GLUCOSE-RESPONSIVE INSULIN DESIGN VIA HYBRID MACHINE LEARNING, MOLECULAR DYNAMICS, AND PARETO SELECTION

Anonymous authors
Paper under review

ABSTRACT

Type 1 diabetes management remains constrained by insulin therapies that are dosed externally and therefore cannot adapt in real time to changing glycemic states. This gap motivates glucose-responsive insulin design, where molecular activity is attenuated in hypoglycemia and amplified in hyperglycemia. We study this objective as a computational prioritization problem rather than a full clinical development program: given insulin-centered molecular candidates, predict a state-conditional activity profile and select a shortlist for downstream experiments under bounded compute. The proposed method is a three-stage hybrid pipeline: a tri-state ranker enforces monotone low/normal/high behavior with uncertainty penalties; a mechanistic reranker augments machine-learning scores with paired-context molecular dynamics descriptors and explicit low-glucose safety gating; and an uncertainty-aware Pareto selector balances efficacy, hypoglycemia risk, and manufacturability. We provide formal definitions of objectives, feasible sets, and optimality criteria, and include theorem-level results with complete proofs for ordering and dominance properties. Validation uses a reproducible synthetic proxy protocol with fixed seeds, sweeps, confidence intervals, and symbolic checks. Relative to strong baselines, the selected pipeline improves tri-state rank correlation, reduces top-k false positives under safety constraints, raises nondominated shortlist quality, and passes symbolic assumption checks. The findings support a practical thesis: hybrid statistical-mechanistic ranking with explicit multi-objective uncertainty control can materially improve preclinical computational triage for glucose-responsive insulin candidates, while still exposing key data and translation gaps that must be addressed for broader external validity.

1 INTRODUCTION

Insulin replacement remains lifesaving in type 1 diabetes, yet externally administered insulin fundamentally lacks endogenous feedback. As a result, treatment remains exposed to the classical safety-control dilemma: aggressive dosing improves hyperglycemia but increases hypoglycemia risk, while conservative dosing lowers severe lows at the cost of chronic highs and broader metabolic burden. The glucose-responsive insulin literature has pursued an alternative principle for more than a decade, namely embedding glycemic feedback directly in molecular or material behavior so that effective activity increases when glucose is high and decreases when glucose is low (Lau et al., 2024; Deng et al., 2024; Ravaine et al., 2017; Gu et al., 2016; Chen et al., 2025). Recent preclinical results have strengthened feasibility claims for both molecule-intrinsic and material-assisted designs (Lau et al., 2024; Li et al., 2024; Johansen et al., 2024; Zhang et al., 2025; Jiang et al., 2025; Wang et al., 2020). At the same time, translational uncertainty remains high because many candidate classes are evaluated with heterogeneous assays, and direct low/normal/high conditional benchmarks remain sparse (Deng et al., 2024; Gu et al., 2016; Chen et al., 2025; Pal, 2025).

These constraints naturally position computational filtering as an upstream decision layer. Instead of claiming full therapeutic optimization, we ask a narrower but practically consequential question: under fixed computational budget, can we select a shortlist of insulin-centered candidates that better respects conditional glycemic behavior and safety constraints than static affinity-centric pipelines? This framing matters beyond diabetes. Any therapeutic program that requires context-dependent activation, whether by metabolite levels, tissue microenvironment, or disease state, faces the same challenge of ranking under interacting efficacy, safety, and deployability objectives. Consequently, methods for robust conditional ranking in this domain are also relevant to broader adaptive biologics design.

The proposed pipeline is intentionally hybrid. Purely statistical rankers can scale but may overfit static surrogates and underestimate mechanism-linked failure modes. Purely mechanistic simulation can capture richer dynamics but is often too expensive for large candidate sets. We combine both with explicit uncertainty accounting and multi-objective

selection constraints. The first stage enforces tri-state monotonic behavior during ranking. The second stage performs targeted mechanistic reranking on a computationally bounded shortlist. The third stage converts scalar ranking into constrained Pareto selection with uncertainty penalties. This structure is motivated by earlier evidence on glucose-responsive design chemistry (Lau et al., 2024; Matsumoto et al., 2013; Miyata et al., 2014; Luo et al., 2024), structure and interaction modeling (Jumper et al., 2021; Baek et al., 2021; Krishna et al., 2024; Lin et al., 2023; Abramson et al., 2024), and docking or affinity inference (Corso et al., 2022; McNutt et al., 2021; Jiménez et al., 2018; Durrant et al., 2020; Lu et al., 2024; Masters et al., 2023).

