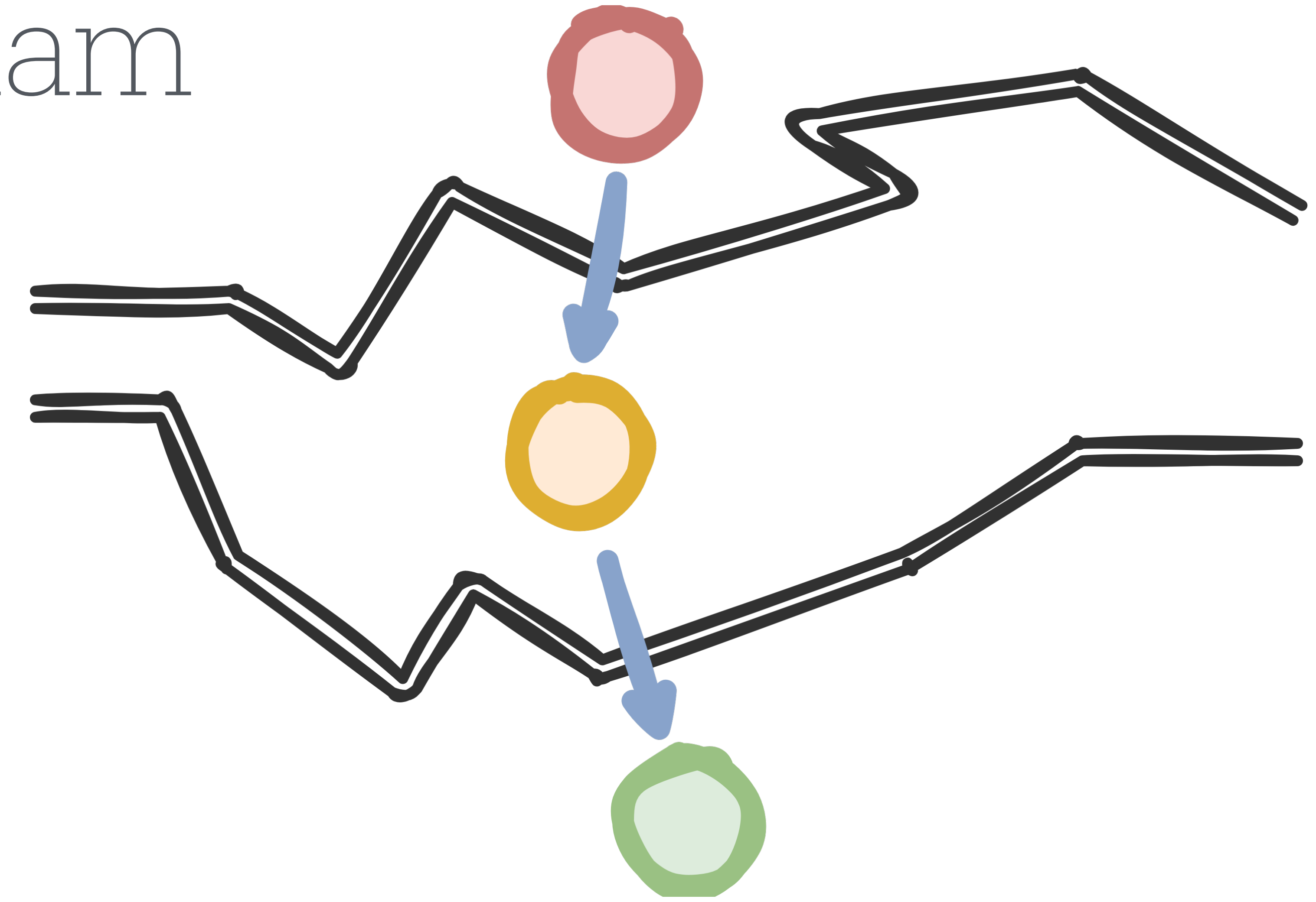
The Computational Complexity of Circuit Discovery for Inner Interpretability

