POTTSENERGY: LEARNING INTERACTIONS FOR COMBINATORIAL PROTEIN STABILITY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting the stability effects of combined mutations requires interaction models that generalize beyond individual substitutions. To meet this need, we introduce PottsEnergy, a compact framework that jointly designs structural encoding, explicit Potts parameterization, and sitewise self-supervision. Specifically, a backbone graph network predicts fields and pair couplings, while native-context pseudolikelihood and masked-context residue reconstruction train the same potentials used for mutation scoring. As a result, this design turns residue-level learning into a reusable sequence energy that combines single-site effects with explicit pair corrections. To evaluate this representation, we curate a higher-order stability benchmark from ten data sources. Even without stability-label training, PottsEnergy achieves state-of-the-art aggregate zero-shot ranking on this benchmark among the evaluated model families. Matched additive controls further show that the predictive contributions of the pair corrections are dataset-dependent. Separate diagnostics then distinguish these gains from interaction-sign accuracy. In addition, single-mutant adaptation tests the compact representation further. Together, these results connect residue reconstruction, combinatorial stability prediction, and candidate design through an explicit interaction-energy representation. More broadly, coupling learned interaction energies with quantum-assisted sampling could broaden the exploration of combinatorial sequence space for protein design.

1 INTRODUCTION

Combinatorial protein design couples two problems: predicting the stability effects of interacting substitutions and finding favorable combinations in a rapidly growing sequence space. A mutation can change its effect across sequence backgrounds, and combinations can reverse the effects of their component singles (Olson et al., 2014; Dieckhaus and Kuhlman, 2025). Choosing k substitutions in a protein of length L already defines $\binom{L}{k}19^k$ candidates. An effective representation must capture these interactions and make their evaluation reusable during search.

Structure-conditioned residue reconstruction learns sequence constraints without stability measurements (Dauparas et al., 2022; Hsu et al., 2022; Yang et al., 2023). Extending it to combination prediction requires a consistent score across mutation contexts. Isolated-single addition omits interactions between mutated sites, while successive conditional predictions depend on the contexts in which they are evaluated. Explicit Potts fields and couplings provide a common sequence energy. Structure-to-Potts networks establish this construction (Li et al., 2023; Birnbaum and Keating, 2026); the opportunity is to make sitewise interaction learning serve both zero-shot stability ranking and efficient candidate search.

We introduce PottsEnergy, a 1.052-million-parameter framework that jointly designs backbone encoding, explicit interaction heads and sitewise self-supervision (Figure 1). The structural encoder excludes residue identities and side-chain coordinates. Native-context pseudolikelihood and masked-context residue reconstruction update the same fields and couplings later used to score substitutions. Interactions learned while reconstructing individual residues therefore remain available as pair corrections when multiple sites change.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

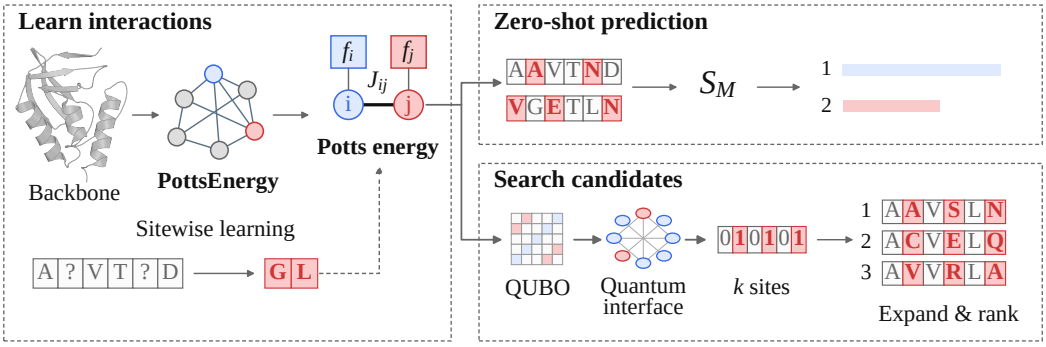


Figure 1: **PottsEnergy links residue learning, combinatorial prediction and candidate search.** Sitewise reconstruction trains structure-conditioned fields and couplings, and the resulting energy scores combined substitutions without stability-label training. The position QUBO supplies a sampling interface; proposed masks then undergo full-alphabet expansion and energy ranking. Figure 3 illustrates the workflow; Section 4.5 evaluates it with a physical CIM and a separately single- and double-supervised ensemble. All illustrations are schematic.

In summary, our primary contributions are as follows:

1. **Interaction learning for combinatorial prediction.** PottsEnergy combines backbone-only encoding, explicit Potts fields and couplings, and native- and masked-context site reconstruction in one scoring energy. Without stability-label training, it leads the aggregate zero-shot ranking among 13 evaluated model families; component ablations and matched additive controls identify the contributions of learned interactions.
2. **A structure-mapped higher-order benchmark.** We assemble 306,014 variants with at least three substitutions from ten data sources into 679 assay–condition groups. Exact structure mapping, source-overlap audits and endpoint-specific scoring make comparisons traceable while keeping thermodynamic measurements separate from abundance proxies.
3. **Candidate search through reusable interaction energies.** We turn precomputed interactions into position proposals, lower-bound pruning, full-alphabet expansion and original-energy ranking. On ten proteins, a separately supervised ensemble recovers the frozen-energy Top-1 and Top-5 sequences for two- and three-site mutations. After position proposal and pruning, it evaluates 475- and 2,747-fold fewer complete sequences than naive full-alphabet enumeration, respectively.