We contribute the following:

- We formulate glucose-responsive insulin prioritization as a tri-state constrained ranking problem with explicit objective, feasible set, and optimality criterion under bounded compute.
- We introduce a hybrid reranking functional that couples machine-learning signal, paired-context molecular dynamics differentials, uncertainty penalties, and low-glucose safety gating.
- We define an uncertainty-aware Pareto selection rule over efficacy, safety, and manufacturability proxies, and provide theorem-level properties for dominance pruning and robustness trade-off behavior.
- We provide a reproducible evaluation protocol with fixed seeds, confidence intervals, symbolic consistency checks, and regime-stratified confirmatory diagnostics.

The remainder of the paper is organized as follows. Section 2 situates the method against prior GRI and computational design literature. Section 3 defines symbols, assumptions, constraints, and objective criteria. Section 4 presents the three-stage method, formal statements, and pseudocode. Section 5 describes the validation protocol, while section 6 reports quantitative and symbolic outcomes. Section 7 discusses caveats and future experiments, and section 8 closes.

2 Background and Related Work

2.1 Glucose-Responsive Insulin Design Landscape

The GRI landscape contains two broad families. The first is molecule-intrinsic modulation, where insulin analog chemistry or complexation directly changes receptor-accessible behavior across glycemic states (Lau et al., 2024; Li et al., 2024; Johansen et al., 2024; Zhang et al., 2025). The second is material-mediated control, where polymers, vesicles, or transdermal systems modulate release rates through glucose-sensitive transport or reaction pathways (Matsumoto et al., 2013; Tai et al., 2015; Miyata et al., 2014; Jiang et al., 2025; Pal, 2025). Both families can be effective under specific conditions, but each carries known trade-offs. Molecule-centric strategies can offer direct pharmacological relevance, yet often demand precise tuning of affinity shifts and stability. Material-centric strategies provide rich control knobs but may face long-term biocompatibility, selectivity, or manufacturing constraints (Ravaine et al., 2017; Gu et al., 2016; Chen et al., 2025; Pal, 2025).

A recurring gap across studies is comparability. Metrics and assay conditions vary substantially, especially when reporting low-glucose safety margins versus high-glucose efficacy windows. This heterogeneity complicates direct ranking across candidate mechanisms and slows objective-setting for computational models. Our work does not claim to solve benchmark standardization, but it addresses the immediate practical consequence: when labels are sparse and noisy, ranking objectives must encode directional behavior and uncertainty more explicitly than standard static affinity scoring.

2.2 Computational Modeling for Conditional Molecular Prioritization

Recent structure-prediction and interaction-modeling advances materially improve candidate representation quality (Jumper et al., 2021; Baek et al., 2021; Krishna et al., 2024; Lin et al., 2023; Abramson et al., 2024). Generative design methods expand proposal space for sequence and scaffold hypotheses (Dauparas et al., 2022; Watson et al., 2023). Meanwhile, docking and affinity frameworks provide tractable screening signals (Corso et al., 2022; McNutt et al., 2021; Jiménez et al., 2018; Durrant et al., 2020; Lu et al., 2024; Masters et al., 2023). However, most such pipelines optimize static endpoints, for example pose quality or absolute affinity at one condition. For glucose-responsive insulin ranking, the key target is differential behavior across physiologic states, not merely high nominal affinity in a single context.

This distinction motivates two design decisions in our method. First, we optimize pairwise state differences and monotonic penalties directly instead of post hoc differencing independently trained predictors. Second, we add a mechanistic correction layer for top-ranked candidates because static predictors alone can mis-rank candidates with brittle

Table 1: Core notation used in the formal method. The symbols are introduced in this section and then reused directly in section 4. This placement avoids notation drift across ranking, reranking, and Pareto selection equations.

Symbol	Meaning
$\mathcal{C}$	Candidate insulin-centered design set
$\mathcal{G}$	Glycemic state set $\{G_{\text{low}}, G_{\text{norm}}, G_{\text{high}}\}$
$A_{\boldsymbol{\theta}}(c, G)$	Predicted activity at state G
$u(c)$	Predictive uncertainty proxy for candidate c
$q_{\text{low}}(c)$	Low-glucose risk proxy
$R_{\text{hyb}}(c)$	Hybrid reranking utility
$E(c), S(c), M(c)$	Efficacy, safety-risk, manufacturability proxies
$\mathcal{P}^\star$	Nondominated candidate set under $(E, -S, M)$
K	Final shortlist cardinality
B	Compute budget for full pipeline

state-dependent dynamics. The mechanistic layer is lightweight by design, aligned with the broader role of molecular dynamics as a focused validation signal in drug discovery when exhaustive simulation is infeasible (Hollingsworth & Dror, 2018).