Federico Adolfi, Martina Vilas, Todd Wareham

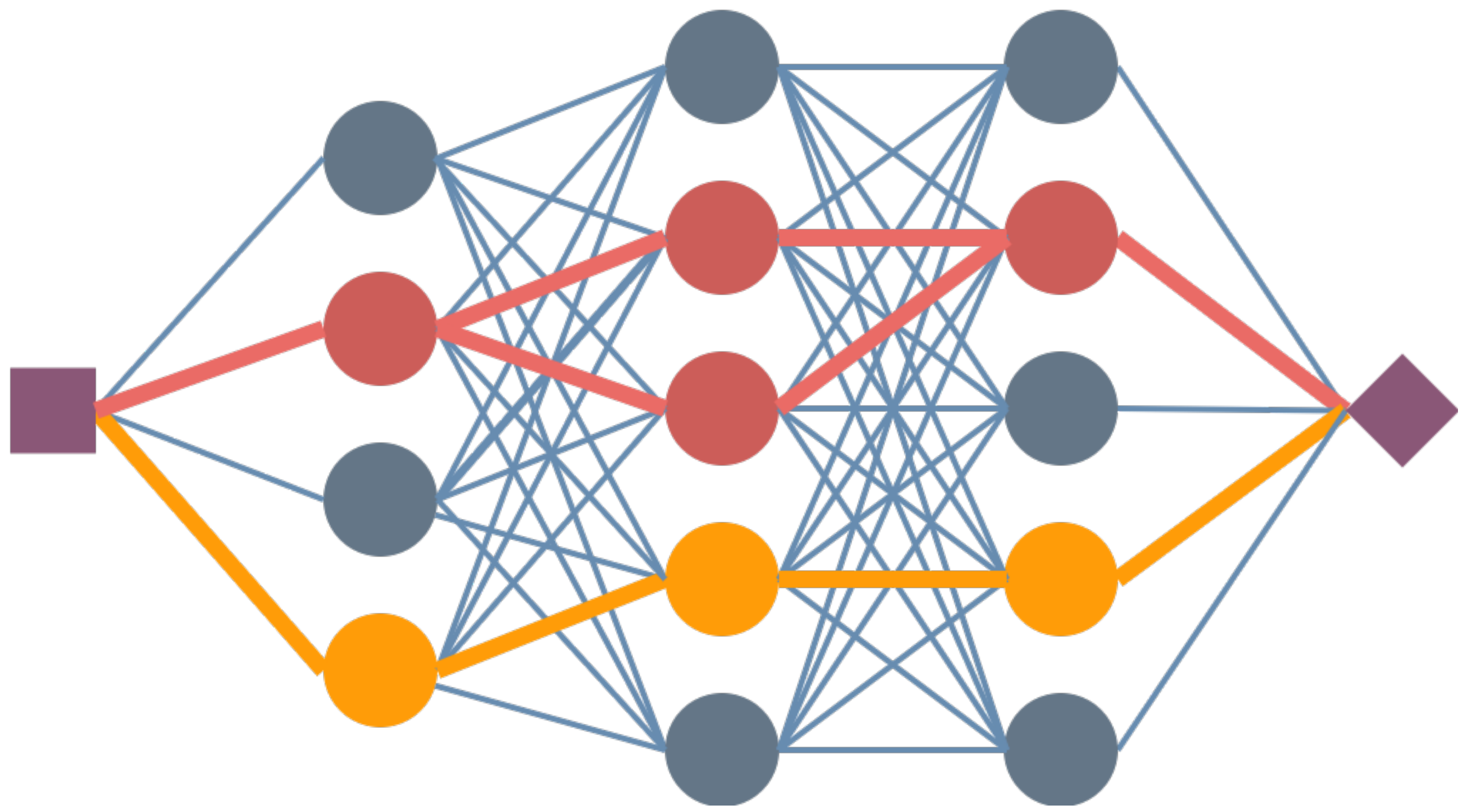


ICLR 2025

Spotlight

Circuit Discovery

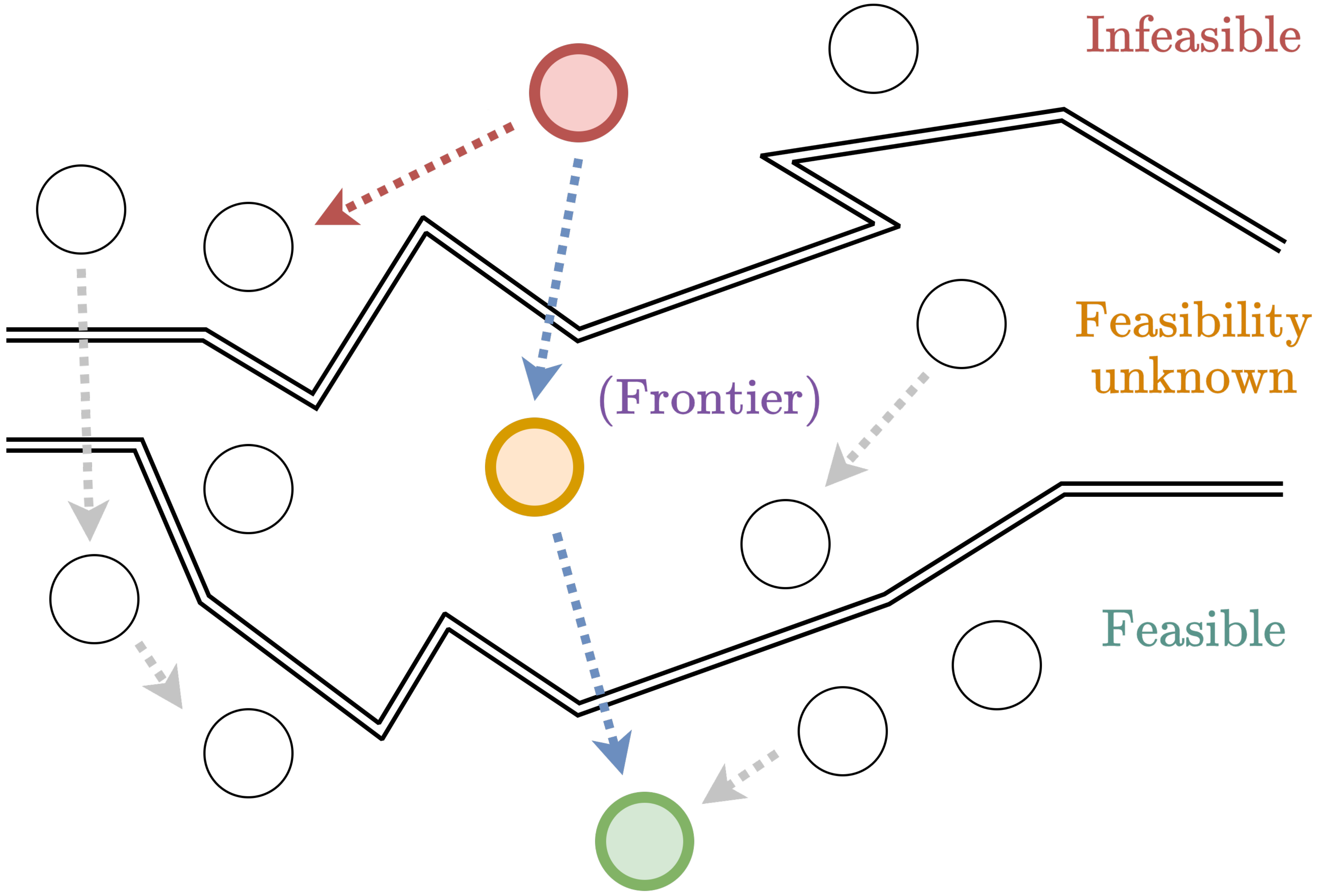
- Many applications of neural networks depend on the **feasibility of inner interpretability** through circuit discovery.
- **Circuit hypothesis:** networks might implement their capabilities via small circuits.
- This calls for empirical and theoretical explorations of viable interpretability queries and procedures to answer them.



Challenges and opportunities

- Automation and scalability
- Global/local faithfulness

- Theoretical exploration of viable interpretability queries is lacking
- Breakdown of scalability: computational complexity
- Intrinsic complexity of interpretability queries is unknown



Contributions

- Conceptualization of circuit queries in terms of affordances for description, explanation, prediction and control
- Formalization of a comprehensive set of queries and a formal framework for analysis
- Complexity-theoretic results for query variants, parameterizations, relaxations and approximation schemes in MLPs
- (Find other contributions in the paper)

Generating interpretability queries

Problem 0. PROBLEMNAME (PN)
Input: A multi-layer perceptron $\mathcal{M}$, `CoverageIN`, `SizeIN`.
Output: A `Property` circuit $\mathcal{C}$ of $\mathcal{M}$, `SizeOUT`, s.t. `CoverageOUT` $\mathcal{C}(\mathbf{x}) = \mathcal{M}(\mathbf{x})$, `Suffix`.

