Vertical Reasoning Enhanced Learning, Generation and Scientific Discovery

Yexiang Xue

Department of Computer Science

Purdue University

yexiang@purdue.edu

Intelligent Systems Integrate Learning and Reasoning

Perception

Learning

Knowledge

Reaction

Reasoning

Machine learning:

- **Bottom-up:** Learn predictive models from data
- **Challenging** in providing formal guarantees
- May **violate constraints** in rare and unseen situations

Automated reasoning:

- **Top-down:** Build models from problem description
- **Rigid models**: problem formulation must be agreed a-priori
- **Difficult to adapt** to data distributions

Data

Action

Intelligent Systems Integrate Learning and Reasoning

Perception

Learning

Knowledge

Reaction

Reasoning

Machine learning:

Automated reasoning:

Challenge for next generation AI:

How to **integrate** learning with reasoning

odels

iption

lem

- May **violate constraints** in rare and unseen situations

- **Difficult to adapt** to data distributions

Data

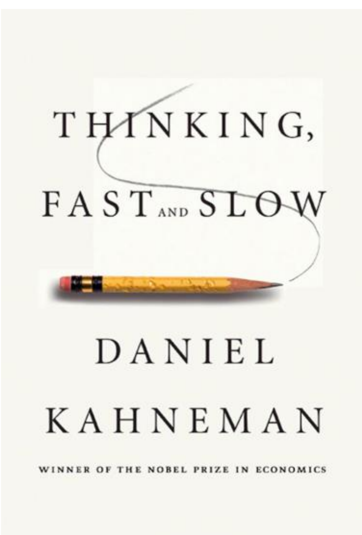
Action

Generalist Systems; Think Fast and Slow

Input Specifications:
- Add a blue microwave right of the oven
- Add a green toaster left of the oven and below the sink

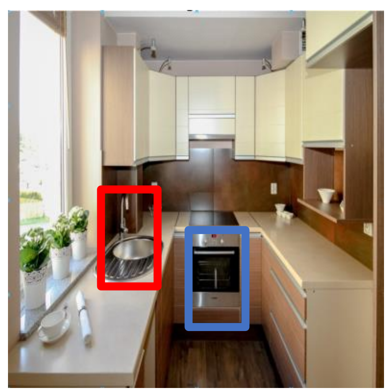
Reasoning *&* ***learning*** *are in charge of different cognitive systems.*

Need both for building a generalist AI.



Learning

System 1 perception (fast thinking)



Reasoning
+ Learning

System 2 planning & generation (slow thinking)

Integrate Reasoning into Design Generation

Existing Kitchen Env:

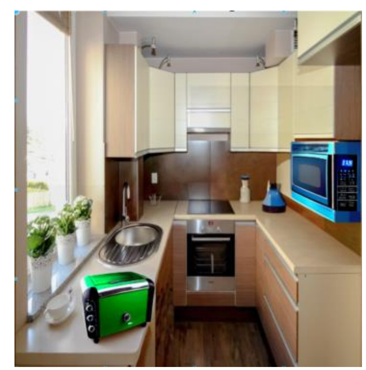
Input Specifications:

- Add a blue microwave right of the oven
- Add a green toaster left of the oven and below the sink

(stated in propositional logic)

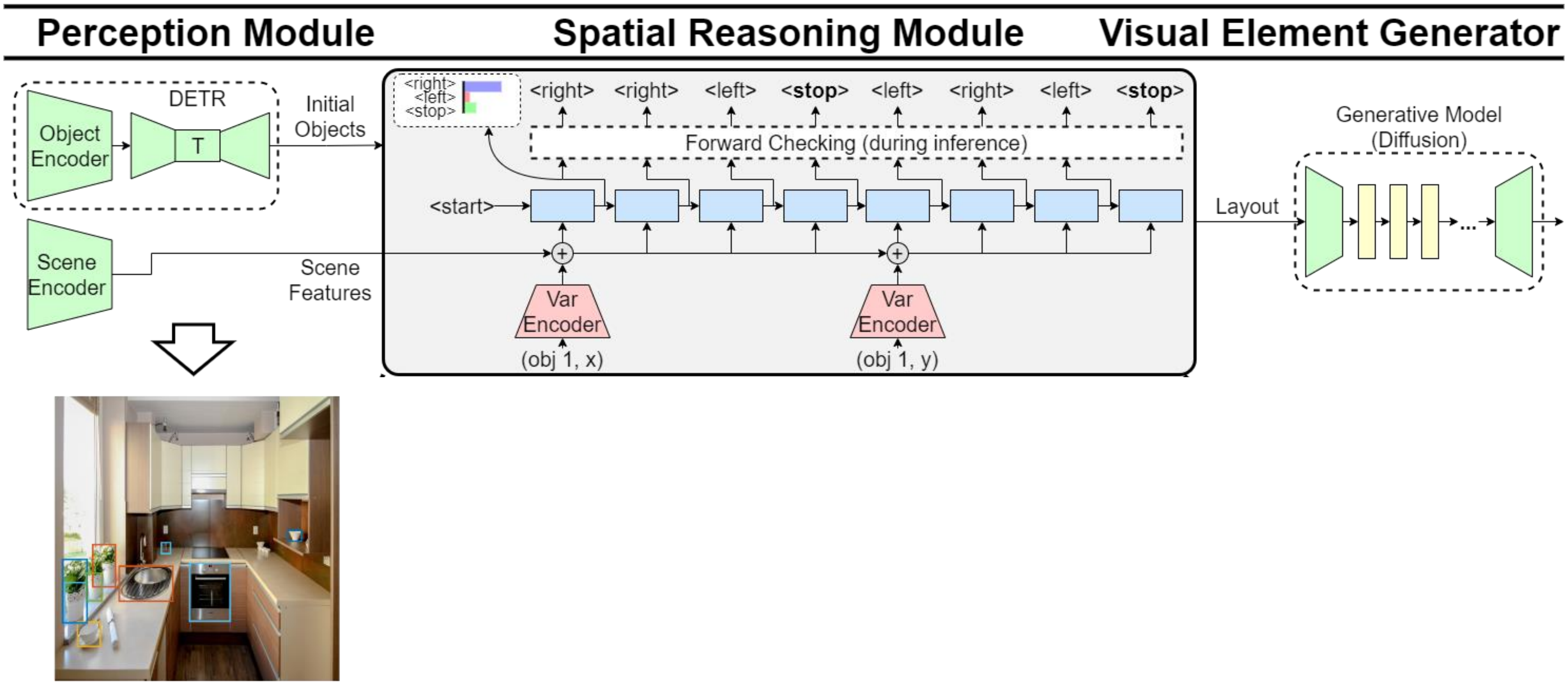
- Good designs need to meet industry standards and user needs, while capturing subtle aspects such as aesthetics and convenience.
- **Complete constraint reasoning approach**: satisfy design specifications, but cannot capture visual information. In fact, such info cannot be encoded in objective functions.
- **Complete ML approach**: generate beautiful designs, but cannot meet specifications.

Baseline (Stable Diffusion)