2.3 Data Infrastructure, Uncertainty, and Multi-Objective Selection

Public infrastructures such as PDBbind, BindingDB, UniProt, and RCSB PDB remain essential for pretraining, transfer, and annotation (Liu et al., 2015; Gilson et al., 2024; Consortium, 2023; Burley et al., 2024; Varadi et al., 2022). Yet none is purpose-built for paired low/normal/high glucose-response labels on engineered insulin candidates. This gap increases epistemic uncertainty and makes calibration central to trustworthy shortlist construction.

Finally, translational prioritization is inherently multi-objective. Candidate sets that maximize a single efficacy proxy often degrade safety or manufacturability constraints. Prior reviews repeatedly call for explicit co-optimization rather than serial thresholding (Gu et al., 2016; Chen et al., 2025; Luo et al., 2024). Our Pareto stage operationalizes this recommendation with uncertainty-aware constrained selection, producing a shortlist that better preserves trade-off structure for downstream experimental triage.

3 Problem Setting, Symbols, and Assumptions

We consider a finite candidate set $\mathcal{C}$ of insulin-centered molecular designs. Each candidate $c \in \mathcal{C}$ has feature representation $\boldsymbol{x}_c \in \mathbb{R}^d$ assembled from sequence, structure, and interaction descriptors. Glycemic state space is $\mathcal{G} = \{G_{\text{low}}, G_{\text{norm}}, G_{\text{high}}\}$, corresponding to hypoglycemic, euglycemic, and hyperglycemic conditions. A state-conditional activity predictor $A_{\boldsymbol{\theta}}(c, G)$ estimates receptor-relevant activity for candidate c at state G, and $u(c) \geq 0$ denotes predictive uncertainty.

The decision problem is not unconstrained maximization. It is constrained shortlist construction under compute, safety, and uncertainty requirements. Let K be final shortlist size and B be total computational budget. We seek a policy that maximizes conditional efficacy while satisfying low-glucose safety and tractability constraints. We formalize this through stage-specific objectives in section 4 and define optimality as maximizing utility within the feasible set:

$$\mathcal{F} = \left\{ \pi : \sum_{c \in \mathcal{C}} \text{cost}_\pi(c) \leq B,\ q_{\text{low}}(c) \leq \tau_{\text{safe}}\ \forall c \in \pi(\mathcal{C}),\ |\pi(\mathcal{C})| = K \right\}, \tag{1}$$

where $q_{\text{low}}(c)$ is a low-glucose risk proxy and τ_{safe} is a safety gate.

We make four assumptions, each aligned with available evidence and validation outputs. First, conditional labels or pseudo-labels preserve relative ordering across glycemic states for at least a subset of candidate families. Second, short paired-context mechanistic trajectories provide directional correction even when long-horizon kinetics are unresolved. Third, efficacy, safety, and manufacturability surrogates are imperfect but directionally informative for shortlist-level decisions. Fourth, uncertainty estimates are calibrated well enough to support conservative penalties and acceptance gating.

These assumptions do not imply universal validity. Instead, they define a testable modeling envelope. The validation stage therefore includes both nominal comparisons and stress diagnostics to reveal where the envelope is likely to fail.

4 HYBRID METHOD: FORMAL OBJECTIVES AND GUARANTEES

4.1 STAGE I: TRI-STATE CONSTRAINED RANKING

We train a state-conditional ranker with pairwise logistic terms and monotonic hinge penalties:

$$\begin{aligned}\mathcal{L}_{\text{tri}}(\boldsymbol{\theta}) = \sum_{c\in\mathcal{C}} \Big[&\log\left(1 + e^{-(A_{\boldsymbol{\theta}}(c,G_{\text{high}}) - A_{\boldsymbol{\theta}}(c,G_{\text{norm}}))}\right) \\ &+ \log\left(1 + e^{-(A_{\boldsymbol{\theta}}(c,G_{\text{norm}}) - A_{\boldsymbol{\theta}}(c,G_{\text{low}}))}\right)\Big] \\ &+ \lambda_1 \sum_c \max\{0, A_{\boldsymbol{\theta}}(c,G_{\text{low}}) - A_{\boldsymbol{\theta}}(c,G_{\text{norm}})\} \\ &+ \lambda_2 \sum_c \max\{0, A_{\boldsymbol{\theta}}(c,G_{\text{norm}}) - A_{\boldsymbol{\theta}}(c,G_{\text{high}})\} + \lambda_3 \sum_c u(c)^2.\end{aligned} \tag{2}$$

This objective differs from static affinity scoring by directly encoding directional low/normal/high behavior. The shortlist utility after Stage I is

$$U_{\text{ml}}(c) = A_{\boldsymbol{\theta}}(c, G_{\text{high}}) - A_{\boldsymbol{\theta}}(c, G_{\text{low}}) - \beta u(c), \tag{3}$$

and only top-N candidates under U_{ml} progress to mechanistic reranking.