- Coverage
- Size
- Local / Global Minimality
- Necessity
- Sufficiency

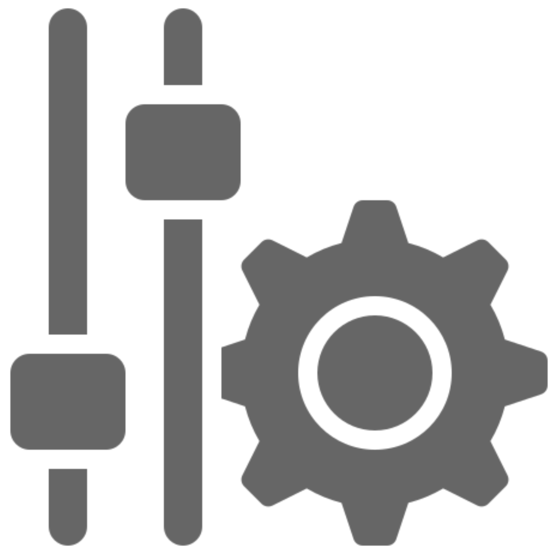
Table 2: Generating query variants from problem templates.

Description variables	**Query variants**					
	Local			*Global*		
	Bounded	*Unbounded*	*Optimal*	*Bounded*	*Unbounded*	*Optimal*
`CoverageIN`	an input $\mathbf{x}$	an input $\mathbf{x}$	an input $\mathbf{x}$	"__"	"__"	"__"
`CoverageOUT`	"__"	"__"	"__"	$\forall_{\mathbf{x}}$	$\forall_{\mathbf{x}}$	$\forall_{\mathbf{x}}$
`SizeIN`	int. $u \leq \lvert\mathcal{M}\rvert$	"__"	"__"	int. $u \leq \lvert\mathcal{M}\rvert$	"__"	"__"
`SizeOUT`	size $\lvert\mathcal{C}\rvert \leq u$	"__"	min. size	size $\lvert\mathcal{C}\rvert \leq u$	"__"	min. size
`Property`	minimal / "__"	minimal / "__"	"__"	minimal / "__"	minimal / "__"	"__"
`Suffix`	if it exists, otherwise $\perp$	"__"	"__"	if it exists, otherwise $\perp$	"__"	"__"