2 Related Work

Structural encoding and Potts parameterization. ProteinMPNN uses backbone node–edge message passing and random-order autoregressive reconstruction (Dauparas et al., 2022); ESM-IF1 combines geometric processing and a Transformer (Hsu et al., 2022). Evolutionary Potts models fit fields and couplings to homologs (Ekeberg et al., 2013; Hopf et al., 2017). TERMinator predicts Potts parameters using tertiary motifs, COORDinator uses coordinates, and PottsMPNN extends backbone message passing with Potts outputs (Li et al., 2023; Birnbaum and Keating, 2026). PottsEnergy couples explicit node and edge representations to a shared energy for residue learning, mutation scoring and search.

Learning interactions from residue context. Pseudolikelihood learns conditional residue probabilities (Ekeberg et al., 2013). TERMinator/COORDinator use pair composite pseudolikelihood; PottsMPNN combines it with autoregressive reconstruction and also evaluates single-site Potts training (Li et al., 2023; Birnbaum and Keating, 2026). MIF uses masked inverse folding, MIF-ST adds sequence-model transfer (Yang et al., 2023), and Caliby learns Potts potentials from partial sequence and side-chain context (Shuai et al., 2025). Our backbone-only encoder supplies potentials jointly trained by native- and masked-context site reconstruction, then reused directly for combinatorial stability prediction.

Prediction and search. Tables 1 and 2 summarize the evaluated model families. Single-mutant adaptation transfers representations to stability labels (Dieckhaus et al., 2024; Li and Luo, 2026), while nonlinear phenotype mappings (Otwinowski et al., 2018) and assay-range effects (Reeves and Kalyaanamoorthy, 2026) motivate separate interaction diagnostics. Binary quadratic optimization (Kochenberger et al., 2014) supplies the search interface: we combine position-level proposals with full-alphabet expansion and original-energy ranking. TERMinator and Caliby provide methodological precedents rather than evaluated predictors.

Table 1: External models used for zero-shot mutation scoring.

Model	Input	Representation and learning or scoring principle
ESM-1v / ESM-2 (Meier et al., 2021; Lin et al., 2023)	Sequence	Masked protein language models; residue likelihoods reflect sequence constraints.
SaProt (Su et al., 2024)	Sequence, structure tokens	Masked modeling with a joint amino-acid and structural vocabulary.
ProteinMPNN (Dauparas et al., 2022)	Backbone, sequence context	Geometric message passing with random-order autoregressive sequence reconstruction.
ESM-IF1 (Hsu et al., 2022)	Backbone, sequence context	Geometric vector encoder and Transformer trained for inverse folding.
MIF / MIF-ST (Yang et al., 2023)	Backbone, visible sequence	Masked inverse folding; MIF-ST also transfers pretrained sequence-model outputs.
COORDinator (Li et al., 2023)	Backbone	Graph network predicting Potts tables, trained by pair composite pseudolikelihood.
PottsMPNN (Birnbaum and Keating, 2026)	Backbone	Joint autoregressive reconstruction and explicit Potts learning.
VenusREM (Tan et al., 2025)	Sequence, structure, homologs	Retrieval-enhanced sequence/structure representations combined with homolog information.
StructureDCA[RSA] (Tsishyn et al., 2026)	Alignment, structure	Contact-restricted Potts couplings with relative solvent accessibility reweighting.
EvoEF2 (Huang et al., 2020)	Atomic structure	Empirical energy function combining physical and statistical terms.

Paired names denote separately evaluated models. Zero-shot refers to the absence of stability-label fitting in this evaluation; EvoEF2 is an energy-function reference. The primary PottsMPNN comparison uses the no-MSA checkpoint; its MSA-trained variant is a sensitivity analysis. Benchmark-specific readouts are defined in Appendix A.7.

Table 2: Supervised stability models and regression controls included in this study.

Model	Core representation	Stability supervision and use here
SPURS (Li and Luo, 2026)	ProteinMPNN and ESM-2 linked by an adapter	Experimental stability labels; single-trained weights in the matched comparison.
ThermoMPNN (Dieckhaus et al., 2024)	ProteinMPNN features with light attention and a regressor	Experimental stability labels; single-mutant checkpoint used here.
ProteinMPNN reg. (Dauparas et al., 2022)	ProteinMPNN features and a regression head	Study-specific readout trained on experimental single-mutant labels.
ESM-2 reg. (Lin et al., 2023)	ESM-2 features and a regression head	Study-specific readout trained on experimental single-mutant labels.
RaSP (Blaabjerg et al., 2023)	Local atomic-environment encoder and a small regressor	Original regressor trained on Rosetta estimates; retrained here on B1 single-mutant labels and used as an additive reference.

The regression rows cite their pretrained backbone models; they are distinct from the zero-shot versions in Table 1. No double-trained SPURS-Multi or ThermoMPNN-D checkpoint enters the matched comparison. RaSP has no joint inference extension here (Appendix C.8).