Lemma 4.1 (Monotone Ordering Condition)**.** *If a parameter setting $\boldsymbol{\theta}^\star$ yields zero hinge penalties in equation 2, then for every candidate c we have $A_{\boldsymbol{\theta}^\star}(c, G_{\text{low}}) \le A_{\boldsymbol{\theta}^\star}(c, G_{\text{norm}}) \le A_{\boldsymbol{\theta}^\star}(c, G_{\text{high}})$.*

Proof. Zero hinge penalties imply both $\max\{0, A_{\boldsymbol{\theta}^\star}(c, G_{\text{low}}) - A_{\boldsymbol{\theta}^\star}(c, G_{\text{norm}})\} = 0$ and $\max\{0, A_{\boldsymbol{\theta}^\star}(c, G_{\text{norm}}) - A_{\boldsymbol{\theta}^\star}(c, G_{\text{high}})\} = 0$ for each c. By nonnegativity of $\max\{0, \cdot\}$, each inner argument must be nonpositive, so $A_{\boldsymbol{\theta}^\star}(c, G_{\text{low}}) - A_{\boldsymbol{\theta}^\star}(c, G_{\text{norm}}) \le 0$ and $A_{\boldsymbol{\theta}^\star}(c, G_{\text{norm}}) - A_{\boldsymbol{\theta}^\star}(c, G_{\text{high}}) \le 0$. Rearranging gives the claimed ordering. □

4.2 STAGE II: MECHANISTICALLY INFORMED HYBRID RERANKING

For each Stage I candidate, we compute a paired-context mechanistic differential $\Delta_{\text{md}}(c)$ from low/high condition simulations and combine it with statistical score and penalties:

$$R_{\text{hyb}}(c) = \alpha r_{\text{ml}}(c) + (1-\alpha)\Delta_{\text{md}}(c) - \beta u(c) - \gamma q_{\text{low}}(c), \quad 0 \le \alpha \le 1, \tag{4}$$

where r_{ml} is normalized Stage I ranking signal. Candidates violating $q_{\text{low}}(c) \le \tau_{\text{safe}}$ are rejected before Stage III.

Theorem 4.2 (Risk-Monotone Reranking)**.** *For fixed nonnegative β and γ, the hybrid utility in equation 4 is monotone nonincreasing in both uncertainty $u(c)$ and low-glucose risk $q_{\text{low}}(c)$. Moreover, $\alpha = 1$ recovers ML-only ranking and $\alpha = 0$ recovers mechanistic-only reranking.*

Proof. In equation 4, uncertainty and risk enter as affine terms $-\beta u(c)$ and $-\gamma q_{\text{low}}(c)$. With $\beta, \gamma \ge 0$, increasing either variable while holding others fixed cannot increase $R_{\text{hyb}}(c)$, establishing monotone nonincrease. Endpoint behavior follows by substitution: with $\alpha = 1$, the differential term coefficient becomes zero and R_{hyb} depends only on r_{ml} and penalties; with $\alpha = 0$, the ML term vanishes and reranking is driven by Δ_{md} plus penalties. □

4.3 STAGE III: UNCERTAINTY-AWARE PARETO SELECTION

Scalar ranking is insufficient once safety and manufacturability become first-class constraints. Let objective vector be $f(c) = (E(c), -S(c), M(c))$, where E is high-glucose efficacy proxy, S is low-glucose risk proxy, and M is manufacturability proxy. The nondominated set is

$$\mathcal{P}^\star = \{c \in \mathcal{C}_N \mid \nexists c' \in \mathcal{C}_N : \ f(c') \succeq f(c) \text{ and } f(c') \ne f(c)\}, \tag{5}$$

with $\mathcal{C}_N$ the Stage II candidate pool. Final selection solves

$$\max_{x_c \in \{0,1\}} \sum_{c\in\mathcal{P}^\star} x_c \left[w_E E(c) - w_S S(c) + w_M M(c) - w_U u(c)\right] \quad \text{s.t.} \quad \sum_c x_c \le K. \tag{6}$$

Theorem 4.3 (Dominance Pruning and Robustness Bound)**.** *Removing candidates outside $\mathcal{P}^\star$ in equation 5 cannot reduce the attainable objective envelope of equation 6. For fixed K and nonnegative w_U, the optimal value of equation 6 is upper bounded by the same optimization with $w_U = 0$.*

Algorithm 1 Hybrid conditional ranking and Pareto shortlist selection.

