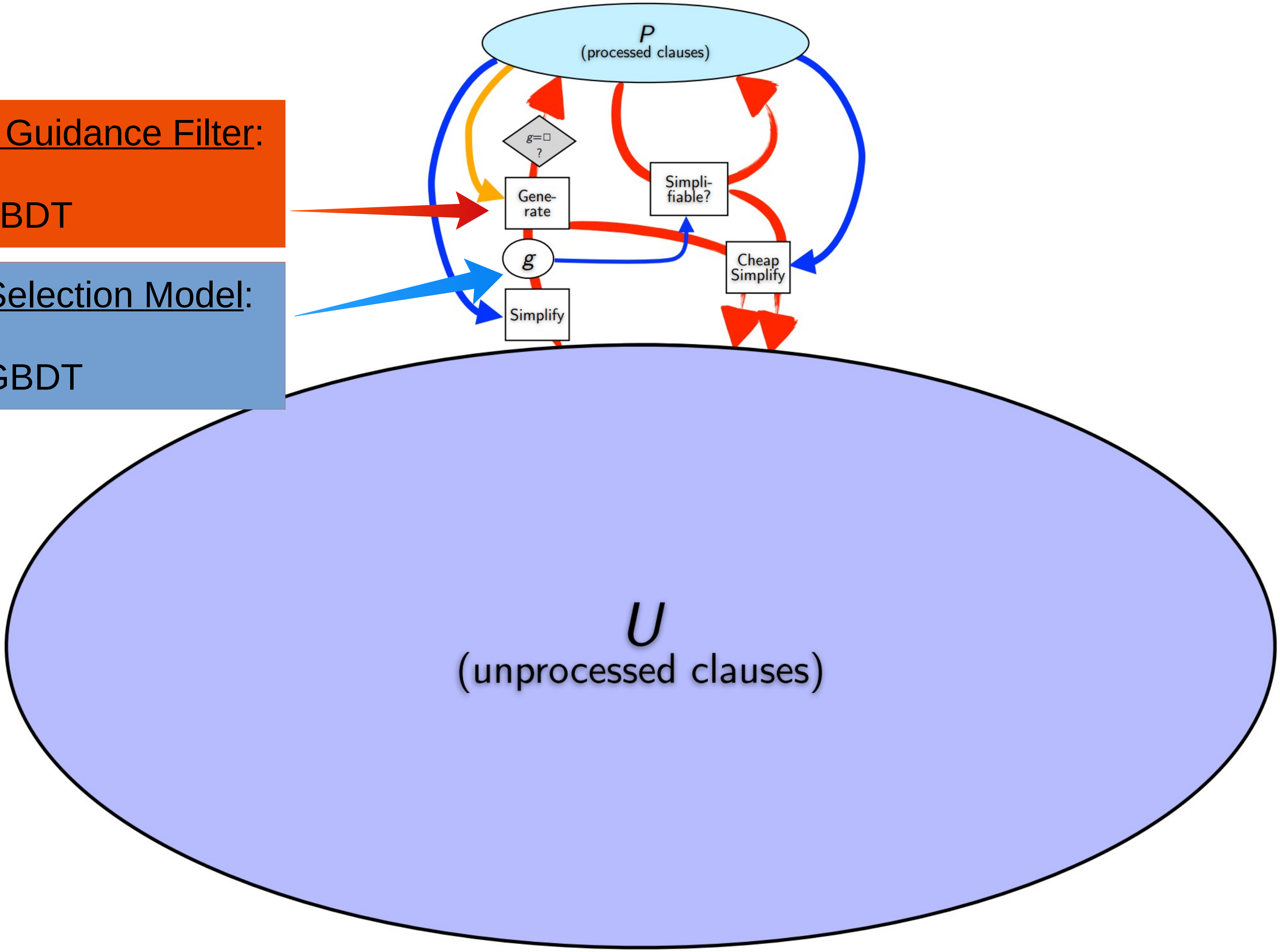
Contribution 3

__Parental Guidance Filter__:

Fast – GBDT

__Clause Selection Model__:

Fast – GBDT