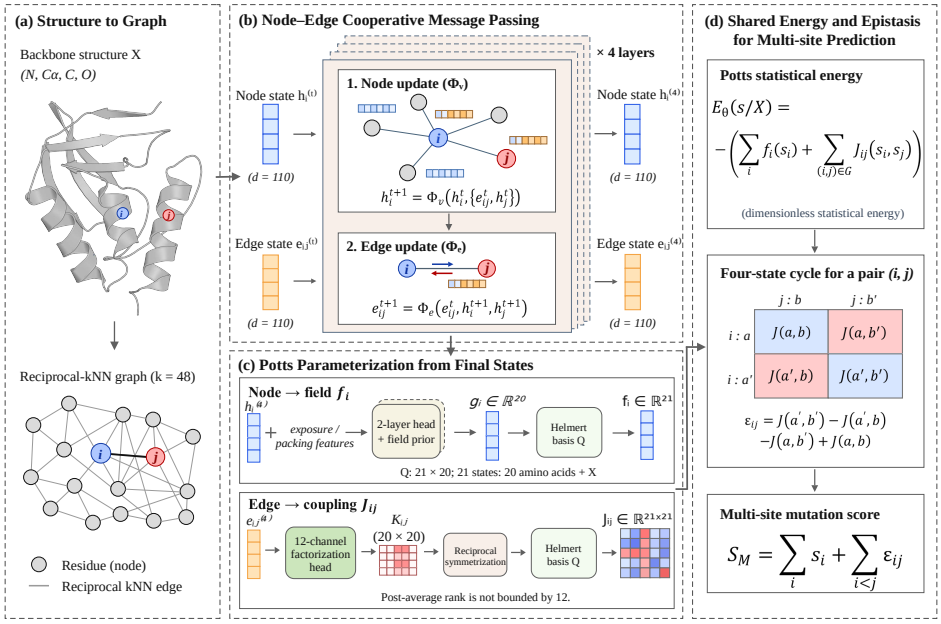


Figure 2: **PottsEnergy maps a reciprocal-kNN structure graph to a shared field–coupling energy.** **a**, A fixed backbone defines a reciprocal 48-neighbor graph. **b**, Four width-110 node–edge co-updates propagate structural context while retaining explicit edge states. **c**, Site-field and factorized coupling heads produce reduced coordinates; reciprocal symmetrization and a fixed Helmert basis impose the zero-sum gauge. **d**, The resulting f_i and J_{ij} jointly define the sequence energy and its four-state pair correction for multi-site scoring.

3 METHOD

3.1 ONE ENERGY FOR SINGLE AND COMBINED SUBSTITUTIONS

Given backbone X, PottsEnergy predicts fields f_i and pair couplings J_{ij} on an undirected interaction graph G, counting each residue pair once in the energy. For wild-type sequence s^0 and mutation set M, define

$$E_\theta(s \mid X) = -\sum_i f_i(s_i) - \sum_{(i,j)\in G} J_{ij}(s_i, s_j), \qquad S_M = E_\theta(s^0 \mid X) - E_\theta(s^M \mid X). \quad (1)$$

Larger S_M indicates greater stability; E_θ is a dimensionless statistical energy. For substitutions $a \to a'$ and $b \to b'$ at sites i, j,

$$S_{ij} = S_i + S_j + \epsilon_{ij}, \qquad \epsilon_{ij} = J_{ij}(a', b') - J_{ij}(a', b) - J_{ij}(a, b') + J_{ij}(a, b). \quad (2)$$

The four-state cycle cancels fields and unchanged-neighbor interactions. Each single effect already includes its native-context couplings; ϵ_{ij} accounts for changing both target sites. Summing singles and pair corrections gives the exact energy difference for any mutation set in this pairwise model. The matched additive control removes only target–target corrections, which are zero outside G.

3.2 STRUCTURAL ENCODING AND EXPLICIT INTERACTION HEADS

Backbone atoms N, C_α, C and O define a reciprocal 48-neighbor graph. Four width-110 node–edge updates produce structural representations (Figure 2). Final node states feed the field head; edge states produce two 20×12 factors for a reduced 20×20 coupling table. Reverse-edge transpose averaging enforces reciprocity. A fixed Helmert basis $Q \in \mathbb{R}^{21\times 20}$ maps these outputs to $f_i = Qg_i$ and $J_{ij} = QK_{ij}Q^\top$ in a common zero-sum gauge. The 21 states include 20 amino acids and X; an exposure/packing prior enters the fields. Appendix C.1 specifies the architecture.

The 1,052,014 shared parameters generate protein-specific energy tables once per backbone. Table storage scales as $O(|G|q^2)$ for alphabet size q, and cached single effects require $O(Lq)$ values. A

candidate combination is then scored by table lookup and summation, amortizing neural computation across mutation sets.

3.3 SITEWISE SELF-SUPERVISION TRAINS THE SCORING ENERGY

The energy induces conditional logits $\ell_i(a) = f_i(a) + \sum_j J_{ij}(a, s_j)$. Native-context pseudolikelihood trains the fields and couplings to reconstruct each residue. A second objective masks 15% of the sequence context and reconstructs masked residues from fields and visible-neighbor couplings. Residue identities and side-chain coordinates are excluded from the encoder; visible identities enter the conditional readout.