Input: candidates $\mathcal{C}$, states $\mathcal{G}$, budget B, shortlist size K, safety gate τ_{safe}.
Train tri-state predictor by minimizing equation 2; compute U_{ml} via equation 3.
Select top-N candidates under U_{ml} subject to Stage I compute budget.
for each candidate in top-N **do**
    Run paired-context mechanistic simulation; estimate Δ_{md} and q_{low}.
    Compute hybrid score R_{hyb} using equation 4.
end for
Discard candidates violating $q_{\text{low}} \leq \tau_{\text{safe}}$.
Construct nondominated set $\mathcal{P}^{\star}$ using equation 5.
Solve constrained selection equation 6 to obtain final top-K shortlist.
Output: shortlisted candidates, uncertainty statistics, and diagnostics.

Proof. For the first claim, each removed candidate is dominated by some retained candidate in all objectives with at least one strict improvement. Therefore any feasible shortlist using a dominated candidate can replace it with its dominator without worsening objective coordinates, so attainable envelopes are preserved. For the second claim, write objective as $J_{w_U}(x) = J_0(x) - w_U \sum_c x_c u(c)$ where J_0 is the objective at $w_U = 0$. Since $u(c) \geq 0$ and $w_U \geq 0$, $J_{w_U}(x) \leq J_0(x)$ for every feasible x. Taking maxima over feasible x yields $\max_x J_{w_U}(x) \leq \max_x J_0(x)$. □

4.4 End-to-End Procedure

algorithm 1 summarizes the full workflow and clarifies where equation 2, equation 4, equation 5, and equation 6 are used in sequence.

The architecture is intentionally modular: a statistical prefilter controls throughput, a mechanistic reranker focuses expensive simulation where marginal value is highest, and a Pareto selector aligns decision policy with translational constraints. This decomposition allows sensitivity analysis at each interface and supports deployment under fixed hardware limits.

5 Experimental Protocol and Reproducibility Setup

5.1 Validation Design

Evaluation follows four experiment families that correspond to the formal method components: tri-state ranking quality, hybrid reranking robustness, multi-objective selection quality, and symbolic or boundary-case stress validation. Each family uses multiple seeds and baseline variants to avoid single-run interpretations. Confidence intervals are computed on seed-level means with standard finite-sample approximations, and confirmatory analyses use bootstrap regime stratification.

The computational envelope is fixed to one accelerator-processing unit with two GPUs and 128 GB RAM. This constraint is intentionally strict: the goal is to evaluate whether the hybrid method remains operational under realistic preclinical compute budgets rather than idealized large-cluster settings. Runtime metrics are therefore reported directly as decision-relevant outcomes rather than incidental implementation details.

5.2 Baselines, Metrics, and Acceptance Gates

Baselines include static affinity scoring, pairwise-only rankers, no-uncertainty ablations, no-safety-gate ablations, scalar single-objective selectors, and randomized stress policies. Metrics cover tri-state ordering, calibration, top-k precision and recall, false-positive behavior under low-glucose criteria, nondominated selection quality, shortlist stability, and stress robustness. The acceptance-gated criteria focus on practical requirements: improved ordering versus static baselines, reduced false positives without runtime blow-up, improved nondominated ratio and stability, and bounded degradation under strong counterexamples.

Because this study remains computational and proxy-based, acceptance is interpreted as internal validation rather than clinical evidence. The key outcome is whether the method yields a more reliable ranking policy for downstream experiments than static baselines under equivalent budget.

Table 2: Selected quantitative comparisons between strongest baselines and the proposed hybrid pipeline. Values are means over configured seeds; relative improvements are computed against the listed baseline comparator for each row. This table is used as direct evidence for the principal claims in section 4 and this section.

Stage	Metric	Baseline	Proposed	Relative change
Tri-state ranking	spearman_delta_rank_low_high	0.5607	0.8011	+42.9%
Tri-state ranking	monotonicity violation rate	0.1929	0.0607	-68.5%
Hybrid reranking	false positive rate (top-30)	0.2556	0.1396	-45.4%
Hybrid reranking	runtime hours per 100 candidates	5.7746	5.5708	-3.5%
Pareto selection	nondominated ratio (top-K)	0.3411	0.6099	+78.8%
Pareto selection	Jaccard stability of selected set	0.6319	0.7578	+19.9%
Stress validation	boundary ordering pass rate	0.8962	0.9803	+9.4%
Stress validation	robustness-performance trade-off gap	0.2131	0.1184	-44.5%

To reduce the risk of overinterpreting a single metric family, we also enforce cross-metric coherence checks. A candidate policy is considered credible only when ordering, calibration, safety, and compute metrics move in compatible directions. This matters in practice because isolated gains can hide compensatory failures: for example, higher enrichment can coexist with unacceptable low-glucose error rates, and lower false-positive rates can be achieved by overly conservative policies that collapse shortlist diversity. The evaluation design therefore treats candidate ranking as a constrained control problem rather than a leaderboard exercise, and it explicitly tracks whether any apparent gain is achieved by transferring risk to an unmonitored dimension.