Given Clause Loop in E + ML Guidance

Contribution 4

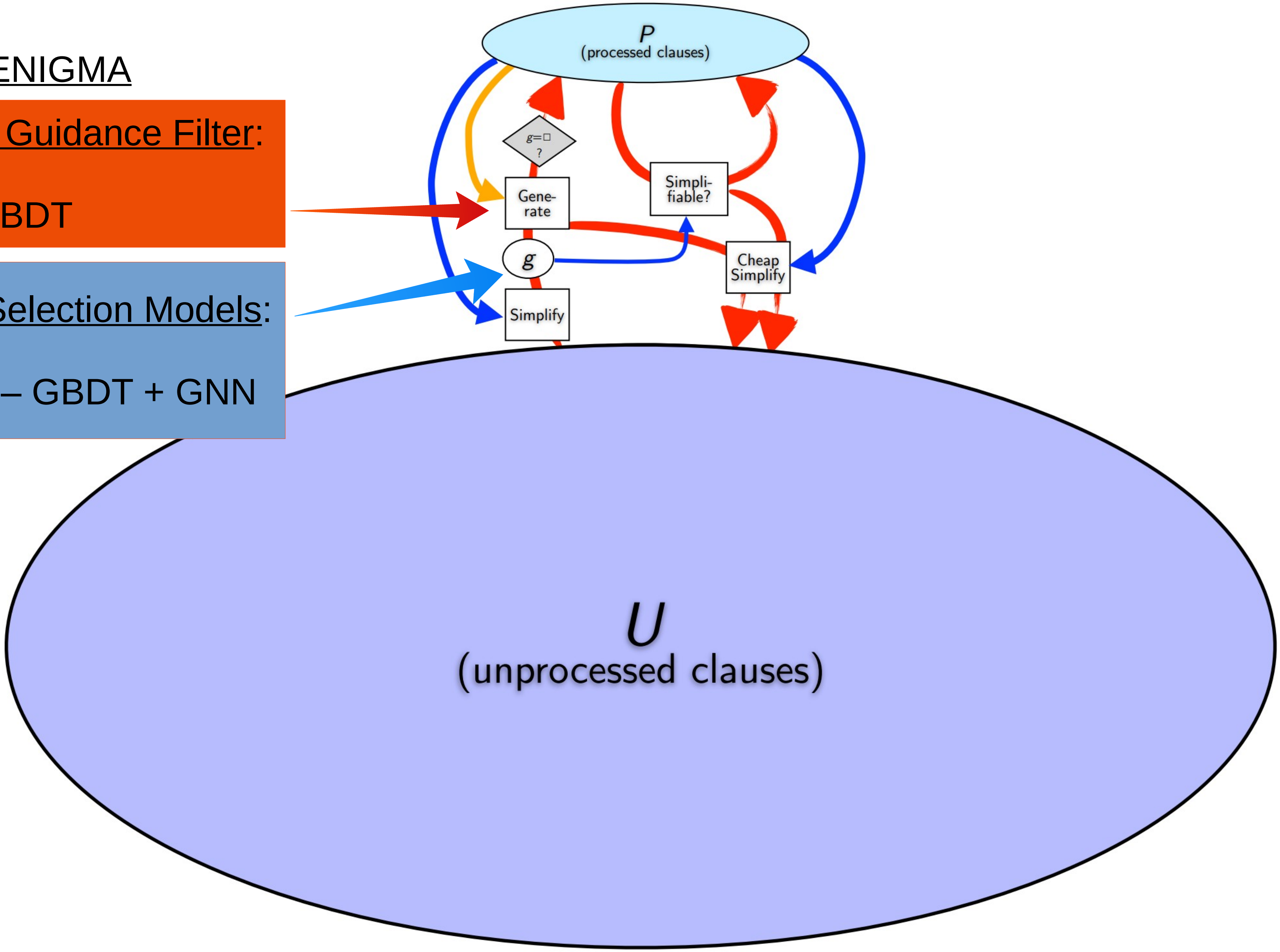
3-phase ENIGMA

Parental Guidance Filter:

Fast – GBDT

Clause Selection Models:

2-phase – GBDT + GNN



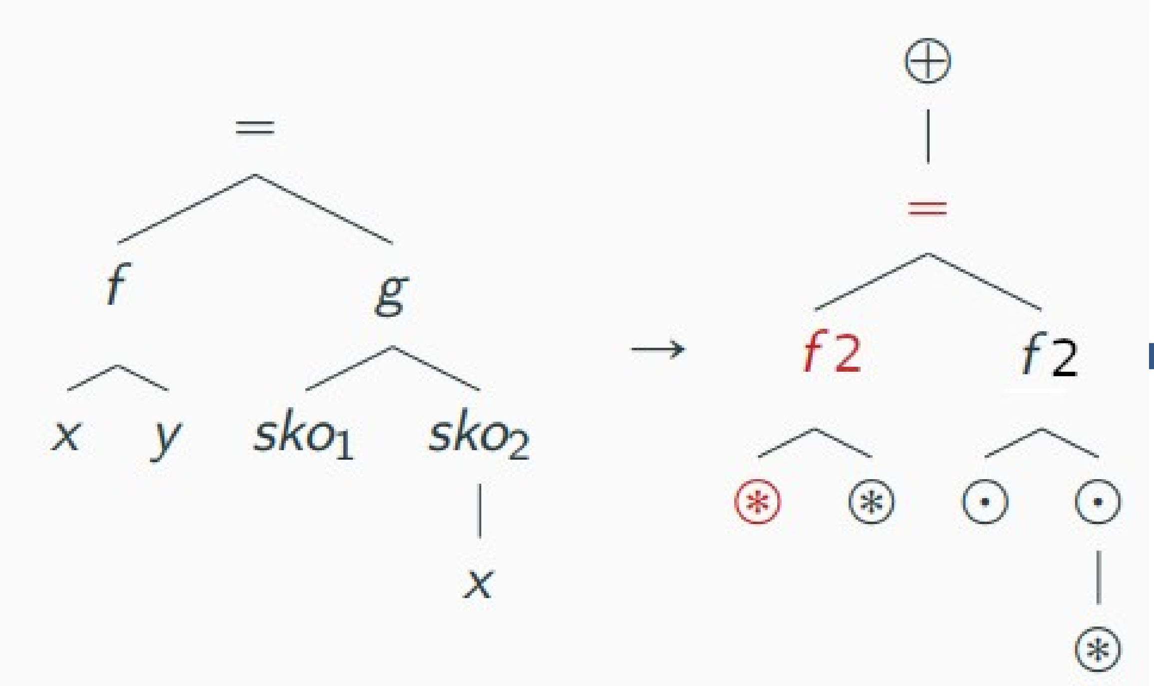
ENIGMA Anonymous

- Statistical machine learning for <u>clause selection</u>.
  - LightGBM (gradient boosted decision trees)
  - GNN (Graph Neural Network)
- Learns from given clauses:
  - *Positive* if in a proof
  - *Negative* otherwise
- Guides E via a weight function.
- Features: given clause + conjecture + theory

Featurization: clauses ➜ vectors

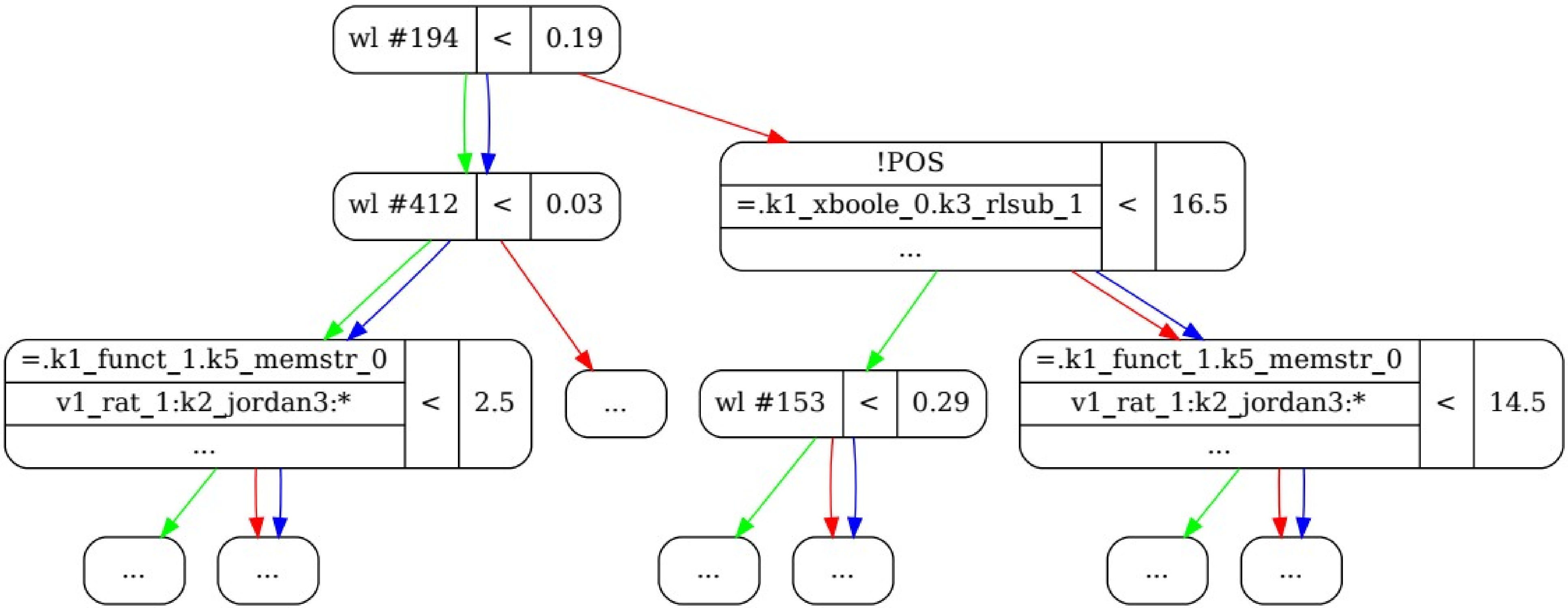
- Treat clauses as trees.
- Abstract vars and skolem symbols.
- Anonymize function and predicate symbols of arity *n* with "**f***n*" or "**p***n*".
- Hash features to reduce dimensionality.
- The clause vector consists of feature counts.