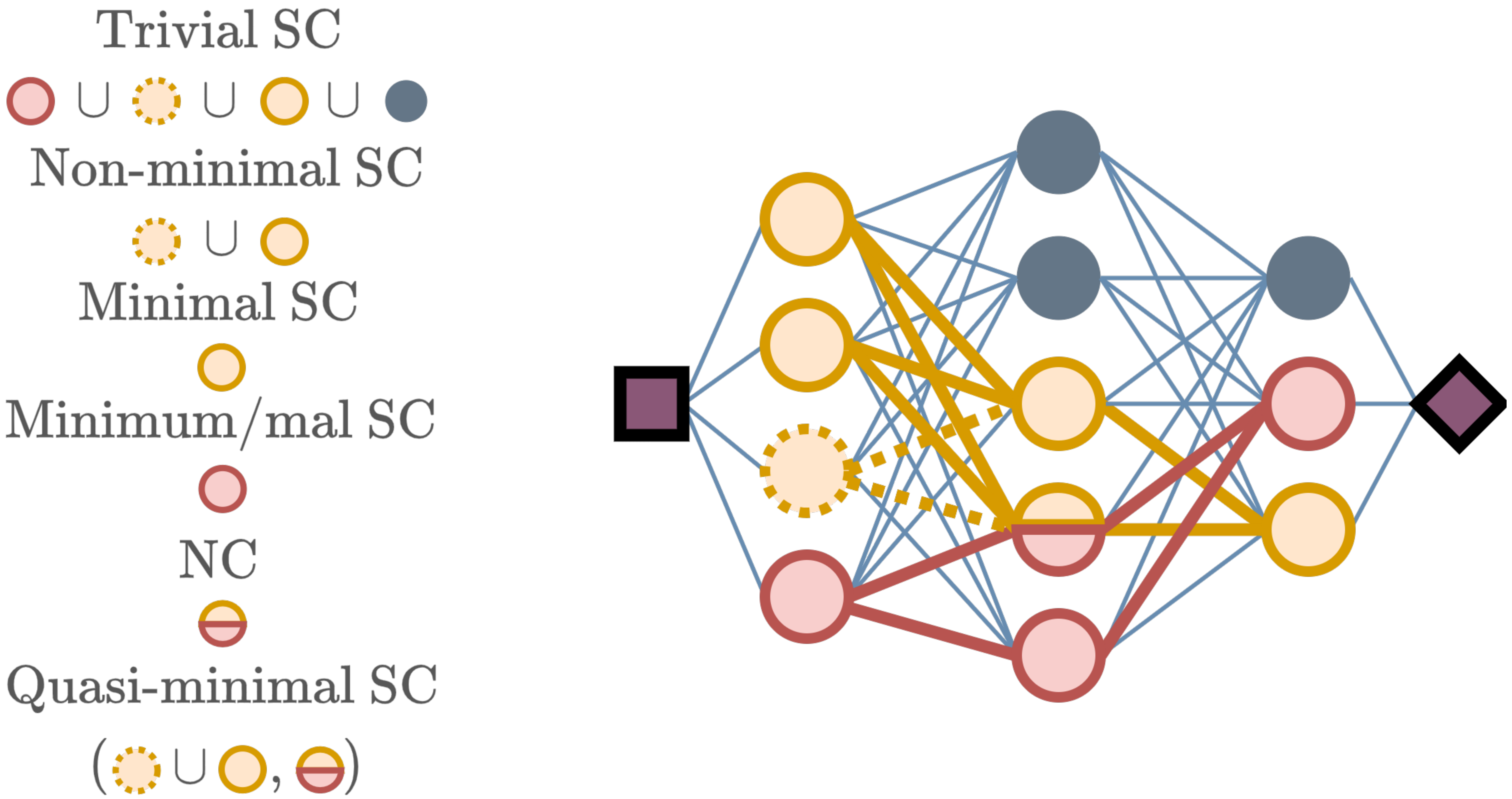
Circuits for explanation and control

Table 1: Circuit affordances for description, explanation, prediction, and control.

Circuit	**Affordance**	
	Description / Explanation	*Prediction / Control*
Sufficient Circuit	Which neurons suffice in isolation to cause a behavior? *Minimum:* shortest description.	Inference in isolation. *Minimal:* ablating any neuron breaks behavior of the circuit.
Quasi-minimal Sufficient Circuit	Which neurons suffice in isolation to cause a behavior and which is a breaking point?	Ablating the breaking point breaks behavior of the circuit.
Necessary Circuit	Which neurons are part of all circuits for a behavior? Key subcomputations?	Ablating the neurons breaks behavior of any sufficient circuit in the network.
Circuit Ablation & Clamping	Which neurons are necessary in the current configuration of the network?	Ablating/Clamping the neurons breaks behavior of the network.
Circuit Robustness	How much redundancy supports a behavior? Resilience to perturbations.	Ablating any set of neurons of size below threshold does not break behavior.
Patched Circuit	Which neurons drive a behavior in a given input context, i.e., are control nodes?	Patching neurons changes network behavior for inputs of interest. Steering; Editing.
Quasi-minimal Patched Circuit	Which neurons can drive a behavior in a given input context and which neuron is a breaking point?	Patching neurons causes target behavior for inputs of interest; Unpatching breaking point breaks target behavior.
Gnostic Neurons	Which neurons respond preferentially to a certain concept?	Concept editing; guided synthesis.



Circuits for explanation and control



Query approximations, parameterizations, relaxations

- Approximation
  - Additive
  - Multiplicative
  - Probabilistic
- Relaxation
  - Quasi-minimality

Table 3: Model and circuit parameterizations.