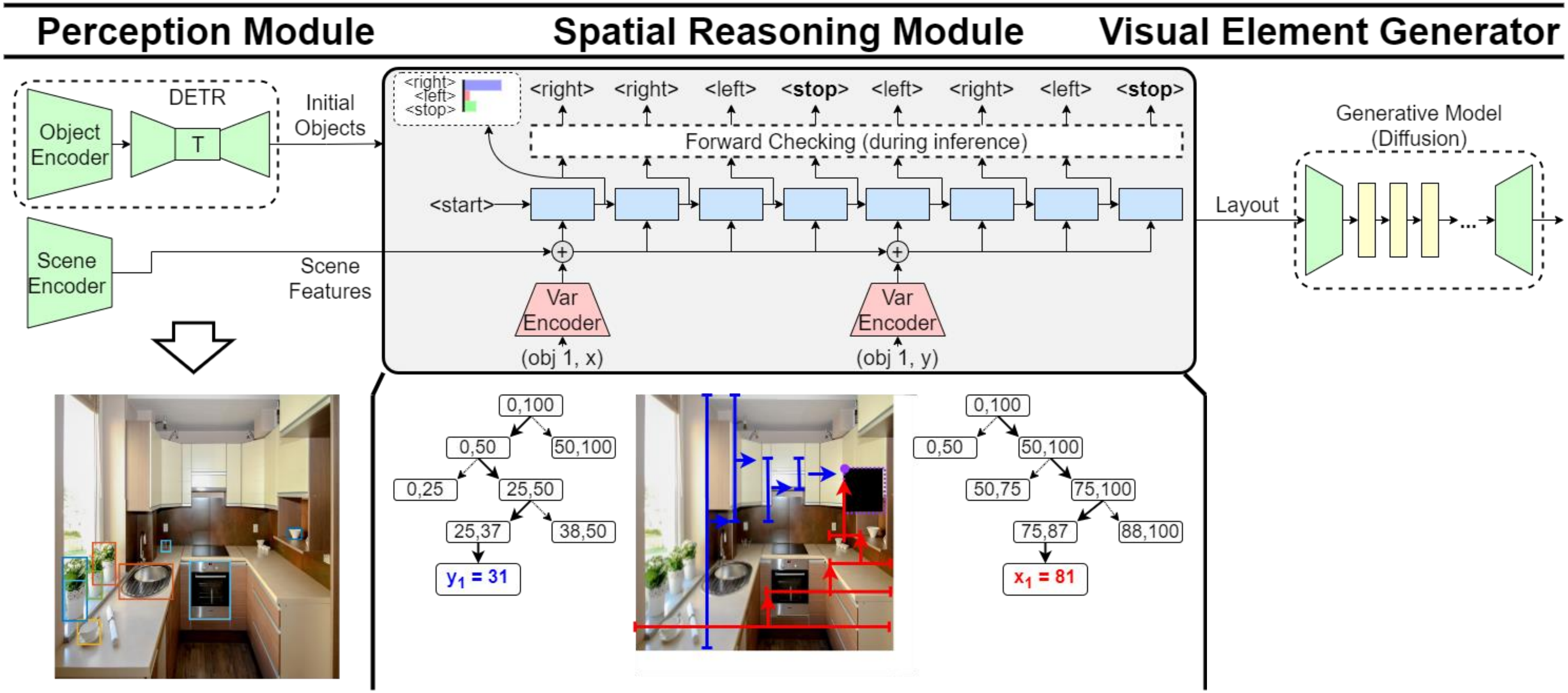


Ours (CORE)