For example: $f(x, y) = g(\mathrm{sko}_1, \mathrm{sko}_2(x))$



#	feature	count
1	(⊕,=,a)	0
⋮	⋮	⋮
11	(⊕,=,f2)	1
12	(⊕,=,f2)	1
13	(=,f2,⊛)	2
14	(=,f2,⊙)	2
15	(f2,⊙,⊛)	1
⋮	⋮	⋮

Gradient Boosted Decision Tree



*XGBoost tree with non-anonymized watchlist features

ENIGMA-GNN

- Graph Neural Network
- Directed hypergraph for a set of clauses
- Anonymized symbol names
- Nodes: clauses, functions and predicate symbols, unique (sub)terms, and literals
- Hyperedges:
  1) Clauses and literals
  2) Functions and predicates with subterms
- Message passing rounds → clause embedding
- Prediction layer

ENIGMA-GNN

- Application $a = f(x_1, x_2, \ldots, x_n)$ is represented by a set of 4-ary hyperedges $(f, a, x_1, x_2), (f, a, x_2, x_3), \ldots, (f, a, x_{n-1}, x_n)$.

Hyperedges

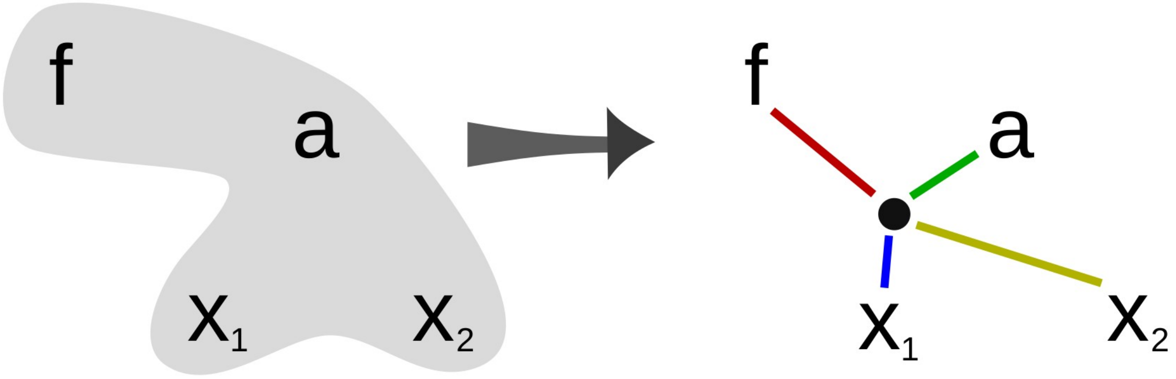


Image courtesy of
http://arg.ciirc.cvut.cz/teaching/mlr20/slides/ai4reason-mlr20-gnn.pdf

GPU Server for Faster GNN Eval

- Persistent multi-threaded GPU server
- E clients send batches of clauses for evaluation
- GPU start-up costs are amortized
- More E clients can run in parallel
  (while waiting for the GPU).

2-phase ENIGMA

- Use a GBDT model as a pre-filter for the GNN
- Clauses with high scores are given high weights
- Otherwise, the GNN scores the clauses