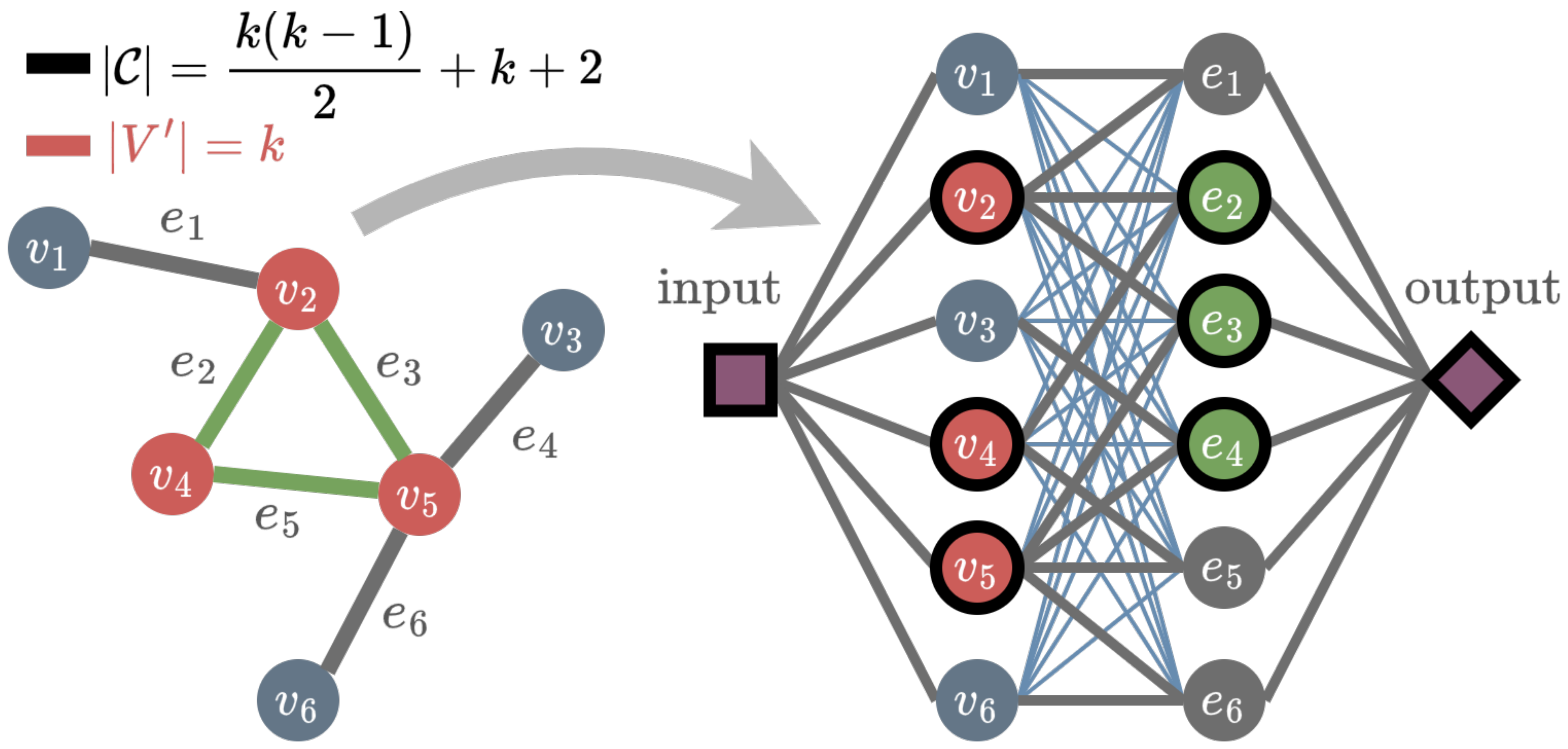
Parameter	*Model (given)*	*Circuit (requested)*
Number of layers (depth)	$\hat{L}$	$\hat{l}$
Maximum layer width	$\hat{L}_w$	$\hat{l}_w$
Total number of units[2]	$\hat{U} = \lvert\mathcal{M}\rvert \leq \hat{L} \cdot \hat{L}_w$	$\lvert\mathcal{C}\rvert = \hat{u}$
Number of input units	$\hat{U}_I$	$\hat{u}_I$
Number of output units	$\hat{U}_O$	$\hat{u}_O$
Maximum weight	$\hat{W}$	$\hat{w}$
Maximum bias	$\hat{B}$	$\hat{b}$

Problem 7. UNBOUNDED QUASI-MINIMAL LOCAL CIRCUIT PATCHING (UQLCP)

Input: A multi-layer perceptron $\mathcal{M}$, an input vector $\mathbf{y}$, and a set $\mathcal{X}$ of input vectors.

Output: A subset $\mathcal{C}$ in $\mathcal{M}$ and a neuron $v \in \mathcal{C}$, such that for the $\mathcal{M}^*$ induced by patching $\mathcal{C}$ with activations from $\mathcal{M}(\mathbf{y})$ and $\mathcal{M} \setminus \mathcal{C}$ with activations from $\mathcal{M}(\mathbf{x})$, $\forall_{\mathbf{x}\in\mathcal{X}} : \mathcal{M}^*(\mathbf{x}) = \mathcal{M}(\mathbf{y})$, and for $\mathcal{M}'$ induced by patching identically except for $v \in \mathcal{C}$, $\exists_{\mathbf{x}\in\mathcal{X}} : \mathcal{M}'(\mathbf{x}) \neq \mathcal{M}(\mathbf{y})$.