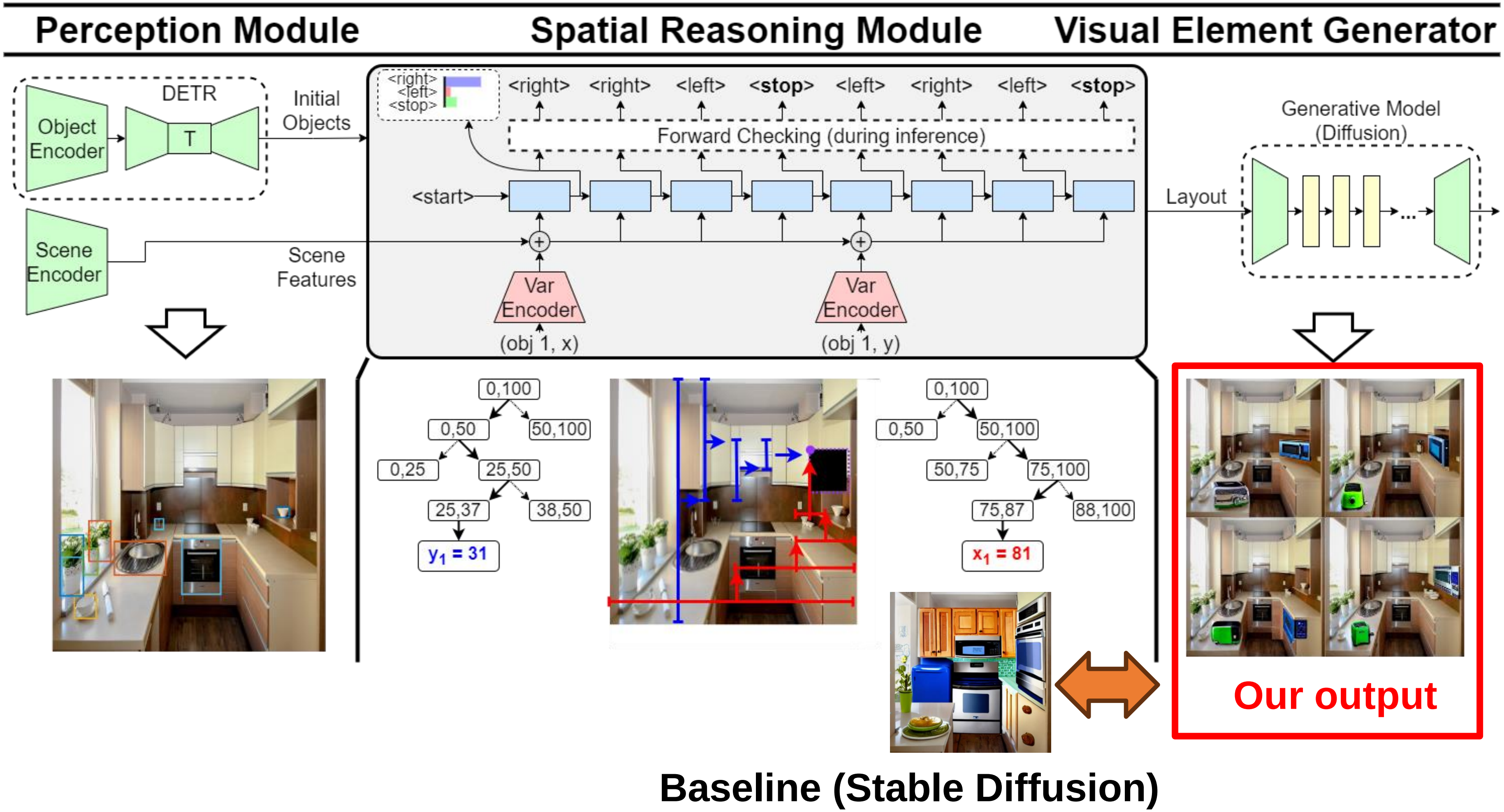
CORE Applied to Design Generation



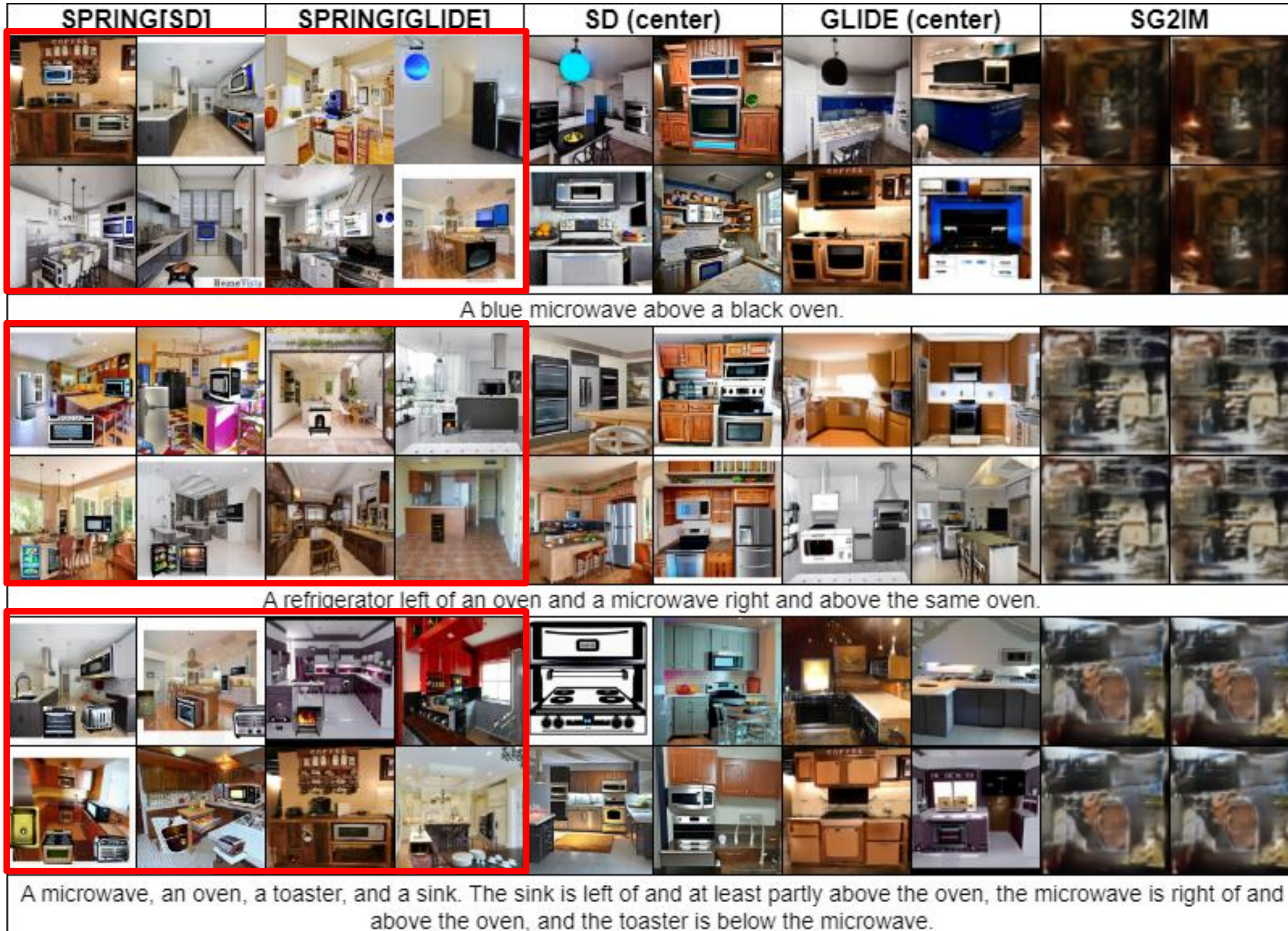
CORE Applied to Design Generation



CORE Applied to Design Generation

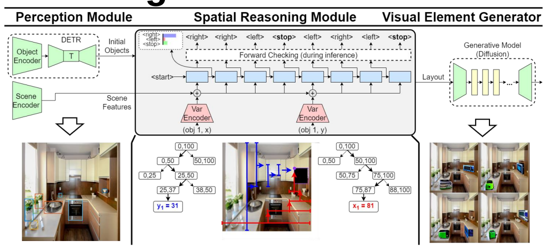


CORE for Design Generation

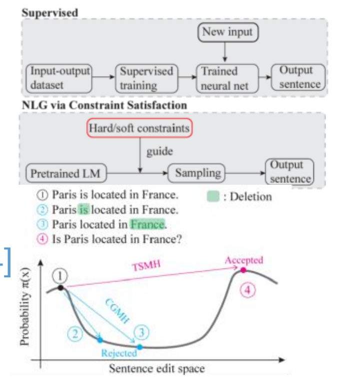


Fruitful Expedition on Integrating Reasoning with Learning

Design Generation



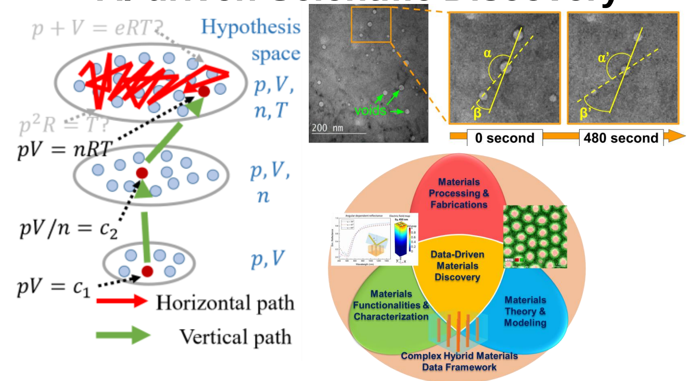
Language Generation



[IJCAI-19, Preprint-24]

[NeurIPS-18, WWW-22 EMNLP-20, JMLR-22]

AI-driven Scientific Discovery



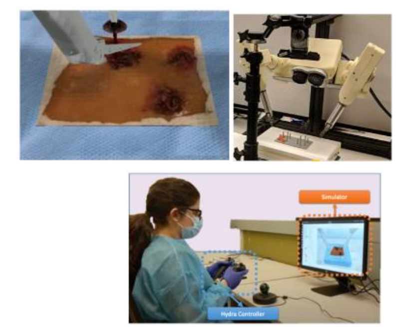
[MRS Comm-19, NeurIPS-21, ECML-PKDD-21,23 UAI-22, IAAI-24, AAAI-24, ICASSP-19, JOM-20, JNM-22]

Operational Research

[CPAIOR-19, UAI-21, JMLR-22]

Automated Reasoning ⟺ **Machine Learning**

Robotic Surgery

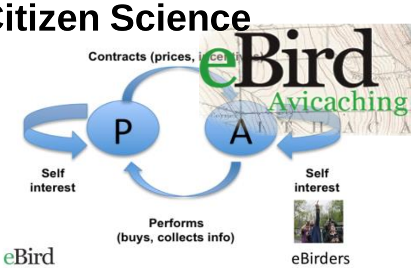


[IROS-19, ICRA-21, Roman-21, MHSRS-20,22, journal-20, IEEE Trans-23 MilMed-23]

Computational Sustainability

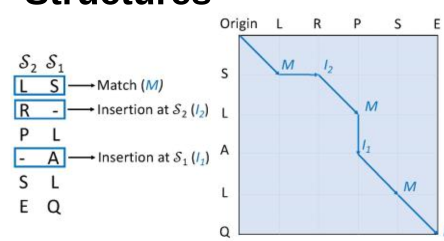
[Nature Comm-19, CACM-19, Science-22]

Leader-follower Games Citizen Science



[AAMAS-23, UAI-22, ICMLA-22, SMC-22]

Learn Combinatorial Structures



[AAAI-23,24]

[ICML-21, ECAI-20, UAI-18, 21, PGM-20, Brief-Bioinfo-22]

Content

- Introduction
- Vertical Reasoning Enhanced Neural Generation
- Vertical Reasoning Driven Scientific Discovery
- Vertical Reasoning Solving Satisfiability Modulo Counting (SMC) Integrating Symbolic & Statistical AI with Provable Guarantees
- Conclusion

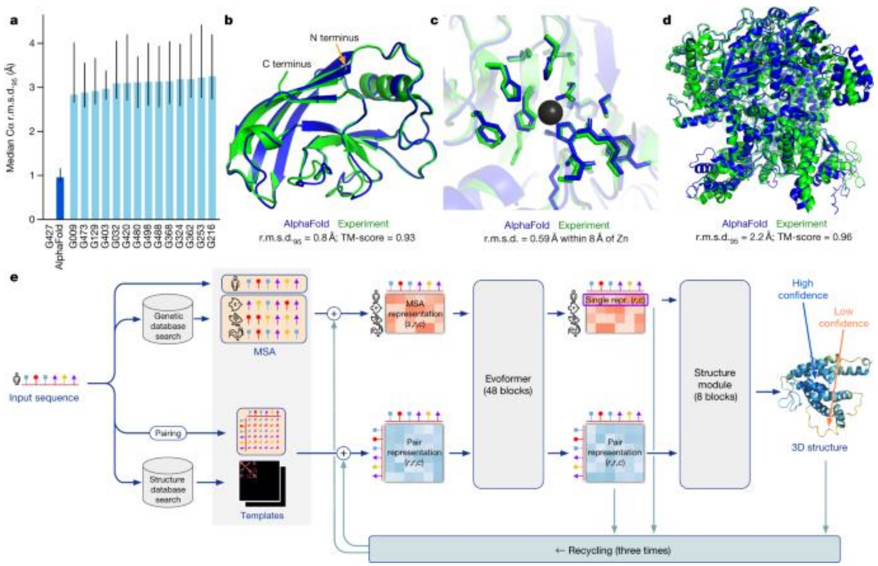
Content

- Introduction ✓
- Vertical Reasoning Enhanced Neural Generation ✓
- Vertical Reasoning Driven Scientific Discovery
- Vertical Reasoning Solving Satisfiability Modulo Counting (SMC) Integrating Symbolic & Statistical AI with Provable Guarantees
- Conclusion

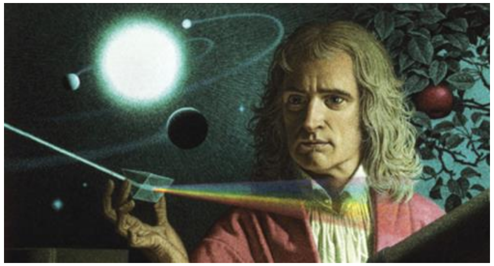
AI-driven Scientific Discovery (Learning) Needs Reasoning

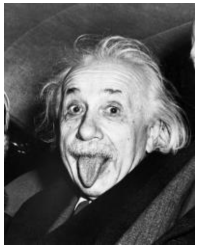
- Exciting progress in deep learning

Especially in science domains [AlphaFold]