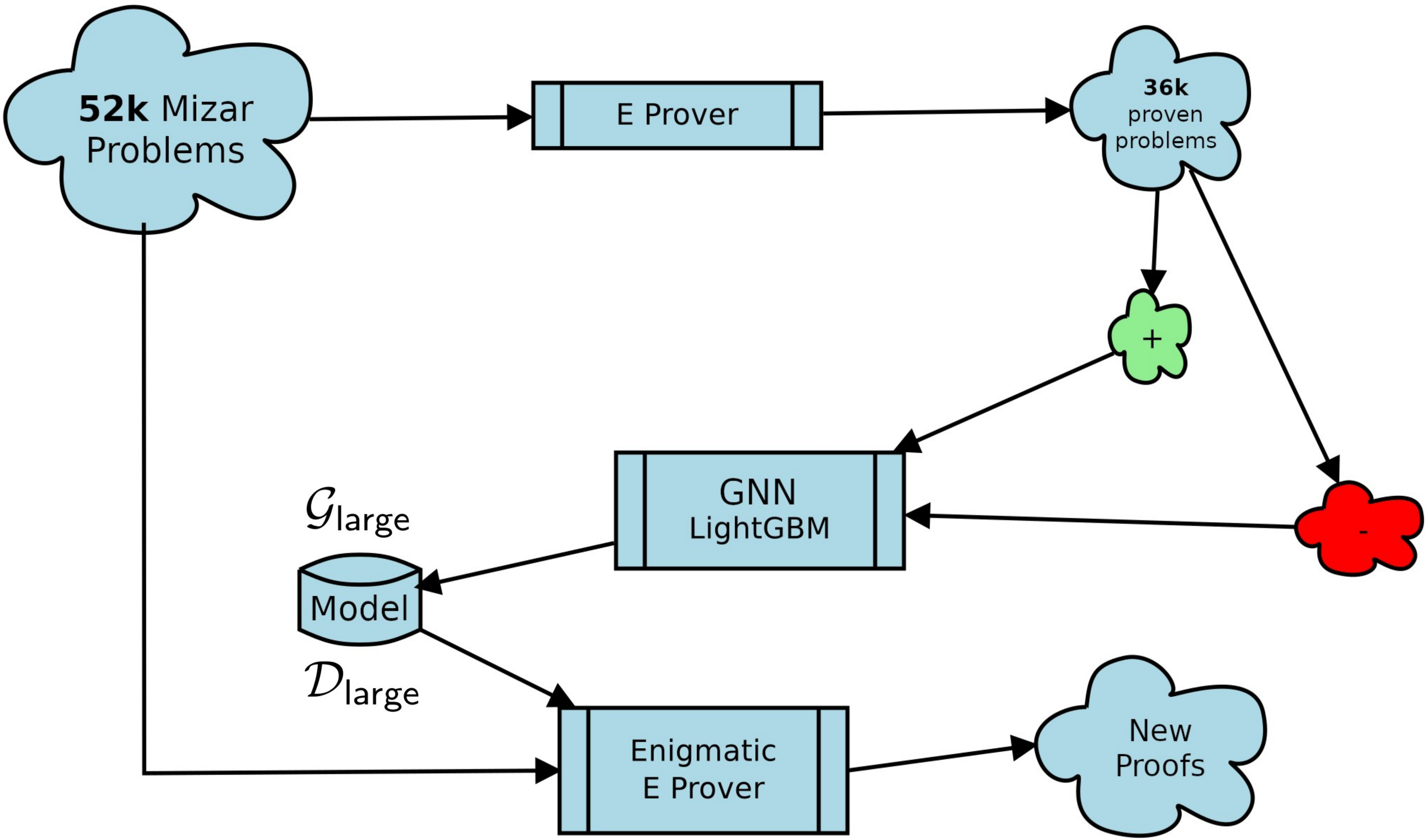
Parental Guidance

- Clause evaluation based on parent clause features.
- Score only *valid* pairs of parents with GBDT model.
- Freeze all clauses with scores below *threshold*.
- Unfreeze clauses if the unprocessed set empties.
- Run with $\mathcal{D}_{\mathsf{large}}$ to select from unfiltered clauses.
- Feature vector options:
  - $\mathcal{P}_{\mathsf{fuse}}$: merge parent vectors into one
  - $\mathcal{P}_{\mathsf{cat}}$: concatenate parent vectors

Mizar Experiment Settings

- 90-5-5% train-development-holdout split
  - Training: 52k problems
    - Small trains: 5792 problems
  - Development: 2896 problems
    - Small devel: 300 problems
  - Holdout: 2896 problems
- Already have 36k problems solved on *training.*
- Train baseline models: GDBT, $\mathcal{D}_{\text{large}}$ ($\mathcal{D}_{\text{small}}$), and GNN, $\mathcal{G}_{\text{large}}$ ($\mathcal{G}_{\text{small}}$)

Mizar Experiment Settings



Parental Guidance Training Data

- Need to judge *all generated clauses*!
- What are responsible parents?

$\mathcal{P}^{\text{proof}}$) Parents of proof clauses are *positive*.

$\mathcal{P}^{\text{given}}$) Parents of processed clauses are *positive*.

Parental Guidance Training Data

- Need to judge *all generated clauses*!
- What are responsible parents?

$\mathcal{P}^{\text{proof}}$) Parents of proof clauses are *positive*.

$\mathcal{P}^{\text{given}}$) Parents of processed clauses are *positive*.

- Run $\mathcal{D}_{\text{large}}$ or $\mathcal{G}_{\text{large}}$ on 52k training set to create data
- **Pos-neg-ratio**: with $\mathcal{P}^{\text{proof}}$ data, is 1:192!

