Minor Embedding for Max-Cut: Quantum Annealing's Combinatorial Bottleneck

"Minor Embedding is not so Minor"

Ron Lavi Alberto Moraglio Ofer M. Shir

presenting author underlined

MIGAL Galilee Research Institute LTD.

UNIVERSITY OF EXETER

TEL-HAI University of Kiryat Shmona in the Galilee

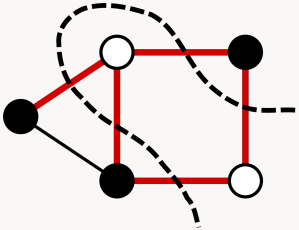
The 2026 IEEE World Congress on Computational Intelligence (WCCI)

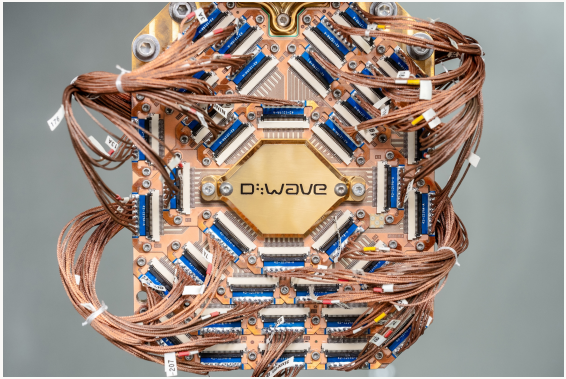
CEC'26 | June 2026, Maastricht, NL

IEEE WCCI 2026 Maastricht

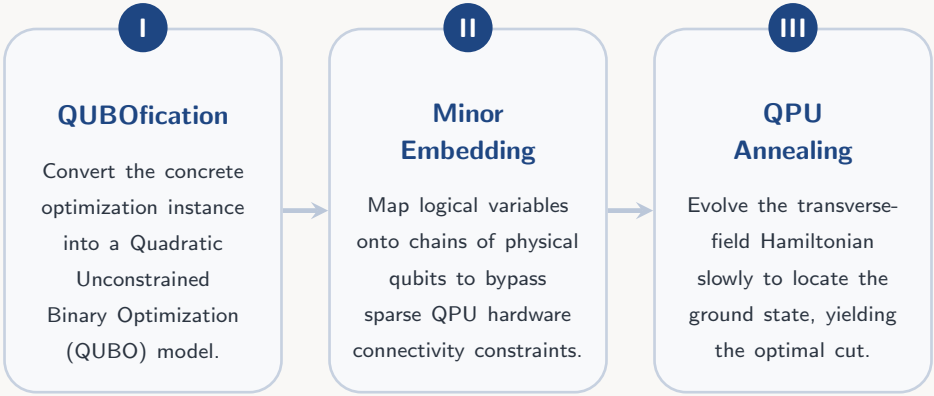
Outline

Introduction & Motivation
Problem Formulation
Approach & Methodology
Results
Conclusions



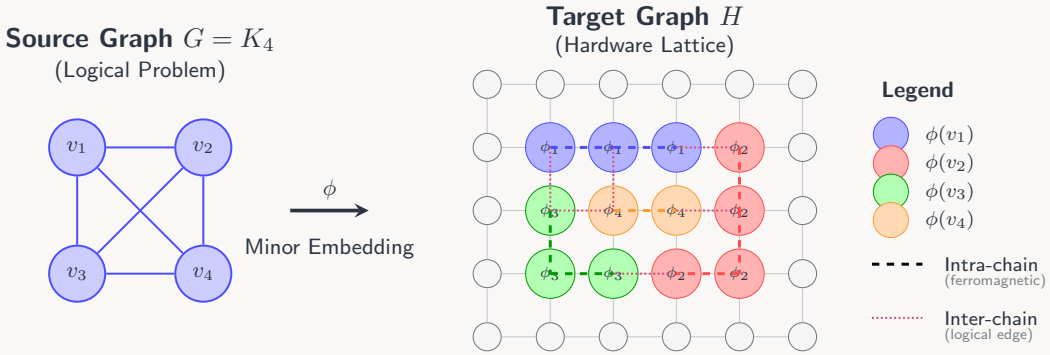


The QPU pipeline for combinatorial optimization



Sparse QPU connectivity restricts native mapping, forcing logical variables into physical chains. This introduces a $4.2\times$ **qubit overhead** and a 15% **drop in success rate**, with solution quality decaying exponentially if average chain lengths exceed the critical threshold of 20 **qubits**.