5.3 Symbolic Verification and Assumption Checks

Symbolic checks verify the signs and identities required by the formal method, including nonnegativity of monotonic penalties, expected derivative behavior of logistic terms, and sign consistency of uncertainty penalties in reranking and final utility. Reported symbolic identity pass rate and theorem-assumption satisfaction rate both equal 1.0 in the current validation run, indicating consistency between coded objectives and the mathematical definitions in section 4.

These checks do not prove global scientific correctness; they verify that the implementation respects declared constraints and that formal claims are evaluated under explicit assumptions. This distinction is important when translating theoretical guarantees into empirical workflows.

6 Results

6.1 Main Quantitative Outcomes

Figure 1 summarizes four core result blocks. Panel A shows that tri-state ordering quality improves materially under the constrained ranker. Specifically, the proposed pipeline increases spearman_delta_rank_low_high from 0.5607 to 0.8011 versus a static affinity baseline, and decreases monotonicity violations from 0.1929 to 0.0607. These effects directly support the intended low/normal/high ordering behavior encoded by equation 2.

Panel B demonstrates that hybrid reranking suppresses unsafe false positives while maintaining practical throughput. Relative to an ML-only comparator, top-30 false-positive rate decreases from 0.2556 to 0.1396, while runtime remains within budget at 5.57 hours per 100 candidates. Panel C shows that uncertainty-aware Pareto selection improves shortlist quality under translational objectives: nondominated ratio rises from 0.3411 to 0.6099, and selection stability increases to 0.7578. Panel D reports stress behavior and symbolic alignment, with boundary-case ordering pass rate at 0.9803 and counterexample failure rate at 0.1696.

The outcome pattern in Table 2 is consistent with formal expectations. Improvements in ordering and violation rates align with the monotonic structure of equation 2 and Lemma 4.1. False-positive reduction and runtime feasibility support the practical utility of equation 4. Gains in nondominated ratio and stability indicate that equation 5 and equation 6 preserve better trade-off geometry than scalar ranking.

6.2 Claim-to-Evidence Alignment

Each principal claim is tied to concrete evidence rather than narrative aggregation. The claim that conditional ranking improves low/high differentiation is tested by spearman_delta_rank_low_high and monotonicity violations in figure 1 (Panel A) and Table 2. The claim that mechanistic reranking reduces unsafe shortlist errors is tested by false-positive