Pretraining combines pseudolikelihood, masked reconstruction and coupling regularization (Equation S11). These objectives train the same potentials used for mutation scoring, so sitewise reconstruction produces combination scores without stability labels or an additional prediction head. Optional single-mutant adaptation updates the model through $E(WT) - E(MUT)$ using ranking, Huber and pseudolikelihood losses; its fitting and selection use singles (Appendices A and C.1).

3.4 POSITION PROPOSALS AND FULL-ALPHABET CANDIDATE SEARCH

We turn the explicit energy into a search objective that separates position selection from amino-acid assignment. For a fixed keep/substitute choice at each site, the energy is a quadratic unconstrained binary optimization (QUBO) problem (Kochenberger et al., 2014):

$$E(x) = C + \sum_i q_i x_i + \sum_{i<j} q_{ij} x_i x_j, \qquad q_i = -S_i, \quad q_{ij} = -\epsilon_{ij}. \tag{3}$$

To search all amino-acid alternatives, let $u_i(a)$ and $v_{ij}(a, b)$ denote single and pair energy changes relative to an iterated conditional modes (ICM) reference c (Besag, 1986). For $k \geq 2$ substitutions, define

$$B_{ij}^{(k)} = \min_{a \neq c_i,\, b \neq c_j} \left[\frac{u_i(a) + u_j(b)}{k-1} + v_{ij}(a, b) \right], \qquad P_k(x) = \sum_{i<j} B_{ij}^{(k)} x_i x_j + \lambda \left(\sum_i x_i - k \right)^2. \tag{4}$$

Here x_i selects a position and λ enforces the mutation count. On exactly-k masks, the unpenalized position objective equals the minimum energy over amino-acid assignments for $k = 2$ and lower-bounds that minimum for higher orders. It converts amino-acid pair costs into a position-level proposal problem.

The search pipeline filters and deduplicates proposed masks, prunes masks whose original-energy lower bounds cannot enter the current Top-5, and expands all 19^k assignments for each remaining mask (Figure 3). Expansion enforces a consistent amino-acid assignment across pairs; final ranking uses the original float64 energy. For m distinct proposed masks, expansion considers at most $m19^k$ assignments within the full fixed-order space of $\binom{L}{k}19^k$. This architecture makes binary sampling reusable across mutation orders while retaining the learned energy for sequence ranking.

4 EXPERIMENTS

4.1 BENCHMARKS AND EVALUATION

The primary zero-shot comparison evaluates 13 model families on 497,099 multi-site records: two double-mutant and eleven higher-order source–endpoint blocks. Our ten-source higher-order collection contains 306,014 structure-mapped records in 679 assay–condition groups (Appendix C.3). Thermodynamic endpoints and abundance proxies retain separate metric blocks; MegaScale (Tsuboyama et al., 2023) and ProteinGym (Notin et al., 2023) share provenance.

Each model supplies its own interaction-aware score. Scalar sequence models use single and four-state differences; conditional models average both updated-background paths; explicit Potts models use their energies. Complete-score sensitivity additionally evaluates full-sequence differences where defined (Appendix A.7). Native single-site results are reported separately over four source blocks.

Spearman rank correlation (SPCC) and source-native normalized discounted cumulative gain (NDCG) (Järvelin and Kekäläinen, 2002) measure ranking. The aggregate normalized rank $R = 100(13 - r)/12$, where r is within-block family rank, weights metric blocks equally. Supervised comparisons additionally use area under the precision–recall curve (AUPRC) (Davis and Goadrich, 2006). Component ablations use source-macro SPCC and an equal average across the single, double and higher-order groups.

4.2 State-of-the-art aggregate zero-shot prediction

PottsEnergy leads the multi-site comparison with $R = 85.58$, ahead of VenusREM (75.00) and PottsMPNN (70.51), and achieves the highest higher-order rank of 84.09 (Table 3). Its higher-order lead persists under complete-sequence scoring and inclusion of the MSA-trained PottsMPNN variant (Tables S24 and S9). VenusREM leads the double-mutant rank. These results establish the aggregate zero-shot advantage of the compact interaction model across the evaluated families.

Table 3: **Native single-site controls and matched multi-site zero-shot inference.** R is normalized within-block rank; larger is better.

	Native control	Model-specific second-order inference		
Model family	Single $R\uparrow$	Double $R\uparrow$	≥ 3 $R\uparrow$	Multi-site $R\uparrow$
PottsEnergy	**98.96**	93.75	**84.09**	**85.58**
VenusREM	67.71	**95.83**	71.21	75.00
PottsMPNN	84.38	77.08	69.32	70.51
StructureDCA[RSA]	45.83	54.17	65.91	64.10
ProteinMPNN	60.42	56.25	52.27	52.88
ESM-2	21.88	18.75	49.62	44.87
COORDinator	60.42	56.25	42.42	44.55
ESM-1v	6.25	6.25	48.48	41.99
ESM-IF1	66.67	66.67	36.36	41.03
SaProt	42.71	62.50	33.71	38.14
MIF-ST	30.21	29.17	36.74	35.58
MIF	57.29	31.25	27.65	28.21
EvoEF2	7.29	2.08	32.20	27.56