An embedding, concretely: K_4 onto the hardware



The dense logical K_4 becomes four qubit *chains*; intra-chain couplings hold each chain together, and all *six* logical edges are realized by inter-chain links on orthogonal lattice couplers.

Quantum annealing and where it breaks down

- **Quantum Annealing (QA):** a global *stochastic search* heuristic for combinatorial optimization, exploiting tunneling and superposition.
- Problems are posed as **QUBO** models $\Rightarrow$ a *logical graph* of binary variables and their interactions.
- **The catch:** QA hardware (D-Wave) is *sparse* and *fixed* — Chimera (deg. 6), Pegasus (deg. 15), Zephyr (deg. 20).

Minor Embedding (ME)

Map the dense logical graph onto the sparse hardware by representing each logical variable as a connected *chain* of physical qubits. This pre-solving step decides *feasibility* and *quality* — yet is $\mathcal{NP}$-hard.

Why ME is the bottleneck — and the research gap

Practical consequences

- Quality is measured by *average chain length* (ACL); long chains $\Rightarrow$ *chain breaks* $\Rightarrow$ degraded solutions.
- Embedding overhead often *dominates* total runtime — by orders of magnitude.
- Dense problems: chain-length ratios of $10{:}1$ or worse.

The gap

- ME is mostly *abstracted away* or treated as a nuisance.
- Knowledge sits behind *commercial* walls.
- Prior empirical work: *single topology*, *synthetic* random graphs.

Need

Independent, *cross-topology* study on *established* benchmarks.

Minor embedding: the formal object

Graph minor

G is a *minor* of H ($G \preceq_m H$) if G is obtainable from H by vertex deletions, edge deletions, and edge contractions. (Robertson–Seymour theory)

A mapping $\varphi : V_G \to 2^{V_H}$ assigns each logical vertex a *chain* of qubits, with:

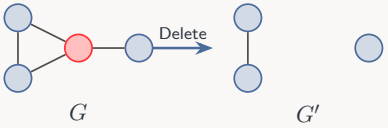
(C1) Disjointness $\varphi(u) \cap \varphi(v) = \emptyset$ for $u \neq v$.

(C2) Connectivity $H[\varphi(v)]$ is connected for every v.

(C3) Edge preservation every logical edge (u, v) is realized by a physical edge between $\varphi(u)$ and $\varphi(v)$.

Each logical vertex → a connected subtree of H; logical edges → inter-chain links.

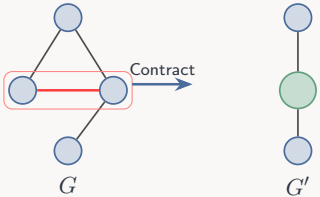
Graph minors: illustrating the elementary operations



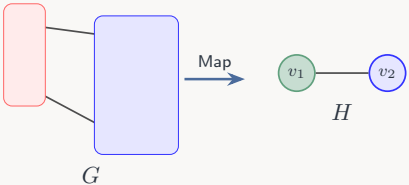
(a) Vertex deletion



(b) Edge deletion



(c) Edge contraction



(d) Minor mapping

Minor embedding as exact combinatorial optimization

Binary $x_{v,i} = 1$ iff qubit i belongs to the chain of logical vertex v. An MTZ-based IP:

Two objectives — economy vs. integrity

$$f_1(\vec{x}) = \sum_v \sum_i x_{v,i} \;\rightarrow\; \min \qquad \text{(total chain length)}$$

$$f_2(\vec{x}) = L_{\max} \;\rightarrow\; \min, \qquad L_{\max} \geq \textstyle\sum_i x_{v,i} \;\; \forall v \qquad \text{(max chain length)}$$

Constraint roles: *disjointness* & *surjectivity* of qubits; *connectivity* via Miller–Tucker–Zemlin subtour elimination; a quadratic *pullback* (logical-edge realization) linearized via *McCormick* relaxation — at a cost of $\mathcal{O}(|E_G| \cdot |E_H|)$ auxiliary variables.