Parameterized complexity analyses



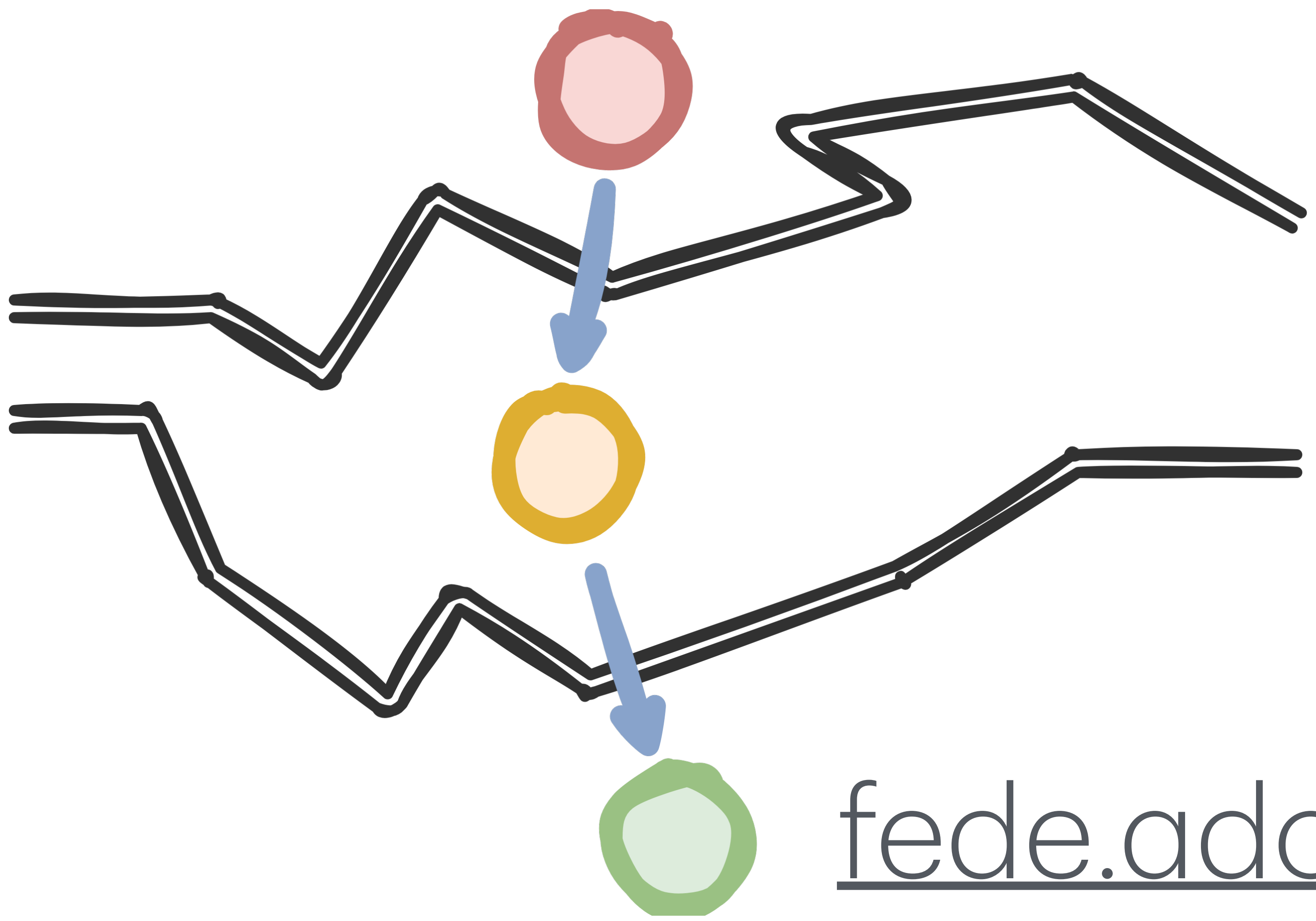
Results

- A challenging complexity landscape for circuit finding in MLPs:
  - Hardness
  - Fixed-parameter intractability
  - Inapproximability
- Transformations that could help tackle some hard problems with better understood heuristics
- Introducing relaxations yields feasible queries with potentially useful properties

Table 4: Classical and parameterized complexity results by problem variant.