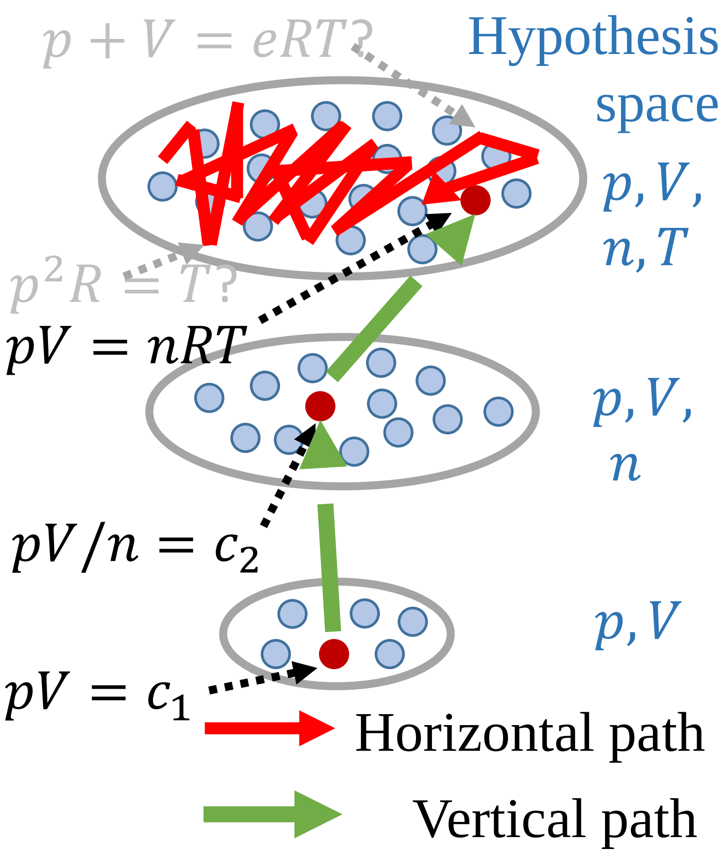
- Human learning (discovery) is **better**!





  - Active exploration with a purpose
  - Learn from an incredibly small set of “surprising” samples
  - Interpretable, elegant models & equations

Vertical vs. Horizontal Discovery



- What can AI learn from human scientists?
- Symbolic regression: learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- State-of-the-art solvers follow horizontal paths
  - Can be challenging because of the exponential size of the hypothesis space.
- We propose: ***vertical paths*** – also scientists' approach!
  - Search in reduced spaces are much easier!
  - Can supercharge AI-driven scientific discovery.

Symbolic regression

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
3.0	-1.0	4.0	1
1.6	3.5	5.2	10.8
1.8	1.0	3.2	5
7.1	8.6	3.8	64.9
1.7	1.0	2.3	4
2.5	2.6	3.1	9.6
8.9	1.1	2.0	11.8
4.2	-1.0	2.2	-2
5.8	1.0	7.2	13
1.6	5.7	1.2	10.3
9.7	-1.0	1.7	-8

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?

Symbolic regression

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
1.8	1.0	3.2	5
1.7	1.0	2.3	4
5.8	1.0	7.2	13

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?
- How about if I only ask you to look into these rows?

$$y = x_1 + x_3?$$

Symbolic regression

X_1	X_2	X_3	Y
3.0	-1.0	4.0	1
4.2	-1.0	2.2	-2
9.7	-1.0	1.7	-8

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?
- How about if I only ask you to look into these rows?

$$y = x_1 + x_3?$$

- How about these rows?

$$y = -x_1 + x_3?$$

Symbolic regression

Red and blue data are two control variable experiment trials in reduced hypothesis spaces (X_2 controlled)!
Vertical discovery ***simplify*** symbolic regression!

X_1	X_2	X_3	Y
2.5	1.0	9.5	12
3.0	-1.0	4.0	1
1.8	1.0	3.2	5
1.7	1.0	2.3	4
4.2	-1.0	2.2	-2
5.8	1.0	7.2	13
9.7	-1.0	1.7	-8

- Learning a symbolic expression from data
  - A good benchmark mimicking scientific discovery process.
- Incredibly difficult because of the large search space of all possible expressions.

- Can you guess which equation $y = f(x_1, x_2, x_3)$ generates the data shown in the left table?
- How about if I only ask you to look into these rows?

$$y = x_1 + x_3?$$

- How about these rows?

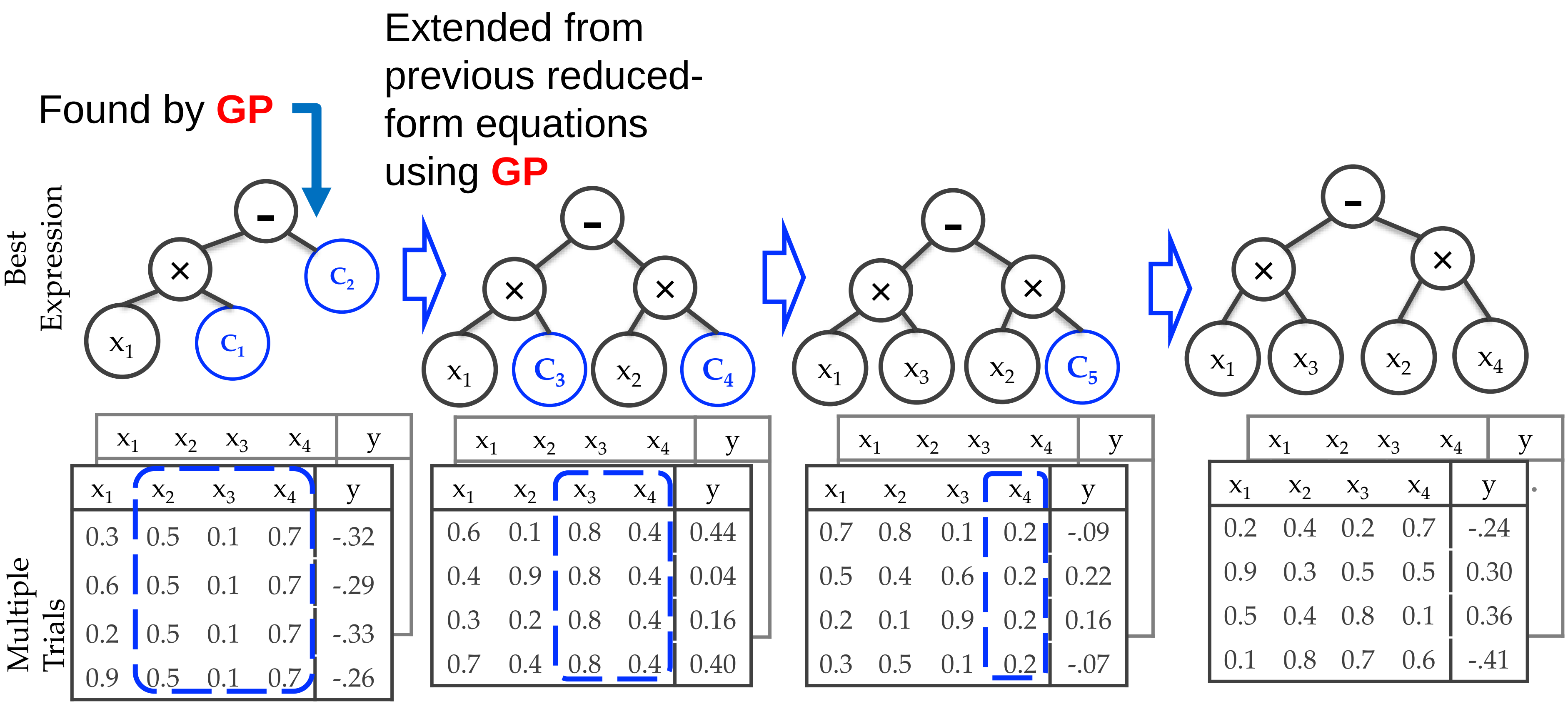
$$y = -x_1 + x_3?$$

- Maybe the equation is:

$$y = x_2 x_1 + x_3?$$

INDEED!