GPU Server Speedup Results

- Compare in the context of parallelization with the CPUs fully saturated:
  - 70-fold parallelization for CPU-only
  - 160-fold parallelization for GPU-server

GPU Server Speedup Results

- Compare in the context of parallelization with the CPUs fully saturated:
  - 70-fold parallelization for CPU-only
  - 160-fold parallelization for GPU-server

set	model	method	time	solved
D	$\mathcal{G}_{\text{large}}$	CPU	30	1311
D	$\mathcal{G}_{\text{large}}$	CPU	60	1380
D	$\boldsymbol{\mathcal{G}}_{\textbf{large}}$	**GPU**	**30**	**1511** (**+9.5%**)

set	model	method	time	solved
H	$\mathcal{G}_{\text{large}}$	CPU	30	1301
H	$\mathcal{G}_{\text{large}}$	CPU	60	1371
H	$\boldsymbol{\mathcal{G}}_{\textbf{large}}$	**GPU**	**30**	**1529** (**+11.5%**)

GPU Server Speedup Results

- Compare in the context of parallelization with the CPUs fully saturated:
  - 70-fold parallelization for CPU-only
  - 160-fold parallelization for GPU-server

set	model	method	time	solved
D	$\mathcal{G}_{\mathsf{large}}$	CPU	30	1311
D	$\mathcal{G}_{\mathsf{large}}$	CPU	60	1380
D	$\mathcal{G}_{\mathsf{large}}$	**GPU**	**30**	**1511** (+**9.5**%)

set	model	method	time	solved
H	$\mathcal{G}_{\mathsf{large}}$	CPU	30	1301
H	$\mathcal{G}_{\mathsf{large}}$	CPU	60	1371
H	$\mathcal{G}_{\mathsf{large}}$	**GPU**	**30**	**1529** (+**11.5**%)

- Generates over 4x the clauses in the time limit!

2-phase ENIGMA

- Parameter grid searches on size 300 devel set:
  - Filter threshold
  - GNN clause query size
  - GNN context size

2-phase ENIGMA

Table 2. Final evaluation of the best combination of $\mathcal{D}_{\mathsf{small}}$ with $\mathcal{G}_{\mathsf{small}}$ on the whole development (D) and holdout (H) datasets of size 2896.

set	model	thresh.	time	query	context	solved
D	$\mathcal{G}_{\mathsf{small}}$	-	30	256	768	1251
D	$\mathcal{D}_{\mathsf{small}}$	-	30	-	-	1011
D	$\mathcal{D}_{\mathsf{small}}+\mathcal{G}_{\mathsf{small}}$	**0.01**	**60**	**512**	**1024**	**1381 (+10.4%)**
D	$\mathcal{D}_{\mathsf{small}}+\mathcal{G}_{\mathsf{small}}$	0.03	60	512	1024	1371 (+9.6%)
D	$\mathcal{D}_{\mathsf{small}}+\mathcal{G}_{\mathsf{small}}$	0.03	30	512	1024	1341 (+7.2%)
D	$\mathcal{D}_{\mathsf{small}}+\mathcal{G}_{\mathsf{small}}$	0.01	30	512	1024	1339 (+7.0%)
H	$\mathcal{G}_{\mathsf{small}}$	-	30	256	768	1277
H	$\mathcal{D}_{\mathsf{small}}$	-	30	-	-	1002
H	$\mathcal{D}_{\mathsf{small}}+\mathcal{G}_{\mathsf{small}}$	**0.01**	**60**	**512**	**1024**	**1392 (+9.0%)**
H	$\mathcal{D}_{\mathsf{small}}+\mathcal{G}_{\mathsf{small}}$	0.03	60	512	1024	1387 (+8.6%)
H	$\mathcal{D}_{\mathsf{small}}+\mathcal{G}_{\mathsf{small}}$	0.01	30	512	1024	1361 (+6.6%)
H	$\mathcal{D}_{\mathsf{small}}+\mathcal{G}_{\mathsf{small}}$	0.03	30	512	1024	1353 (+6.0%)

2-phase ENIGMA

Table 3. Final evaluation of the best combination of $\mathcal{D}_{\mathsf{large}}$ and $\mathcal{G}_{\mathsf{small}}$ on the whole development (D) and holdout (H) datasets of size 2896.

set	model	thresh.	time	query	context	solved
D	$\mathcal{G}_{\mathsf{small}}$	-	30	256	768	1251
D	$\mathcal{D}_{\mathsf{large}}$	-	30	-	-	1397
D	$\mathcal{D}_{\mathbf{large}}+\mathcal{G}_{\mathbf{small}}$	**0.3**	**60**	**2048**	**768**	**1527** (**+9.3%**)
D	$\mathcal{D}_{\mathsf{large}}+\mathcal{G}_{\mathsf{small}}$	0.3	30	2048	768	1496 (+7.1%)
H	$\mathcal{G}_{\mathsf{small}}$	-	30	256	768	1277
H	$\mathcal{D}_{\mathsf{large}}$	-	30	-	-	1390
H	$\mathcal{D}_{\mathbf{large}}+\mathcal{G}_{\mathbf{small}}$	**0.3**	**60**	**2048**	**768**	**1494** (**+7.5%**)
H	$\mathcal{D}_{\mathsf{large}}+\mathcal{G}_{\mathsf{small}}$	0.3	30	2048	768	1467 (+5.5%)