Classical & parameterized queries[3] $\mathcal{P} = \mathcal{P}_{\mathcal{M}} \cup \mathcal{P}_{\mathcal{C}}$ $\mathcal{P}_{\mathcal{M}} = \{\hat{L}, \hat{U}_I, \hat{U}_O, \hat{W}, \hat{B}\}$ $\mathcal{P}_{\mathcal{C}} = \{\hat{l}, \hat{l}_w, \hat{u}, \hat{u}_I, \hat{u}_O, \hat{w}, \hat{b}\}$	**Problem variants**			
	Local		*Global*	
	Decision/Search	*Optimization*	*Decision/Search*	*Optimization*
SUFFICIENT CIRCUIT (SC)	NP-complete	$\mathcal{A}$-inapprox.	Σ_2^p-complete	$\mathcal{A}$-inapprox.
$\mathcal{P}$-SC	W[1]-hard	$\mathcal{A}$-inapprox.	W[1]-hard	$\mathcal{A}$-inapprox.
Minimal SC	NP-complete	?	$\in \Sigma_2^p \mid$ NP-hard	?
$\mathcal{P}$-Minimal SC	W[1]-hard	?	W[1]-hard	?
Unbounded Minimal SC	?	N/A	?	N/A
$\mathcal{P}$-Unbounded Minimal SC	?		?	
Unbounded Quasi-Minimal SC	PTIME		?	
Count SC	#P-complete	N/A	#P-hard	N/A
$\mathcal{P}$-Count SC	#W[1]-hard		#W[1]-hard	
Count Minimal SC	#P-complete		#P-hard	
$\mathcal{P}$-Count Minimal SC	#W[1]-hard		#W[1]-hard	
Count Unbounded Minimal SC	#P-complete		#P-hard	
GNOSTIC NEURON (GN)	PTIME	N/A	?	N/A
CIRCUIT ABLATION (CA)	NP-complete	$\mathcal{A}$-inapprox.	$\in \Sigma_2^p \mid$ NP-hard	$\mathcal{A}$-inapprox.
$\{\hat{L}, \hat{U}_I, \hat{U}_O, \hat{W}, \hat{B}, \hat{u}\}$-CA	W[1]-hard	$\mathcal{A}$-inapprox.	W[1]-hard	$\mathcal{A}$-inapprox.
CIRCUIT CLAMPING (CC)	NP-complete	$\mathcal{A}$-inapprox.	$\in \Sigma_2^p \mid$ NP-hard	$\mathcal{A}$-inapprox.
$\{\hat{L}, \hat{U}_O, \hat{W}, \hat{B}, \hat{u}\}$-CC	W[1]-hard	$\mathcal{A}$-inapprox.	W[1]-hard	$\mathcal{A}$-inapprox.
CIRCUIT PATCHING (CP)	NP-complete	$\mathcal{A}$-inapprox.	$\in \Sigma_2^p \mid$ NP-hard	$\mathcal{A}$-inapprox.
$\{\hat{L}, \hat{U}_O, \hat{W}, \hat{B}, \hat{u}\}$-CP	W[2]-hard	$\mathcal{A}$-inapprox.	W[2]-hard	$\mathcal{A}$-inapprox.
Unbounded Quasi-Minimal CP	PTIME	N/A	?	N/A
NECESSARY CIRCUIT (NC)	$\in \Sigma_2^p \mid$ NP-hard	$\mathcal{A}$-inapprox.	$\in \Sigma_2^p \mid$ NP-hard	$\mathcal{A}$-inapprox.
$\{\hat{L}, \hat{U}_I, \hat{U}_O, \hat{W}, \hat{u}\}$-NC	W[1]-hard	$\mathcal{A}$-inapprox.	W[1]-hard	$\mathcal{A}$-inapprox.
CIRCUIT ROBUSTNESS (CR)	coNP-complete	?	$\in \Pi_2^p \mid$ coNP-hard	?
$\{\hat{L}, \hat{U}_I, \hat{U}_O, \hat{W}, \hat{B}, \hat{u}\}$-CR	coW[1]-hard	?	coW[1]-hard	?
$\{\lvert H \rvert\}$-CR	FPT	FPT	?	?
$\{\lvert H \rvert, \hat{U}_I\}$-CR	FPT	FPT	FPT	FPT
SUFFICIENT REASONS (SR)	$\in \Sigma_2^p \mid$ NP-hard	3PA-inapprox.	N/A	
$\{\hat{L}, \hat{U}_O, \hat{W}, \hat{B}, \hat{u}\}$-SR	W[1]-hard	3PA-inapprox.		

Get in touch



fede.adolfi@bristol.co.uk

fedeadolfi.github.io