Control Variable Genetic Programming (CVGP)



x_1	x_2	x_3	x_4	y
0.3	0.5	0.1	0.7	-.32
0.6	0.5	0.1	0.7	-.29
0.2	0.5	0.1	0.7	-.33
0.9	0.5	0.1	0.7	-.26

(a) Control x_2, x_3, x_4

x_1	x_2	x_3	x_4	y
0.6	0.1	0.8	0.4	0.44
0.4	0.9	0.8	0.4	0.04
0.3	0.2	0.8	0.4	0.16
0.7	0.4	0.8	0.4	0.40

(b) Control x_3, x_4

x_1	x_2	x_3	x_4	y
0.7	0.8	0.1	0.2	-.09
0.5	0.4	0.6	0.2	0.22
0.2	0.1	0.9	0.2	0.16
0.3	0.5	0.1	0.2	-.07

(c) Control x_4

x_1	x_2	x_3	x_4	y
0.2	0.4	0.2	0.7	-.24
0.9	0.3	0.5	0.5	0.30
0.5	0.4	0.8	0.1	0.36
0.1	0.8	0.7	0.6	-.41

(d) No control

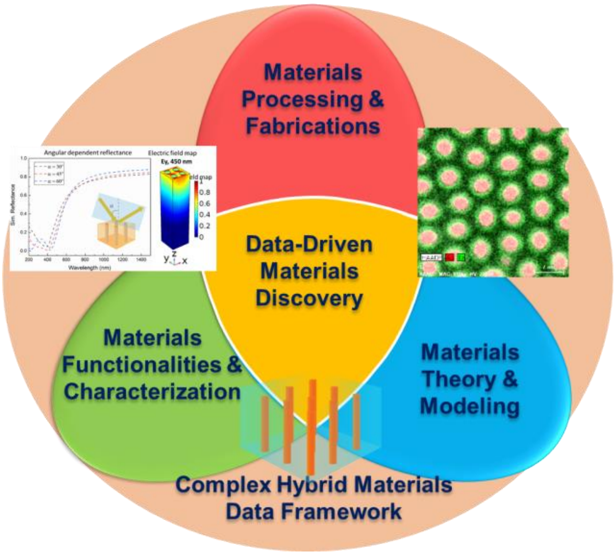
Experiment Results

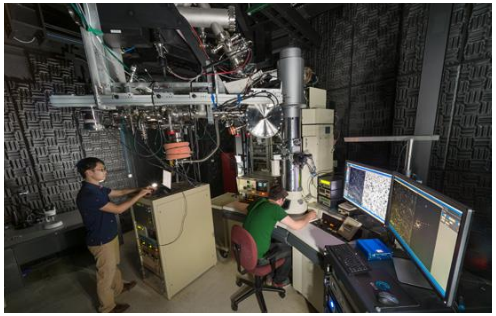
Ops	Dataset configs	CVGP (ours)		GP		DSR		PQT		VPG		GPMeld	
		50%	75%	50%	75%	50%	75%	50%	75%	50%	75%	50%	75%
	(2,1,1)	0.198	0.490	**0.024**	**0.053**	0.032	3.048	0.029	0.953	0.041	0.678	0.387	22.806
	(4,4,6)	**0.036**	**0.088**	0.038	0.108	1.163	3.714	1.016	1.122	1.087	1.275	1.058	1.374
	(5,5,5)	0.076	0.126	**0.075**	**0.102**	1.028	2.270	1.983	4.637	1.075	2.811	1.479	2.855
`inv`	(5,5,8)	**0.061**	**0.118**	0.121	0.186	1.004	1.013	1.005	1.006	1.002	1.009	1.108	2.399
	(6,6,8)	**0.098**	**0.144**	0.104	0.167	1.006	1.027	1.006	1.020	1.009	1.066	1.035	2.671
	(6,6,10)	**0.055**	**0.097**	0.074	0.132	1.003	1.009	1.005	1.008	1.004	1.015	1.021	1.126
	(3,2,2)	**0.098**	**0.165**	0.108	0.425	0.350	0.713	0.351	1.831	0.439	0.581	0.102	0.597
	(4,4,6)	**0.078**	**0.121**	0.120	0.305	7.056	16.321	5.093	19.429	2.458	13.762	2.225	3.754
sin,	(5,5,5)	**0.067**	**0.230**	0.091	0.313	32.45	234.31	36.797	229.529	14.435	46.191	28.440	421.63
cos	(5,5,8)	**0.113**	**0.207**	0.119	0.388	195.22	573.33	449.83	565.69	206.06	629.41	363.79	666.57
	(6,6,8)	**0.170**	**0.481**	0.186	0.727	1.752	3.824	4.887	15.248	2.396	7.051	1.478	6.271
	(6,6,10)	**0.161**	**0.251**	0.312	0.342	11.678	26.941	5.667	24.042	7.398	25.156	11.513	28.439
	(3,2,2)	0.049	**0.113**	**0.023**	0.166	0.663	2.773	1.002	1.992	0.969	1.310	0.413	2.510
	(4,4,6)	**0.141**	**0.220**	0.238	0.662	1.031	1.051	1.297	1.463	1.051	1.774	1.093	1.769
sin,	(5,5,5)	**0.157**	0.438	0.195	**0.337**	1.098	3.617	1.018	5.296	1.012	1.27	1.036	3.617
cos,	(5,5,8)	**0.122**	**0.153**	0.166	0.186	1.009	1.103	1.017	1.429	1.007	1.132	1.07	2.904
`inv`	(6,6,8)	**0.209**	**0.590**	**0.209**	0.646	1.003	1.153	1.047	1.134	1.059	1.302	1.029	3.365
	(6,6,10)	0.139	0.232	**0.073**	**0.159**	1.654	3.408	1.027	1.069	1.009	1.654	1.445	2.106

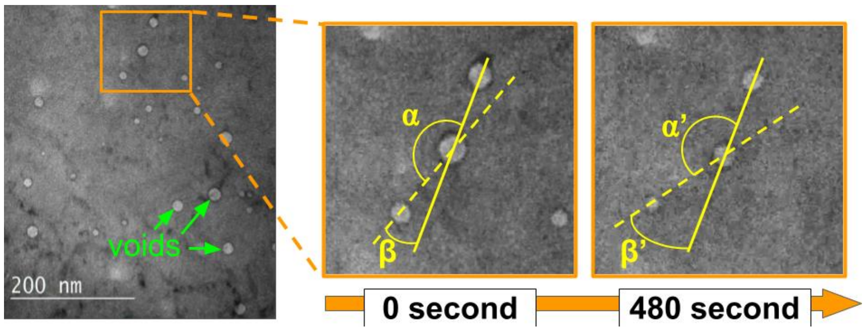
Median (50%) and 75%-quantile NMSE values of the symbolic expressions found by all the algorithms on several noisy benchmark datasets. Our CVGP finds symbolic expressions with the smallest NMSEs.

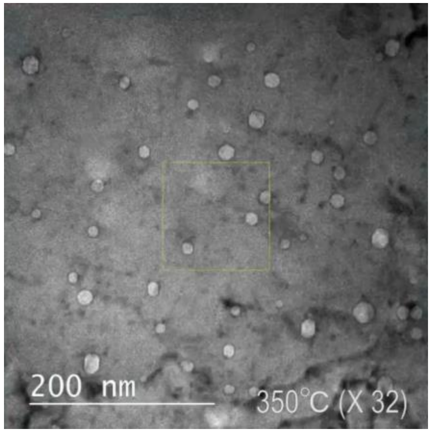
AI Driven Materials Discovery in Extreme Conditions

- Search for strong materials under heavy irradiation and extremely high temperature
- Understand defect formation, migration in extreme conditions
- Better materials for future nuclear reactors
- In-situ experimentation