ME can even be cast *as a QUBO* itself — a self-referential, densely coupled problem.

The pragmatic standard: MinorMiner

The *de facto* industry heuristic (Cai–Macready–Roy)

A randomized, iterative *shortest-path* routine: seed logical vertices onto qubits, then repeatedly route unsatisfied logical edges through the hardware, *tearing and repairing* congested chains, until all edges are realized or a budget is exhausted.

- Per-iteration cost $\approx \mathcal{O}(|E_G|\,(|V_H| + |E_H|))$; embeds $n \lesssim 200$ in seconds.
- **Prioritizes feasibility over chain length** — offers *no* guarantee on the total number of qubits used.
- Highly sensitive to *initial placement*; needs multiple *restarts* as source density approaches the hardware's capacity.

Research questions & contributions

Research questions

RQ1 Does *edge density* ρ — the conventional predictor — remain sufficient across diverse, established benchmark graphs?

RQ2 Which graph-theoretic properties govern *feasibility* vs. *cost*, and do they hold across Chimera, Pegasus, and Zephyr?

Threefold contribution

1. First *cross-topology* study of MinorMiner over ≈ 550 benchmark instances on Chimera, Pegasus, Zephyr.
2. A *structural predictor hierarchy*: the MIS ratio beats edge density for predicting *feasibility*.
3. A *functional dissociation*: MIS ratio governs feasibility; edge density dictates cost — robust across all three generations.

Why Max-Cut is the ideal test case

Given $G = (V, E)$ with weights $w_{ij} \geq 0$, MAX-CUT seeks a bipartition of V maximizing the crossing-edge weight:

$$Q_{ij} = \begin{cases} \deg(i) & i = j, \\ -1 & (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases} \qquad \Longrightarrow \qquad \mathbf{x}^\top Q\,\mathbf{x} = \sum_{(i,j)\in E} \big(x_i - x_j\big)^2.$$

For binary $\mathbf{x} \in \{0,1\}^n$, each edge term is $(x_i - x_j)^2 = \mathbf{1}[x_i \neq x_j]$, so $\mathbf{x}^\top Q\,\mathbf{x}$ *counts the cut*. Hence

$$\mathrm{MaxCut}(G) = \max_{\mathbf{x}\in\{0,1\}^n} \mathbf{x}^\top Q\,\mathbf{x} = -\min_{\mathbf{x}\in\{0,1\}^n} \mathbf{x}^\top(-Q)\,\mathbf{x}.$$

- $\mathcal{NP}$-hard and *well-studied* in both classical and quantum communities.
- **Direct QUBO formulation** — *no* auxiliary variables, *no* penalty terms.
- Isolates *structural* embedding effects from *encoding artifacts*.

Experimental design: no Max-Cut solving — ME in isolation

Benchmark instances (MQLib subset)

- From MQLib ($\approx$3500 Max-Cut graphs), a *size-stratified* set: $N = 548$ in 13 blocks of $\approx$40, $|V| \in [50, 540]$.
- Each instance carries precomputed graph-theoretic metrics.

Target topologies (idealized, defect-free)

- Chimera `C25`, Pegasus `P16`, Zephyr `Z12` via `dwave_networkx`, at comparable qubit scale.

Protocol

- $R = 10$ independent randomized MinorMiner attempts per instance.
- Log success flag, runtime, and chain statistics.
- Chain metrics computed *only* for successful embeddings.

Outcome measures

Feasibility (*success rate*); cost (ACL, MCL); embedding *time*.