Note. All blocks contain 13 families and use $R = 100(13 - r)/12$, averaging ties. SPCC and source-native NDCG have equal block weight. The single column preserves four native-scoring controls; it is not re-inferred under the model-specific sequence scores defined in Appendix A.7. The matched columns cover two double and 11 higher-order blocks (497,099 rows/860 assays per family). Multi-site R excludes all four single blocks; no new across-order Overall R is claimed. PottsMPNN denotes the no-MSA checkpoint; the MSA-trained configuration is separate (Table S9). Bold marks the largest point estimate, not significance. VenusMutHub remains exploratory after 28 local configurations.

4.3 Couplings support prediction across mutation orders

We train nine configurations with seed 101 for 100,000 steps. The mechanism reference M11 uses the backbone encoder and explicit Potts heads, trained by native-context pseudolikelihood (PLL) alone. The full model in Section 3.3 additionally uses masked reconstruction and coupling regularization. The interventions share inputs and recorded batches, with their architecture-specific parameter counts. These single-seed comparisons use a common complete-conditional score (Appendix D). Single-, double- and higher-order SPCC macros average four, two and eleven source blocks, and the balanced score equally averages these three macros.

Removing couplings produces the largest balanced decrease among the five structural interventions: -0.0470, followed by disabling node updates (-0.0375). Replacing Potts heads with a three-layer ProteinMPNN-style decoder decreases all three order-specific scores; the balanced score falls by 0.0170 (Table 4). Node and edge updates contribute most strongly to singles and doubles; their removal raises the higher-order point estimates. The field-free variant slightly improves the balanced common score. This pattern identifies where each component contributes, with coupling removal degrading every mutation-order group.

Table 4: Component ablations of zero-shot stability prediction. Entries are changes in source-macro SPCC relative to the Potts plus pure-PLL reference M11; negative values indicate lower performance. All models use seed 101 and the fixed 100,000-step endpoint.

Intervention	Single	Double	≥ 3	Balanced
Remove couplings	$\mathbf{-0.0779}$	-0.0448	$\mathbf{-0.0183}$	$\mathbf{-0.0470}$
Disable node updates	-0.0703	$\mathbf{-0.0460}$	$+0.0036$	-0.0375
Replace Potts with a 3-layer decoder	-0.0128	-0.0202	-0.0180	-0.0170
Disable edge updates	-0.0502	-0.0099	$+0.0180$	-0.0140
Remove fields and field prior	-0.0049	-0.0021	$+0.0114$	$+0.0014$

The reference scores are 0.5262, 0.3784, 0.3887, 0.4311, in column order. The balanced score equally averages the three mutation-order macros. All columns use complete conditional scoring. Bold marks the largest observed decrease in each column, not significance. Node/edge interventions disable updates while retaining their inputs; the replacement decoder retains the encoder and full-context training. Parameter counts differ. Complete configurations, training alternatives and both scoring rules are reported in Appendix D.

Training alternatives are compared under both common and native scoring in Appendix D. Prefix-trained M10 leads the common balanced score (0.4349 versus M11's 0.4311), while M11 leads with native-energy scoring (0.4472 versus 0.4345). The tested four-sweep Gibbs negative-phase run reaches -0.0073 under the common score. The scoring interface therefore matters alongside the training objective.

At fixed M11 weights, retaining pair corrections improves double- and higher-order macro SPCC by 0.0296 and 0.0317 over matched single-site addition under native-energy scoring (Table S36; Appendix D.4).

4.4 Single-mutant adaptation

Single-mutant supervision provides an additional test of the representation. On 19,682 strict doubles from 20 proteins held out from stability fitting, PottsEnergy reaches SPCC 0.5969 with 1.052 million parameters, compared with 0.5967 for SPURS at 660.965 million (Table 5). SPURS has higher AUPRC. After adaptation, retaining pair corrections raises SPCC from 0.5247 to 0.5969 relative to matched additivity. Separate diagnostics distinguish this total-effect gain from epistasis correlation and sign accuracy, whose paired changes are unresolved (Table S40). Calibration and measured-single anchoring are detailed in Appendix E.

Table 5: Single-supervised prediction on 20 unseen proteins.

Model	Parameters (M)↓	SPCC↑	AUPRC↑
PottsEnergy	**1.052**	**0.5969**	0.2139
SPURS	660.965	0.5967	**0.2566**
ThermoMPNN	4.343	0.5322	0.0964
ProteinMPNN reg.	1.713	0.5331	0.1399
ESM-2 reg.	652.235	0.5484	0.0999

Retained cross-model results: 19,682 strict doubles, 20 proteins. Parameters include frozen encoders. Each baseline uses its own joint inference with single-trained weights; metrics are protein means followed by seed-metric averaging. AUPRC is defined for 10 proteins. Bold identifies point estimates, not significance. Full controls are in Table S4.