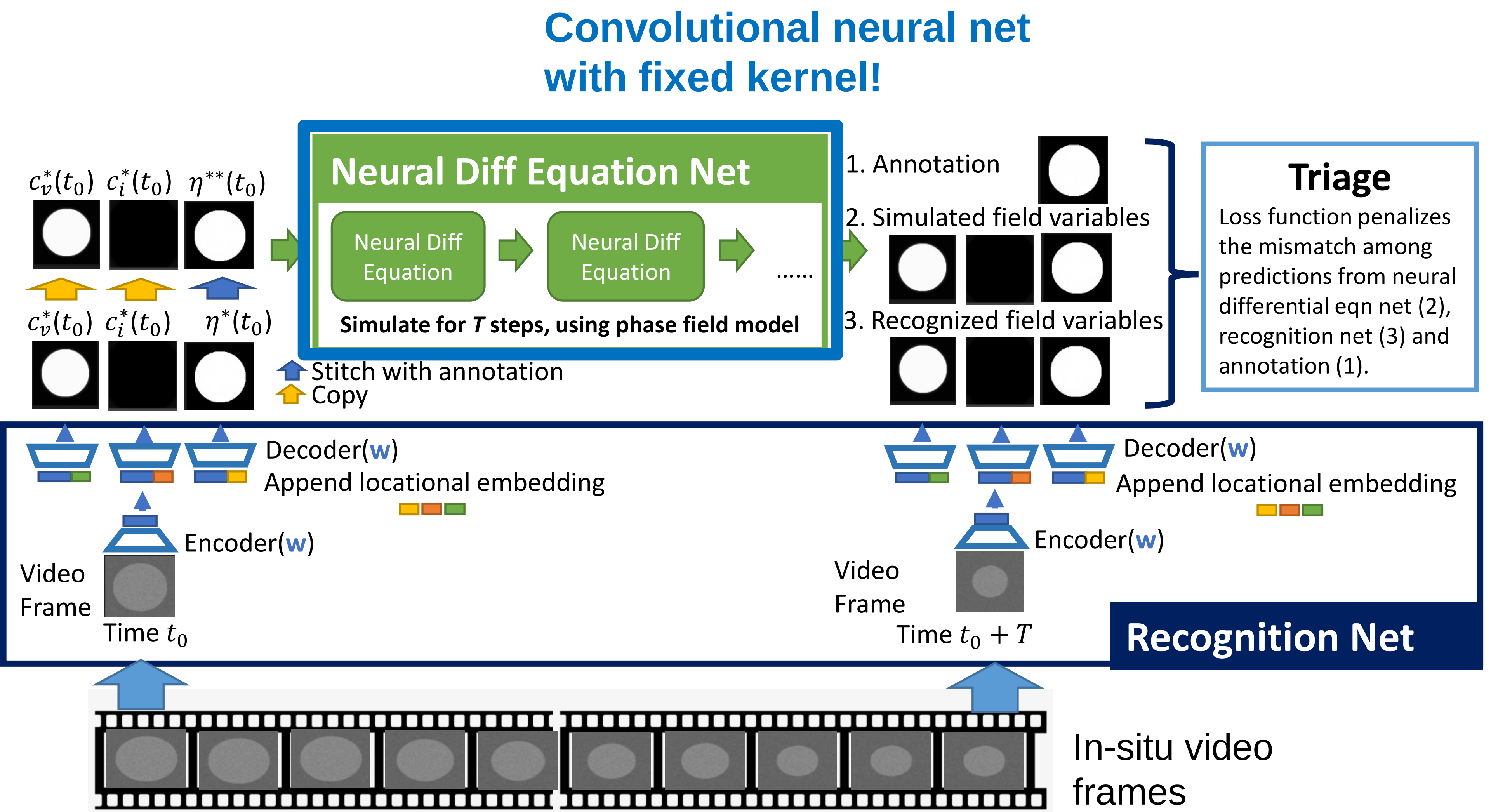






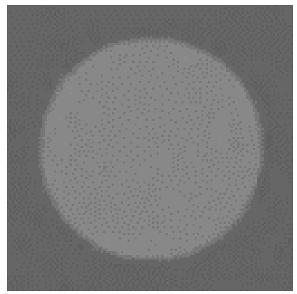


NeuraDiff: High Level Idea

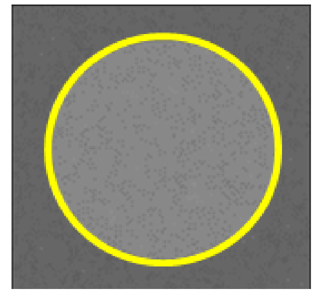


Track Nanovoids + Learn Phase Field Model

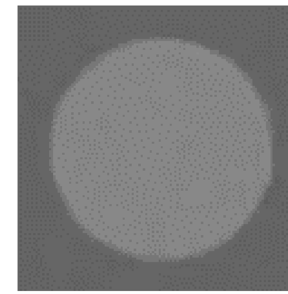
TEM Video



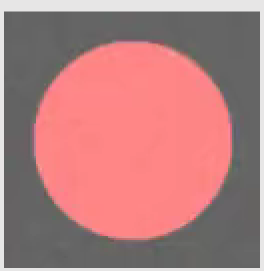
Partial Annotation
(every 10th frame)



NN cannot predict
future well



Nanovoid
Tracking



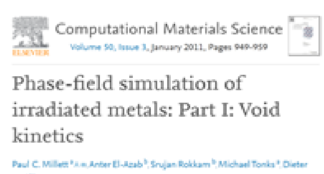
Identify phase field model parameters

$$F = N\int\left[h(\eta)f^{s}(c_v,c_i) + j(\eta)f^{v}(c_v,c_i) + \frac{\kappa_v}{2}|\nabla c_v|^2 + \frac{\kappa_i}{2}|\nabla c_i|^2 + \frac{\kappa_\eta}{2}|\nabla\eta|^2\right]dV,$$

$$\frac{\partial c_v}{\partial t} = \nabla\cdot\left(M_v\nabla\frac{1}{N}\frac{\delta F}{\delta c_v}\right) + \xi + P_v - R_{iv},$$

$$\frac{\partial c_i}{\partial t} = \nabla\cdot\left(M_i\nabla\frac{1}{N}\frac{\delta F}{\delta c_i}\right) + \xi + P_i - R_{iv},$$

$$\frac{\partial \eta}{\partial t} = -L\frac{\delta F}{\delta \eta} + \xi + P_{vi}.$$

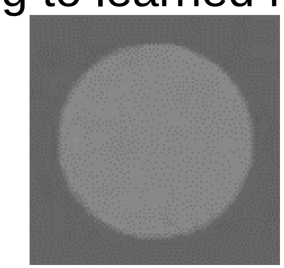


Computational Materials Science
Volume 50, Issue 3, January 2011, Pages 949-959

Phase-field simulation of irradiated metals: Part I: Void kinetics

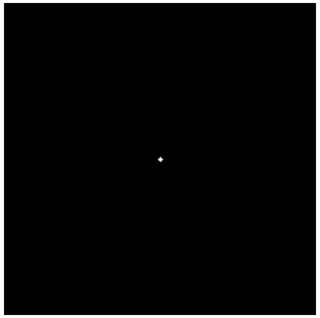
Paul C. Millett, Anter El-Azab, Srujan Rokkam, Michael Tonks, Dieter Wolf

Simulate void evolution
according to learned model



Learning models for dendritic solidification

Ground-truth ϕ



Phase-field model:

$$F(\phi, m) = \int \left(\frac{1}{2}\epsilon^2 |\nabla\phi|^2 + f(\phi, m)\right) dv,$$

$$f(\phi, m) = \frac{1}{4}\phi^4 - \left(\frac{1}{2} - \frac{1}{3}m\right)\phi^3 + \left(\frac{1}{4} - \frac{1}{2}m\right)\phi^2,$$

$$\epsilon = \bar{\epsilon}\sigma(\theta),$$

$$\sigma(\theta) = 1 + \delta\cos(j(\theta - \theta_0)),$$

$$\theta = \tan^{-1}\left(\frac{\partial\phi/\partial y}{\partial\phi/\partial x}\right),$$

$$m(T) = (\alpha/\pi)\tan^{-1}[\gamma(T_{eq} - T)],$$

Dendritic growth follows Allen-Cahn equation:

$$\tau\frac{\partial\phi}{\partial t} = -\frac{\delta F}{\delta\phi}$$

Temperature follows conservation law:

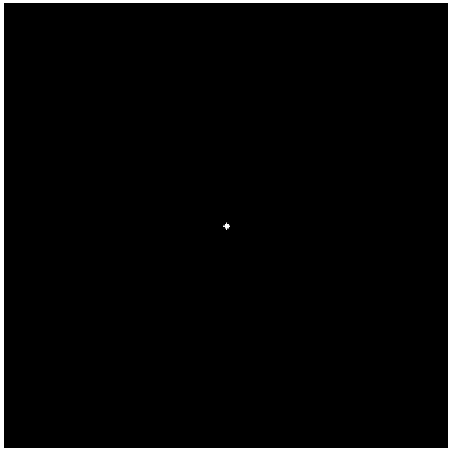
$$\frac{\partial T}{\partial t} = \nabla^2 T + \kappa\frac{\partial\phi}{\partial t}$$

Vertical learning experiment

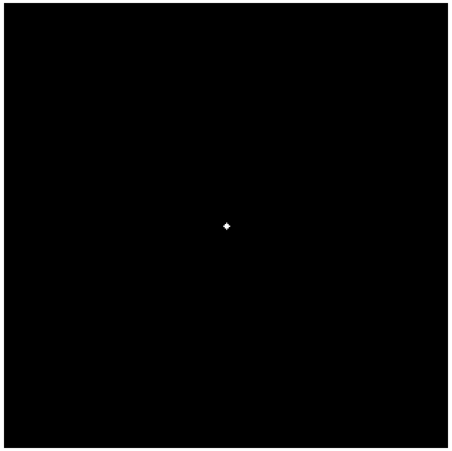
- Intentionally first learn on data in which $\nabla\phi = 0$;
- In this case, blue parameters do not affect dynamics;
- Focus on learning red parameters.