2-phase ENIGMA

Table 4. Final evaluation of the best combination of $\mathcal{D}_{\mathsf{large}}$ and $\mathcal{G}_{\mathsf{large}}$ on the whole development (D) and holdout (H) datasets of size 2896.

set	model	thresh.	time	query	context	solved
D	$\mathcal{G}_{\mathsf{large}}$	-	30	256	768	1511
D	$\mathcal{D}_{\mathsf{large}}$	-	30	-	-	1397
D	$\boldsymbol{\mathcal{D}_{\mathsf{large}}+\mathcal{G}_{\mathsf{large}}}$	**0.1**	**60**	**1024**	**768**	**1648 (+9.1%)**
D	$\mathcal{D}_{\mathsf{large}}+\mathcal{G}_{\mathsf{large}}$	0.1	30	1024	768	1615 (+6.9%)
H	$\mathcal{G}_{\mathsf{large}}$	-	30	256	768	1529
H	$\mathcal{D}_{\mathsf{large}}$	-	30	-	-	1390
H	$\boldsymbol{\mathcal{D}_{\mathsf{large}}+\mathcal{G}_{\mathsf{large}}}$	**0.1**	**60**	**1024**	**768**	**1640 (+7.3%)**
H	$\mathcal{D}_{\mathsf{large}}+\mathcal{G}_{\mathsf{large}}$	0.1	30	1024	768	1602 (+4.8%)

Parental Guidance

- Parameter grid searches on size 300 devel set:
  - Filter threshold (0.005 to 0.5)
  - Pos-neg reduction ratio (1 to 16 or ‘as is’)
  - LightGBM params:
    - Num. trees
    - Num. leaves
    - Max depth
  - Parent vector feature form: $\mathcal{P}_{\mathsf{fuse}}$ vs $\mathcal{P}_{\mathsf{cat}}$
  - Data curation method: $\mathcal{P}^{\mathsf{given}}$ vs $\mathcal{P}^{\mathsf{proof}}$

Parental Guidance

Table 5. The best threshold for each tested reduction ratio. The threshold of 0.03 was identical to 0.05 for all tested ratios with $\mathcal{P}^{\mathsf{given}}_{\mathsf{fuse}}$, whereas there are no ties among thresholds for $\mathcal{P}^{\mathsf{proof}}_{\mathsf{fuse}}$.

$\rho^{\mathsf{given}}_{\mathsf{fuse}}$	–	1	2	4	8	16
threshold	0.05	0.05	0.05	0.05	0.05	0.05
solved	161	161	161	161	161	160

$\rho^{\mathsf{proof}}_{\mathsf{fuse}}$	–	1	2	**4**	8	16
threshold	0.005	0.2	0.2	**0.2**	0.2	0.2
solved	111	164	163	**165**	162	164

- Best $\mathcal{P}_{\mathsf{fuse}}$ model solves 171

Parental Guidance

Table 6. The best threshold for each tested reduction ratio of $\mathcal{P}_{\mathsf{cat}}$.

ρ_{cat}	–	1	2	4	**8**	16
threshold	0.5	0.1	0.05	0.3	**0.1**	0.05
solved	117	168	170	168	**173**	169