4.5 Efficient candidate retrieval with a reusable energy

We evaluate search using a four-model ensemble trained for 60 single-supervision epochs followed by 20 double-supervision epochs (seeds 43–46). Its averaged float64 tables define the ICM reference and frozen search objective. On a fixed random sample of ten of 298 MegaScale proteins, a physical coherent Ising machine (Inagaki et al., 2016) supplies position proposals, with 2,000 successful logical slots per protein and mutation order; each slot can return multiple states (Appendix C.11).

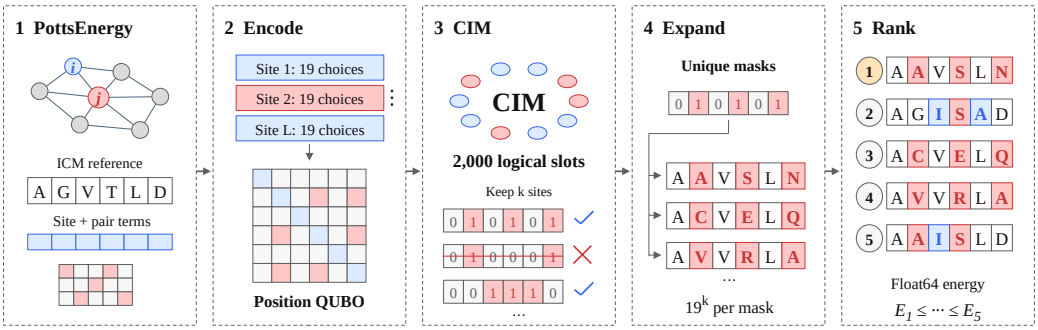


Figure 3: **From a frozen Potts energy to full-alphabet CIM candidates.** 1, The ICM reference anchors site and pair energy changes. 2, Minimizing amino-acid pair costs yields a position-selection QUBO with a mutation-count penalty. 3, Physical CIM receives the Ising matrix at 14-bit coefficient precision, yielding 2,000 successful logical result slots per protein and mutation order; decoded states retain exactly k selected sites. 4, Duplicate masks are removed. Each mask defines 19^k nonreference amino-acid assignments; conservative lower bounds skip masks that cannot enter the Top-5. 5, Candidates are ranked by the original float64 Potts energy. For $k > 2$, the pre-quantization position objective is a lower bound, not an exact full-alphabet sequence encoding. Colored residues mark substitutions; sequences and ranks illustrate $k = 3$, not measured results.

The pipeline recovers every frozen-energy Top-1 and Top-5 sequence for two- and three-site substitutions (Table 6). The exact references rank the same frozen energy over all $\binom{L}{k}19^k$ fixed-order sequences with rigorous pruning. At both orders, position proposals retain the highest-scoring complete sequences while pruning limits amino-acid expansion to a small set of masks.

Table 6: **CIM retrieval, relative computation count, and component time for two- and three-site mutations on 10 proteins.**

Mutated sites k	Recovered@1	Recovered@5	Recall@1 (%)	Recall@5 (%)	Relative computation count (×)	Component time (s)
2	1.00 [1, 1]	5.00 [5, 5]	100.0 [100, 100]	100.0 [100, 100]	475	7.443 [4.338, 8.938]
3	1.00 [1, 1]	5.00 [5, 5]	100.0 [100, 100]	100.0 [100, 100]	2,747	7.292 [5.708, 8.180]

Recovery metrics and component time are protein means [minimum, maximum] ($n = 10$), using 2,000 successful logical result slots per protein and order. Recovered@K counts identities shared with the exact Top-K; Recall@K divides this count by K. Relative computation count is the protein-mean ratio of naive fixed-order full-sequence evaluations to those performed after mask proposal and pruning: $\binom{L}{k}19^k/(m_k 19^k) = \binom{L}{k}/m_k$, where m_k is the number of retained masks; the displayed mean is rounded to the nearest integer. This ratio excludes position-objective construction and is not a runtime ratio. Component time sums measured matrix preparation and host processing with 2,000 platform single-evolution timing proxies (Appendix C.13). Full retrieval campaign: Appendix C.11.

Pruning leaves 3.9 and 21.8 masks on average for two- and three-site expansion. Relative to naive fixed-order enumeration, this reduces explicit full-alphabet sequence evaluations by mean factors of 475 and 2,747, respectively. Cached-energy matrix preparation and host processing average 0.195 and 0.217 seconds, with host replay pinned to one CPU core (Appendix C.13).

Figure 4 shows single-mutant saturation timing and a separate, nominated-pair double-mutant comparison using an assumed 10-ms CIM core (Appendix B.5). Table 6 instead reports full-alphabet candidate retrieval and component-time sums for two- and three-site mutations (Appendix C.13).