- Allow $\nabla\phi$ to vary in the second stage, hence start to learn blue parameters.

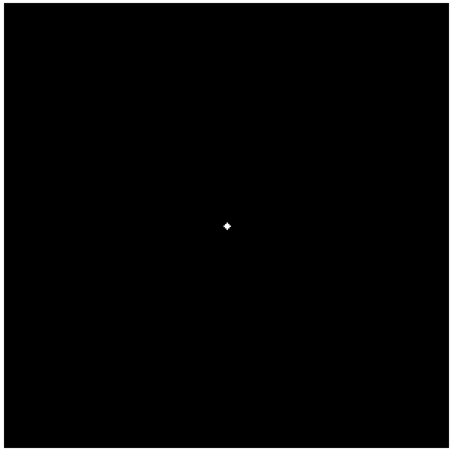
Comparison



Ground-truth ϕ



Learning all parameters at once



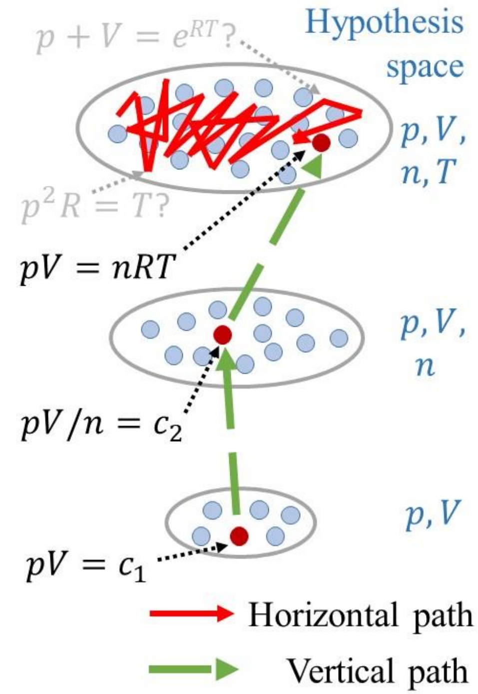
Vertical learning experiments

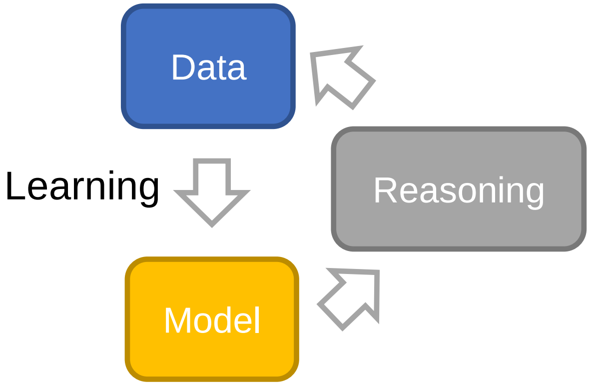
Conclusion

- Vertical symbolic regression
  - Incrementally build complex equations from simple ones using genetic programming
  - Learning from control variable experiments

- Vertical scientific discovery -- learning PDEs from data

- Look into future: integrate active reasoning into learning
  - Science progress resulted from insightful experiment design, courageous hypothesis forming (reasoning) + high-capacity modeling (learning)



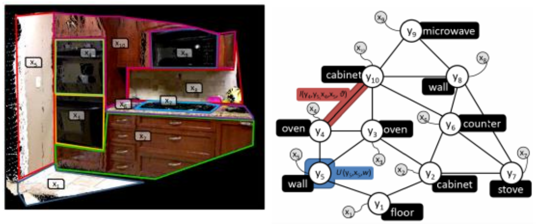


Content

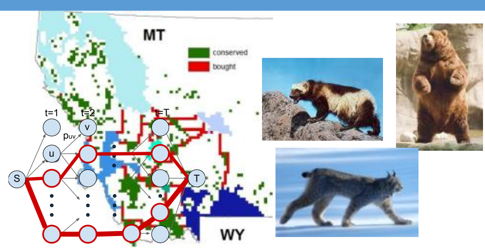
- Introduction ✓
- Vertical Reasoning Enhanced Neural Generation ✓
- Vertical Reasoning Driven Scientific Discovery ✓
- Vertical Reasoning Solving Satisfiability Modulo Counting (SMC) Integrating Symbolic & Statistical AI with Provable Guarantees
- Conclusion

Vertical Reasoning to Solve Satisfiability Modulo Counting (SMC)

Provable likelihood maximization for Markov Random Fields



Stochastic optimization (Network design as an example)



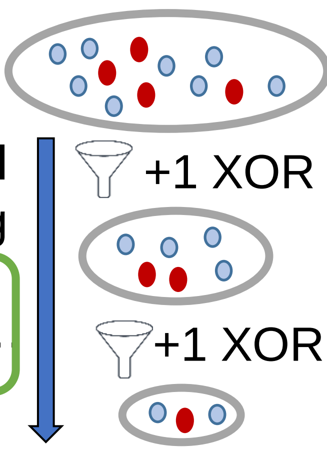
Solving quantal response leader-follower games

All these problems ***integrating symbolic and statistical AI*** are SMC. SMC connects model counting predicates with ∨,∧, ¬, e.g.:

$$\left(\sum_y f_1(x,y) \geq 2^{q_1}\right) \wedge \left(\neg\left(\sum_y f_2(x,y) \geq 2^{q_2}\right) \vee \left(\sum_y f_3(x,y) \geq 2^{q_3}\right)\right)$$

Constant approximation guarantee to solve SMC based on vertical reasoning streamlining XOR constraints.

Vertical reasoning



Initial problem

$$\phi(x,b) \wedge \left(b_1 \Rightarrow \boxed{\sum_{y_1 \in Y_1} f_1(x,y_1) \geq 2^{q_1}}\right) \wedge .$$

XOR-SMC

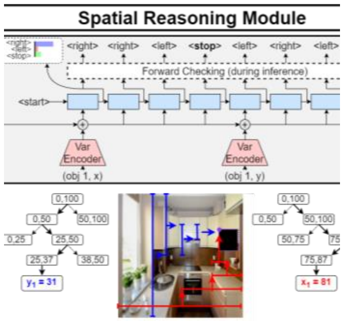
$$\phi(x,b) \wedge \left(b_1 \Rightarrow \boxed{f_1(x,y) \ldots \wedge XOR_{q_1}(y)}\right) \wedge \cdots$$

Content

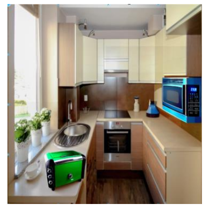
- Introduction ✓
- Vertical Reasoning Enhanced Neural Generation ✓
- Vertical Reasoning Driven Scientific Discovery ✓
- Vertical Reasoning Solving Satisfiability Modulo Counting (SMC) Integrating Symbolic & Statistical AI with Provable Guarantees ✓
- Conclusion

Embedding Reasoning for Learning

Vertical Reasoning Enhanced Neural Generation

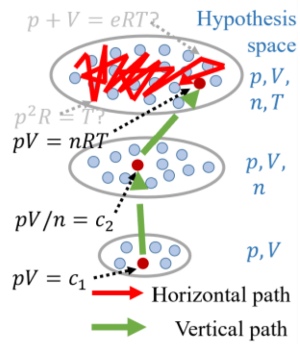


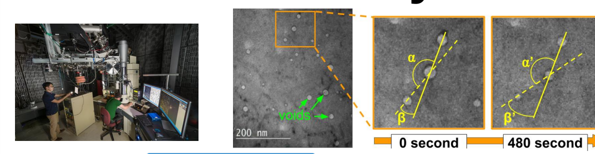
Baseline

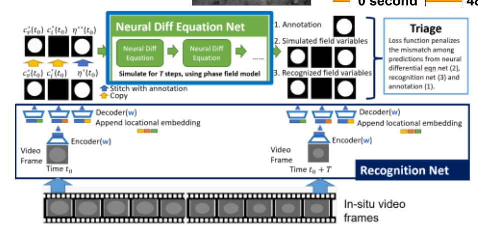


Ours

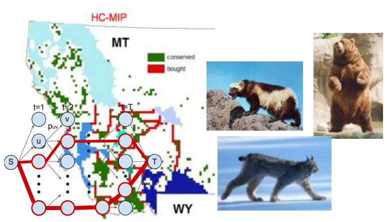
Vertical Reasoning Driven Scientific Discovery