Parental Guidance

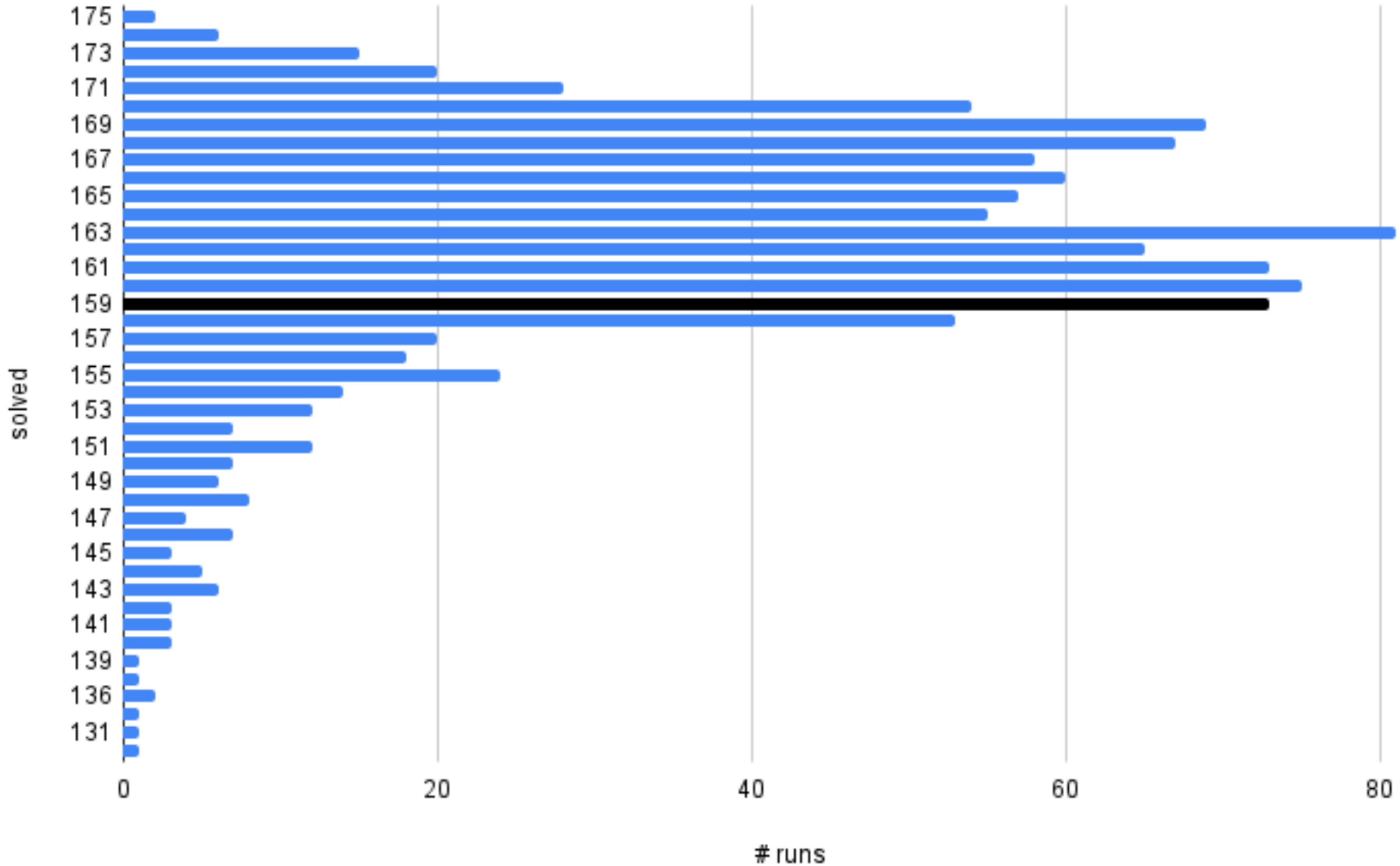


Fig. 1. The number of settings (and runs) corresponding to each number of solutions for the $\mathcal{P}_{\mathsf{cat}}$ grid search. The black bar is 159, the number of problems solved by $\mathcal{D}_{\mathsf{large}}$.

Parental Guidance

Table 8. Final 30s evaluation on small trains (T), development (D), and holdout (H) compared with $\mathcal{D}_{\mathsf{large}}$.

model	threshold	solved (T)	solved (D)	solved (H)
$\mathcal{D}_{\mathsf{large}}$	-	3269	1397	1390
$\mathcal{P}^{\mathsf{given}}_{\mathsf{fuse}}+\mathcal{D}_{\mathsf{large}}$	0.05	3302 (+1.0%)	1411 (+1.0%)	1417 (+1.9%)
$\mathcal{P}^{\mathsf{proof}}_{\mathsf{fuse}}+\mathcal{D}_{\mathsf{large}}$	0.1	3389 (+3.7%)	1489 (+6.6%)	1486 (+6.9%)
$\mathcal{P}_{\mathsf{cat}}+\mathcal{D}_{\mathsf{large}}$	**0.05**	**3452 (+5.6%)**	**1571 (+12.4%)**	**1553 (+11.7%)**

- The model has 100 trees of depth 60 with 8192 leaves.

3-phase ENIGMA

- 2-phase ENIGMA with parental guidance.
- Train a $\mathcal{P}_{\mathsf{cat}}$ model on $\mathcal{G}_{\mathsf{large}}$ data.
- Parental guidance only with GNN:
  - **1621** problems on development
  - **1623** problems on holdout**.**

3-phase ENIGMA

- 2-phase ENIGMA with parental guidance.
- Train a $\mathcal{P}_{\mathsf{cat}}$ model on $\mathcal{G}_{\mathsf{large}}$ data.
- Parental guidance only with GNN:
  - 1621 problems on development in 30s
  - 1623 problems on holdout**.**
- 3-phase**:**
  - **1631** problems on development.
  - **1632** problems on holdout (+17% over $\mathcal{D}_{\mathsf{large}}$).

Conclusion

- A GPU server renders the GNN more usable
- Combining fast and slow models works (+7%)
- Parent guided clause generation works (+11%)
- Combining all three models works even better (despite being ‘slower’) (+17%)

Conclusion

- A GPU server renders the GNN more usable
- Combining fast and slow models works (+7%)
- Parent guided clause generation works (+11%)
- Combining all three models works even better (despite being ‘slower’) (+17%)

- What other models can we integrate into E?