Two predictors, and how we compare them

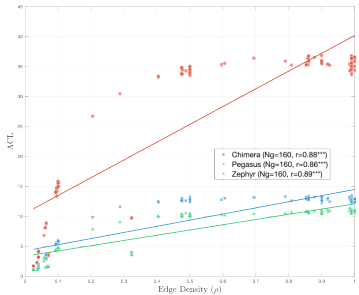
The competing structural metrics

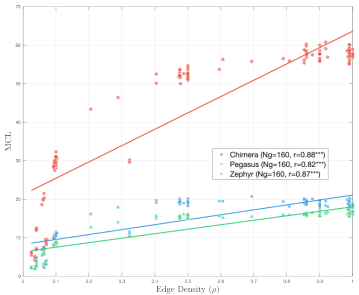
- **Edge density** $\rho = \dfrac{2m}{n(n-1)}$ — a *global, mean-field* measure of connectivity demand.
- **mis ratio** α/n — normalized largest independent set; sensitive to the *local* distribution of edges (structural "slack").

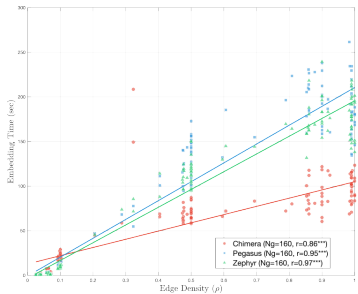
Statistical methodology

- Pearson correlations (r, p) between metrics and outcomes, per architecture.
- **Double-dissociation** via AIC model selection over four models: M_0 (null), M_1 (α/n), M_2 (ρ), M_3 (full).
- $\Delta\text{AIC} = \text{AIC}(M_2) - \text{AIC}(M_1)$: > 0 favors MIS, < 0 favors ρ; $|\Delta\text{AIC}| > 10$ is very strong evidence.
- Feasibility models re-fit with *binomial-logistic* regression as a robustness check.

Analyses of the ME performance measures ($n = 100$)



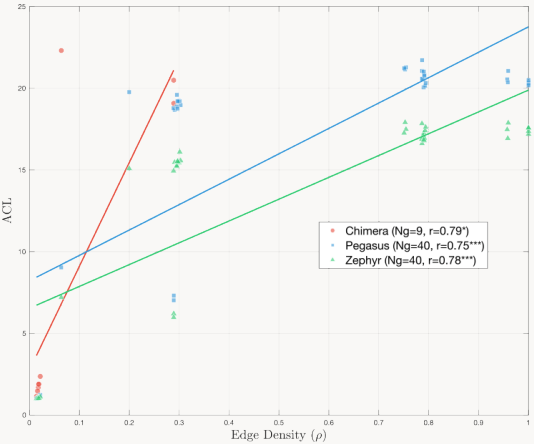




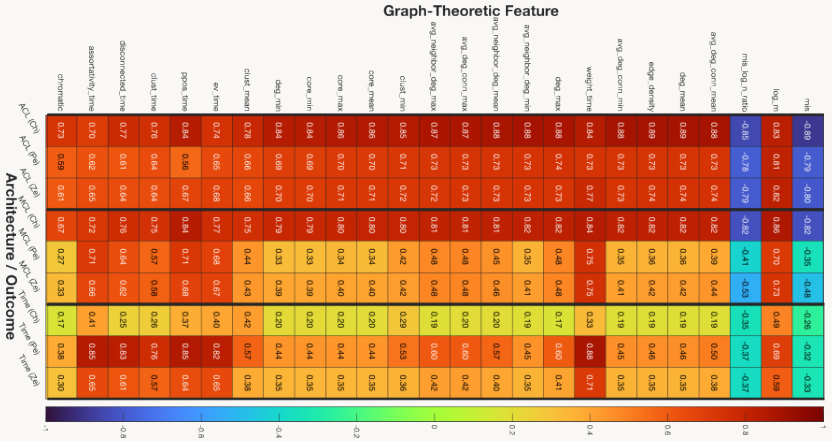
Pearson correlation tests against Edge Density (ρ) for ACL (left), MCL (middle) and Embedding Time (right).

RQ1: edge density is necessary but not sufficient

- Where *all* instances embed (e.g. $n = 100$, 100% success), ρ correlates *strongly* with ACL, MCL, and embedding time.
- But on *harder* subsets (e.g. $n = 150$, lower success), instances of *similar* ρ show *highly dispersed* ACL— density *decouples* from cost.
- There, *independence*-based metrics — led by the MIS ratio α/n — gain predictive power.