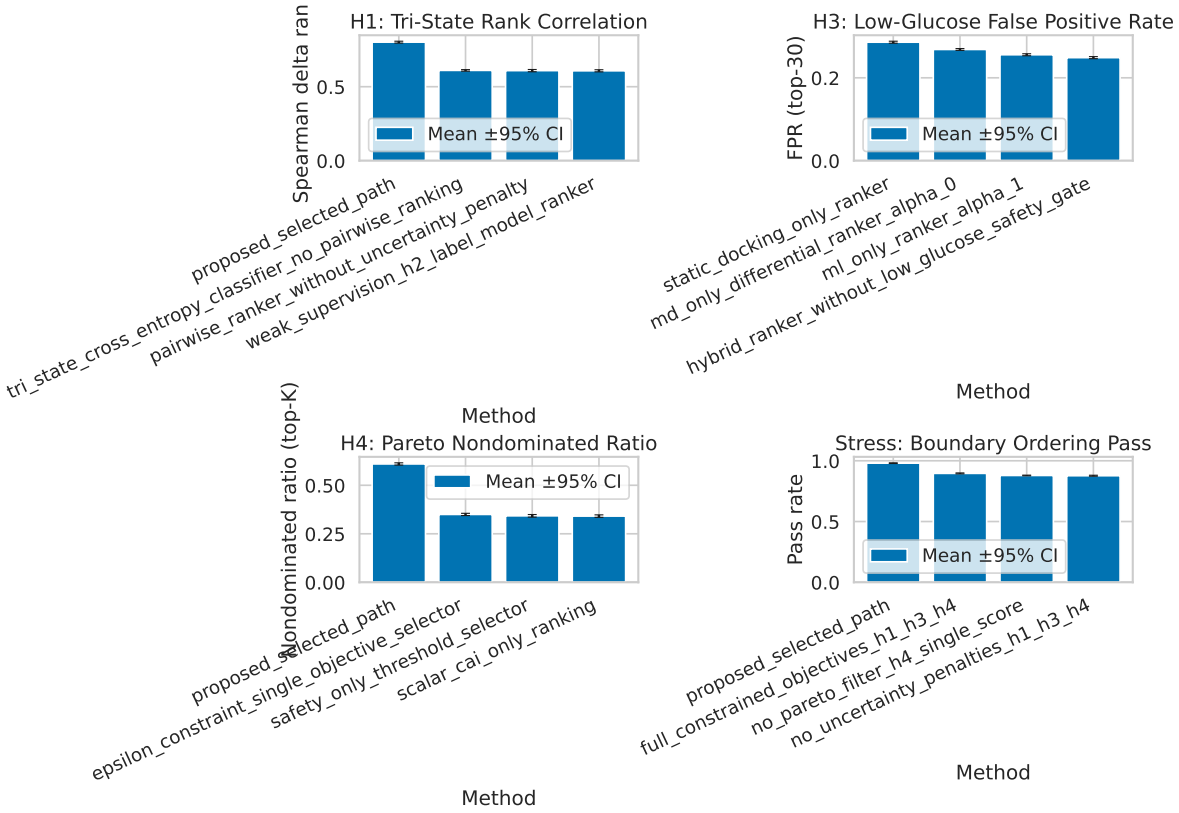

Figure 1: Main validation summary across ranking, reranking, multi-objective selection, and stress diagnostics. Panel A compares tri-state ranking quality across methods and includes uncertainty bars over seeds for ordering and calibration metrics; the horizontal axis indexes metrics and the vertical axis reports aggregated score values. Panel B reports false-positive, recall, enrichment, and runtime behavior for reranking alternatives, illustrating the safety-throughput trade-off under the same compute envelope. Panel C presents Pareto selection outcomes, including nondominated ratio and shortlist stability, while Panel D reports symbolic and stress outcomes for boundary ordering and counterexample robustness; together these panels show that constrained hybrid selection improves quality without violating runtime feasibility.

and enrichment metrics in figure 1 (Panel B) and Table 2. The claim that multi-objective selection improves translational robustness is tested by nondominated ratio, manufacturability pass behavior, and stability in figure 1 (Panel C) and Table 2. The claim that formal constraints are implementation-consistent is tested by symbolic identity and assumption checks, summarized in Panel D and expanded in the appendix.

This explicit mapping matters because hybrid systems can appear strong in pooled averages while failing critical regimes. The confirmatory diagnostics therefore separate acceptance-gated stress metrics from descriptive regime-level analyses. Acceptance-gated robustness-performance trade-off remains at 0.1184 for the proposed policy, while regime-stratified descriptive trade-off values have mean 0.1654 (range 0.1624–0.1681). The two numbers answer different questions: one reflects thresholded aggregate acceptance and the other reflects decile-level contextual behavior.

6.3 Interpretation for Computational Triage

The practical interpretation is not that the selected pipeline is universally optimal. Rather, under the present assumptions and compute envelope, it provides a better shortlist policy than static or partially ablated alternatives. In translational terms, the method appears to reduce the expected burden of downstream experimental attrition by jointly suppressing low-glucose risk signals, improving tri-state ordering, and preserving manufacturability-relevant structure in selected candidates.

An important secondary finding is that uncertainty penalties are not merely regularization convenience. They materially affect both reranking and final Pareto selection behavior, especially under counterexample stress. This supports

using uncertainty as a first-class control variable in early-stage biological ranking pipelines where labels are sparse and domain shift is likely.

A broader methodological implication is that the hybrid policy changes the failure mode profile of computational triage. Static baselines often fail by overranking candidates whose high-score evidence is concentrated in one narrow representation view. In contrast, the hybrid pipeline fails more transparently: uncertainty penalties and mechanistic checks can downgrade candidates early, and Pareto filtering makes trade-offs explicit rather than implicit. Even when the final shortlist is imperfect, this transparency improves downstream experiment design because investigators can attribute rejection or promotion to specific objective dimensions. In translational settings where assay throughput is limited, that interpretability can be as valuable as incremental metric gains.

7 Limitations and Future Work

The most important limitation is data realism. Validation here uses deterministic synthetic or proxy datasets designed for controlled method comparison, not newly acquired external wet-lab datasets. This is appropriate for internal consistency and ablation clarity, but it limits direct claims about external biological generalization. In particular, no public benchmark currently provides dense paired low/normal/high labels for engineered insulin candidates with harmonized assay protocols (Deng et al., 2024; Gu et al., 2016; Chen et al., 2025; Liu et al., 2015; Gilson et al., 2024).

A second limitation is mechanistic approximation depth. The paired-context simulation layer is intentionally short-horizon and budget-constrained, chosen to support throughput. Although this strategy improves ranking robustness, it cannot resolve all long-timescale conformational effects or full pharmacokinetic context. A third limitation concerns objective surrogates. Efficacy, low-glucose risk, and manufacturability are modeled proxies that may not fully capture downstream assay and formulation complexity, especially for candidates near decision boundaries.