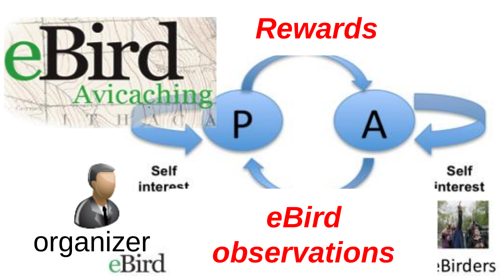






Vertical Reasoning Solving Satisfiability Modulo Counting with Guarantees





Constant approximation guarantee to solve SMC ***integrating symbolic and statistical AI***

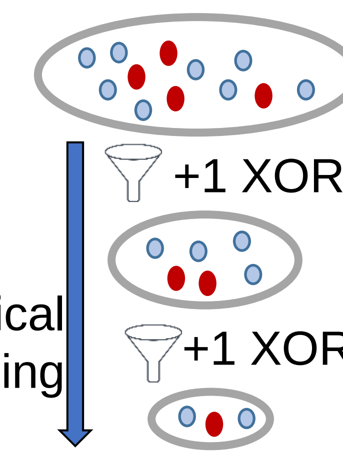
Initial problem

$$\phi(x,b) \wedge \left(b_1 \Rightarrow \sum_{y_1 \in Y_1} f_1(x, y_1) \geq 2^{q_1} \right) \wedge .$$

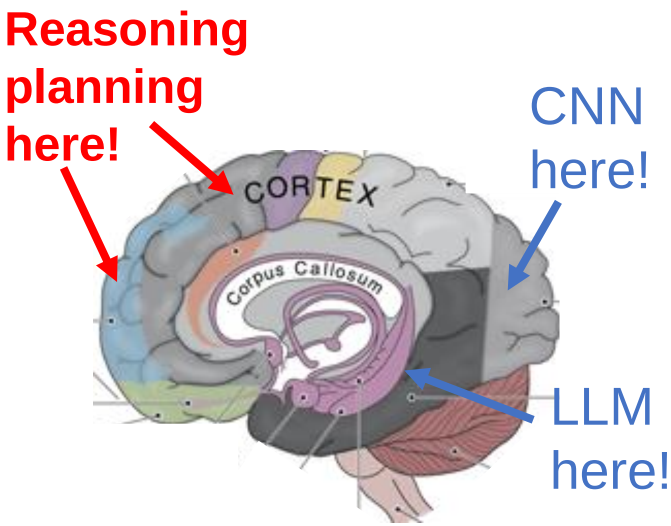
XOR-SMC

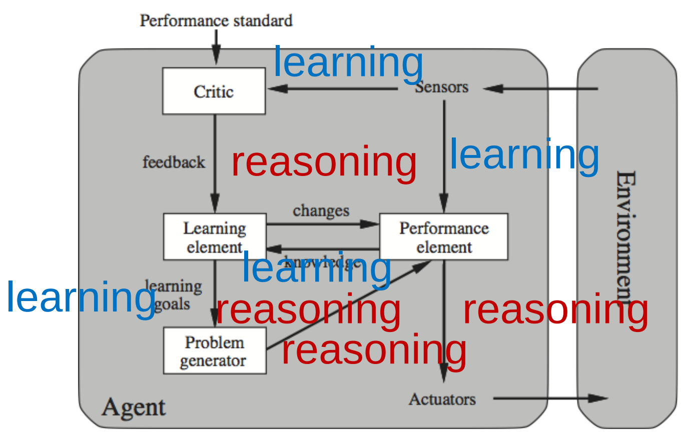
$$\phi(x,b) \wedge \left(b_1 \Rightarrow f_1(x, y) \ldots \wedge XOR_{q_1}(y) \right) \wedge \cdots$$

Vertical reasoning



Conclusion

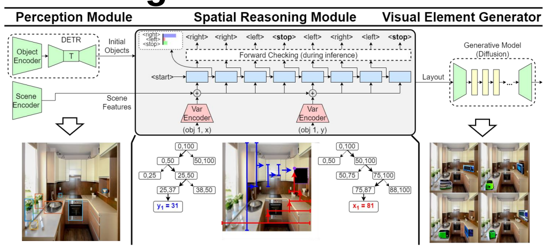




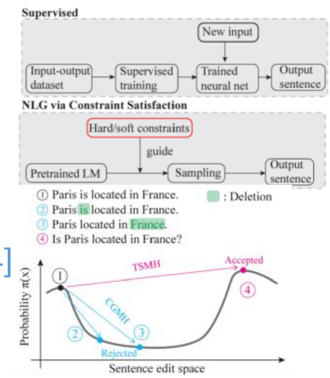
- AI agents (human brains) are integrated systems.
- "Reasoning + Learning" multiplies power than them alone.
- "LLM interfacing coding, web, …" is a good start.
- **Deep** integration offers way more:
  - Reasoning generates designs satisfying user specifications
  - Reasoning expedites learning in scientific discovery
  - Reasoning solves SMC with constant approximation guarantees
- Much more to come, very exciting so far, very busy years ahead.

Fruitful Expedition on Integrating Reasoning with Learning

Design Generation



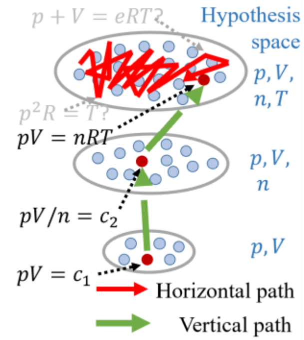
Language Generation

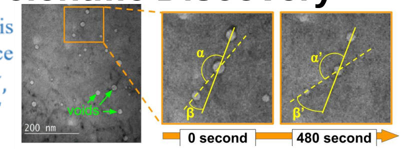


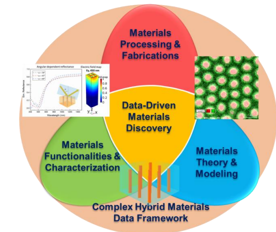
[IJCAI-19, Preprint-24]

[NeurIPS-18, WWW-22 EMNLP-20, JMLR-22]

AI-driven Scientific Discovery

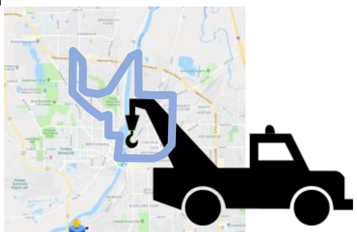






[MRS Comm-19, NeurIPS-21, ECML-PKDD-21,23 UAI-22, IAAI-24, AAAI-24, ICASSP-19, JOM-20, JNM-22]

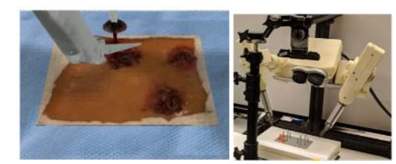
Operational Research



[CPAIOR-19, UAI-21, JMLR-22]

Automated Reasoning ⟺ **Machine Learning**

Robotic Surgery

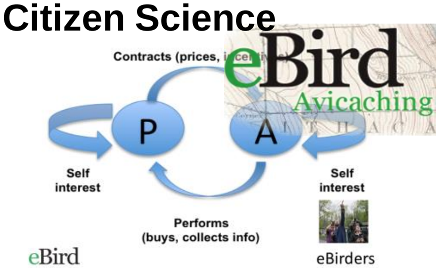


[IROS-19, ICRA-21, Roman-21, MHSRS-20,22, journal-20, IEEE Trans-23 MilMed-23]

Computational Sustainability

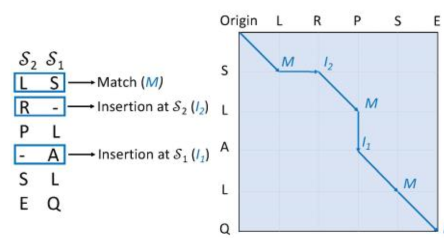
[Nature Comm-19, CACM-19, Science-22]

Leader-follower Games
Citizen Science



[AAMAS-23, UAI-22, ICMLA-22, SMC-22]

Learn Comb Structures



[AAAI-23,24]

[ICML-21, ECAI-20, UAI-18, 21, PGM-20, Brief-Bioinfo-22]