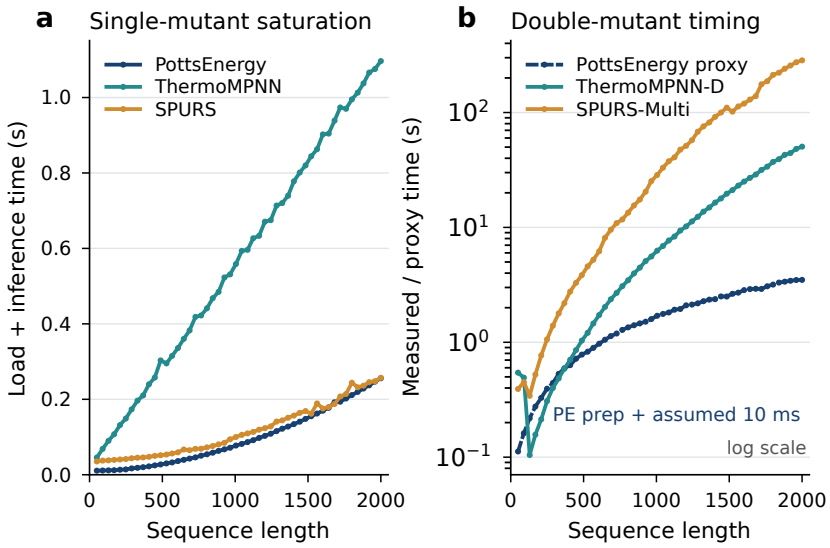


Figure 4: **Runtime scaling for single and double mutants. a**, Measured single-mutant saturation time. **b**, Measured exhaustive scoring of one nominated amino-acid pair per position pair versus PottsEnergy preparation plus an assumed 10-ms CIM core. Both panels use 50 length-selected proteins; the separate full-alphabet candidate-retrieval study reports two- and three-site component times in Table 6.

5 Discussion

The pairwise construction explains why sitewise reconstruction transfers to combination ranking: native-context couplings shape single-mutant scores, while target-pair cycles alter scores when both sites change. The observed correction gains depend on the dataset and endpoint; improved total-effect ranking does not imply more accurate interaction signs (Tables S5 and S40). The search study uses the same energy form but a separately supervised ensemble. Its position objective proposes masks, and full-alphabet expansion resolves amino-acid assignments before final ranking. Optional single-mutant adaptation further tests whether the compact representation accommodates experimental supervision. The frozen energy provides a common target for evaluating future quantum-assisted sampling interfaces.

6 Conclusion

PottsEnergy turns sitewise residue reconstruction into an explicit interaction energy for combinatorial stability scoring. It leads aggregate zero-shot ranking on the ten-source benchmark among the evaluated model families. After separate single- and double-mutant supervision, the same energy form supports candidate retrieval through position proposals, full-alphabet expansion and exact energy ranking. This shared representation connects stability prediction with efficient combinatorial search.

Reproducibility Statement

The Supplementary Material specifies data membership, model and scoring definitions, training protocols and search evaluation. Appendix D includes all nine component configurations and fixed-weight controls; Appendices C.11–C.13 specify the physical-CIM and cross-order search studies. The source package provides frozen numerical inputs, table generators and verification scripts. Experiment manifests bind data, checkpoints, commands and completion records.

AI Use Statement

AI tools were used only to assist literature searches, numerical verification of computational experiments, and language polishing of selected manuscript passages.

References

Beltran, A. et al. Site-saturation mutagenesis of 500 human protein domains. *Nature*, 637:885–894, 2025. doi:10.1038/s41586-024-08370-4.

Besag, J. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B*, 48:259–279, 1986. doi:10.1111/j.2517-6161.1986.tb01412.x.

Birnbaum, F. and Keating, A. E. Beyond native sequence recovery: Improved modeling of the sequence-energy landscape of protein structures. *Proceedings of the National Academy of Sciences*, 123:e2535494123, 2026. doi:10.1073/pnas.2535494123.

Blaabjerg, L. M. et al. Rapid protein stability prediction using deep learning representations. *eLife*, 12:e82593, 2023. doi:10.7554/eLife.82593.

Cianferoni, D., Vizarraga, D., Fernández-Escamilla, A. M., Fita, I., Hamdani, R., Reche, R., Delgado, J., and Serrano, L. Artificial intelligence and first-principle methods in protein redesign: A marriage of convenience? *Protein Science*, 34:e70210, 2025. doi:10.1002/pro.70210.

Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378:49–56, 2022. doi:10.1126/science.add2187.

Davis, J. and Goadrich, M. The relationship between Precision–Recall and ROC curves. In *Proceedings of ICML*, pp. 233–240, 2006. doi:10.1145/1143844.1143874.

Didi, K. et al. FLIP2: Expanding protein fitness landscape benchmarks for real-world machine learning applications. In *Proceedings of ICML*, 2026. URL https://icml.cc/virtual/2026/poster/63684.

Dieckhaus, H., Brocidiacono, M., Randolph, N. Z., and Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences*, 121:e2314853121, 2024. doi:10.1073/pnas.2314853121.

Dieckhaus, H. and Kuhlman, B. Protein stability models fail to capture epistatic interactions of double point mutations. *Protein Science*, 34:e70003, 2025. doi:10.1002/pro.70003.

Ekeberg, M. et al. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87:012707, 2013. doi:10.1103/physreve.87.012707.

Escobedo, A., Voigt, G., Faure, A. J., and Lehner, B. Genetics, energetics, and allostery in proteins with randomized cores and surfaces. *Science*, 389:eadq3948, 2025. doi:10.1126/science.adq3948.

Faure, A. J., Martí-Aranda, A., Hidalgo-Carcedo, C., Beltran, A., Schmiedel, J. M., and Lehner, B. The genetic architecture of protein stability. *Nature*, 634:995–1003, 2024. doi:10.1038/s41586-024-07966-0.

Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35:128–135, 2017. doi:10.1038/nbt.3769.

Hsu, C. et al. Learning inverse folding from millions of predicted structures. In *Proceedings of ICML*, pp. 8946–8970, 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

Huang, X., Pearce, R., and Zhang, Y. EvoEF2: Accurate and fast energy function for computational protein design. *Bioinformatics*, 36:1135–1142, 2020. doi:10.1093/bioinformatics/btz740.

Huber, P. J. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964. doi:10.1214/aoms/1177703732.

Inagaki, T. et al. A coherent Ising machine for 2000-node optimization problems. *Science*, 354:603–606, 2016. doi:10.1126/science.aah4243.

Järvelin, K. and Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20:422–446, 2002. doi:10.1145/582415.582418.

Kochenberger, G. et al. The unconstrained binary quadratic programming problem: A survey. *Journal of Combinatorial Optimization*, 28:58–81, 2014. doi:10.1007/s10878-014-9734-0.

Li, A. J. et al. Neural network-derived Potts models for structure-based protein design using backbone atomic coordinates and tertiary motifs. *Protein Science*, 32:e4554, 2023. doi:10.1002/pro.4554.

Li, G., Yao, S., and Fan, L. Shared-weight graph framework for comprehensive protein stability prediction across diverse mutation types. *Briefings in Bioinformatics*, 26:bbaf190, 2025. doi:10.1093/bib/bbaf190.

Li, W. and Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22:1658–1659, 2006. doi:10.1093/bioinformatics/btl158.

Li, Z. and Luo, Y. Generalizable and scalable protein stability prediction with rewired protein generative models. *Nature Communications*, 17:891, 2026. doi:10.1038/s41467-025-67609-4.

Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023. doi:10.1126/science.ade2574.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *Proceedings of ICLR*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405:442–451, 1975. doi:10.1016/0005-2795(75)90109-9.

Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html.

Montanucci, L. et al. DDGun: An untrained predictor of protein stability changes upon amino acid variants. *Nucleic Acids Research*, 50:W222–W227, 2022. doi:10.1093/nar/gkac325.

Musil, M. et al. FireProtDB 2.0: Large-scale manually curated database of the protein stability data. *Nucleic Acids Research*, 54:D409–D418, 2026. doi:10.1093/nar/gkaf1211.

Notin, P. et al. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023. doi:10.52202/075280-2810.

Olson, C. A., Wu, N. C., and Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24:2643–2651, 2014. doi:10.1016/j.cub.2014.09.072.

Otwinowski, J., McCandlish, D. M., and Plotkin, J. B. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115:E7550–E7558, 2018. doi:10.1073/pnas.1804015115.

Ouyang-Zhang, J., Diaz, D. J., Klivans, A. R., and Krähenbühl, P. Predicting a protein's stability under a million mutations. In *Advances in Neural Information Processing Systems*, 36:76229–76247, 2023. doi:10.52202/075280-3332.

Qin, T., Liu, T.-Y., and Li, H. A general approximation framework for direct optimization of information retrieval measures. *Information Retrieval*, 13:375–397, 2010. doi:10.1007/s10791-009-9124-x.

Reeves, S. and Kalyaanamoorthy, S. Robust multi-mutant protein stability prediction from a fine-tuned evolutionary scale model. *bioRxiv*, 2026. Preprint. doi:10.64898/2026.06.04.730231.

Shuai, R. W., Lu, T., Bhatti, S., Kouba, P., and Huang, P.-S. Ensemble-conditioned protein sequence design with Caliby. *bioRxiv*, 2025. Preprint. doi:10.1101/2025.09.30.679633.

Su, J. et al. SaProt: Protein language modeling with structure-aware vocabulary. In *Proceedings of ICLR*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1c42513b8895ab11fbbb5b7e8e6b6b02-Abstract-Conference.html.

Tan, Y., Wang, R., Wu, B., Hong, L., and Zhou, B. From high-throughput evaluation to wet-lab studies: advancing mutation effect prediction with a retrieval-enhanced model. *Bioinformatics*, 41(Suppl. 1):i401–i409, 2025. doi:10.1093/bioinformatics/btaf189.

Tsishyn, M., Talibart, H., Rooman, M., and Pucci, F. Structure-informed direct coupling analysis improves protein mutational landscape predictions. *bioRxiv*, 2026. Preprint. doi:10.64898/2026.03.27.714804.

Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620:434–444, 2023. doi:10.1038/s41586-023-06328-6.

Xavier, J. S. et al. ThermoMutDB: A thermodynamic database for missense mutations. *Nucleic Acids Research*, 49:D475–D479, 2021. doi:10.1093/nar/gkaa925.

Yang, K. K., Zanichelli, N., and Yeh, H. Masked inverse folding with sequence transfer for protein representation learning. *Protein Engineering, Design and Selection*, 36:gzad015, 2023. doi:10.1093/protein/gzad015.

Zhang, L. et al. VenusMutHub: A systematic evaluation of protein mutation effect predictors on small-scale experimental data. *Acta Pharmaceutica Sinica B*, 15:2454–2467, 2025. doi:10.1016/j.apsb.2025.03.028.