Predictive Hierarchy of Graph-Theoretic Metrics



Comment: Features are ranked by their mean absolute correlation with ACL.

A structural hierarchy emerges

Ranking the full feature space by mean $|\bar{r}|$ with ACL ($N = 548$, all architectures):
The ranking shows the following

- α/n**-related metrics** dominate the top — strong *negative* correlation with chain length.
- Then degree-distribution statistics, then core-structure metrics.
- **Edge density** ρ **is no longer dominant** in this aggregate view.

A structural hierarchy emerges

Ranking the full feature space by mean $|\bar{r}|$ with ACL ($N = 548$, all architectures):
The ranking shows the following

- α/n**-related metrics** dominate the top — strong *negative* correlation with chain length.
- Then degree-distribution statistics, then core-structure metrics.
- **Edge density** ρ **is no longer dominant** in this aggregate view.

Hypothesis

α/n predicts *feasibility*; ρ predicts *cost*.

Double dissociation: gatekeeper vs. toll

AIC model selection, $N = 548$ per architecture — the hypothesis is confirmed in **6/6** tests.

Arch.	Outcome	$R^2_{\alpha/n}$	R^2_{ρ}	ΔAIC	Winner
Feasibility (expected: α/n)					
Chimera	success	.194	.000	$+118.3$	α/n
Pegasus	success	.173	.013	$+96.7$	α/n
Zephyr	success	.173	.014	$+96.5$	α/n
Cost (expected: ρ)					
Chimera	ACL	.236	.784	-422.5	ρ
Pegasus	ACL	.239	.534	-184.0	ρ
Zephyr	ACL	.246	.540	-186.8	ρ

Robustness: binomial-logistic re-fit preserves the α/n sign, with far larger magnitudes (e.g. Chimera $+2414.6$ vs. $+118.3$) — a genuine structural property, not a Gaussian artifact.

Interpreting the dissociation

The *gatekeeper*: α/n

Structural *slack* decides **binary feasibility**. Too little independence $\Rightarrow$ the graph *cannot* fit — *regardless* of density.

The *toll*: ρ

Once *past* the gate, edge density sets the **resource cost** (ACL) of the successful embedding.

- Two *distinct* structural dimensions of "compactness": $\rho =$ global connectivity demand; $\alpha/n =$ local placement availability.
- Overcoming the *gate* needs different properties than minimizing the *toll*.
- **Invariant** across Chimera, Pegasus, and Zephyr.

Conclusions

- We frame ME as a *standalone* combinatorial optimization challenge and give a rigorous formulation (graph-theoretic, IP, and QUBO views).
- Across $500+$ established Max-Cut instances and three hardware generations:
  - ρ predicts *cost* — but only *inside* the feasible region;
  - α/n (MIS ratio) predicts *feasibility* — the gatekeeper.
- A robust *double dissociation*, invariant across topologies.

Take-away for Evolutionary Computation

The *local* distribution of edges — not global density alone — determines embedding difficulty. ME is a clean, structurally-rich *benchmark* for EC heuristics.

Scope & future work

- **Scope is about coverage, not method:** ME is *purely graph-theoretic* — the embedder sees only an abstract G. Our predictors thus generalize to *any* QUBO whose logical graph lies in the studied structural regime.
- Max-Cut-derived graphs may *underrepresent* penalty-induced cliques and auxiliary-variable motifs — a natural extension.
- **Toward a structural ontology:** condense the high-dimensional metric space via latent-structure estimation, deduplication, and semantic clustering into a compact set of *medoid* predictors.
- Goal: an *architecture-agnostic* model predicting QA performance from a problem's topological "fingerprint."

Thank you — Questions welcome.

Code & Correlation Plots

github.com/.../MinorEmbeddingCEC2026

This Presentation PDF

ofersh.github.io/.../ME-for-MaxCut.pdf

Scientific Team Contact:

ronl@migal.org.il · a.moraglio@exeter.ac.uk · ofersh@telhai.ac.il

Meanwhile, in the box, Schrödinger's cat is planning its revenge…