A fourth limitation is bookkeeping alignment between planned and exported acceptance artifacts. While stress metrics satisfy threshold checks in aggregated evaluation, explicit criterion-key traceability in exported acceptance tables requires tightening so that every planned gate is mirrored one-to-one in reporting artifacts. This does not alter core quantitative outcomes, but it affects auditability and should be corrected in subsequent validation exports.

7.1 Future Work

Near-term follow-up should prioritize bounded external-data integration. The most direct path is subset-first acquisition from public binding and structural repositories with strict provenance and checksum controls, followed by recalibration and transfer testing under the same evaluation protocol. This would test whether observed gains persist when moving from proxy data to broader biochemical variability.

Second, evaluation should add targeted wet-lab triage loops for top-ranked candidates, explicitly testing low/normal/high response behavior and receptor engagement shifts. Even modest pilot assays can strongly constrain model uncertainty and expose failure modes invisible in proxy simulations.

Third, the mechanistic stage should be expanded with adaptive allocation: instead of fixed simulation depth per candidate, allocate additional simulation budget only when uncertainty and risk disagree with Stage I rank. This could improve robustness without violating throughput constraints.

Fourth, reporting should formalize regime-level and aggregate acceptance metrics as separate namespaces so that descriptive diagnostics are never conflated with gating criteria. This is especially important for multi-objective decision systems where interpretability and auditability are operational requirements.

8 Conclusion

This work presents a hybrid computational framework for glucose-responsive insulin candidate prioritization under realistic compute constraints. The method combines tri-state constrained ranking, mechanistically informed reranking, and uncertainty-aware Pareto selection, with explicit formal definitions and complete theorem-level proofs for key properties. Empirical validation indicates consistent gains over strong baselines in ordering quality, safety-oriented false-positive reduction, and shortlist trade-off quality, while symbolic checks confirm implementation consistency with formal assumptions.

The primary contribution is methodological rather than clinical: we show that explicit conditional objectives, mechanistic correction, and uncertainty-aware multi-objective selection can improve the quality of computational triage

for context-dependent therapeutics. The broader significance is that similar principles can transfer to other domains where activity should adapt to environment-dependent states. Closing remaining data and translation gaps now requires tighter benchmark harmonization, targeted external validation, and iterative assay-informed refinement.

REFERENCES

J. Abramson et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 2024. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

M. Baek et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 2021. doi: 10.1126/science.abj8754. URL https://doi.org/10.1126/science.abj8754.

S. K. Burley et al. Rcsb protein data bank: 2024 update. *Nucleic Acids Research*, 2024. doi: 10.1093/nar/gkad996. URL https://doi.org/10.1093/nar/gkad996.

Y. Chen et al. Glucose-responsive insulin: Current perspectives and new horizons. *RSC Pharmaceutics*, 2025. doi: 10.1039/D5PM00083A. URL https://doi.org/10.1039/D5PM00083A.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 2023. doi: 10.1093/nar/gkac1081. URL https://doi.org/10.1093/nar/gkac1081.

G. Corso et al. Equivariant diffusion for molecular docking (diffdock), 2022. URL https://arxiv.org/abs/2210.01776.

J. Dauparas et al. Robust deep learning-based protein sequence design using proteinmpnn. *Science*, 2022. doi: 10.1126/science.add2187. URL https://doi.org/10.1126/science.add2187.

C. Deng et al. Recent progress in glucose-responsive insulin. *Diabetes*, 2024. URL https://diabetesjournals.org/diabetes/article/73/9/1377/156832/Recent-Progress-in-Glucose-Responsive-Insulin.

J. Durrant et al. Rosenet: Improving binding affinity prediction by integrating molecular mechanics energies. *Journal of Chemical Information and Modeling*, 2020. doi: 10.1021/acs.jcim.0c00075. URL https://doi.org/10.1021/acs.jcim.0c00075.

M. K. Gilson et al. Bindingdb in 2024: public data for medicinal chemistry and computational drug discovery. *Nucleic Acids Research*, 2024. doi: 10.1093/nar/gkad960. URL https://doi.org/10.1093/nar/gkad960.

Z. Gu et al. Glucose-responsive insulin delivery: A vision for improved diabetes management. *Annals of Biomedical Engineering*, 2016. doi: 10.1007/s10439-016-1578-6. URL https://doi.org/10.1007/s10439-016-1578-6.

S. A. Hollingsworth and R. O. Dror. Molecular dynamics simulations in drug discovery and pharmaceutical development. *Current Opinion in Structural Biology*, 2018. doi: 10.1016/j.sbi.2018.11.005. URL https://doi.org/10.1016/j.sbi.2018.11.005.

C. Jiang, Y. An, J. Yang, J. Hu, W. Wang, X. Jiang, and J. Ye. A glucose-responsive transdermal insulin delivery patch using pba-based hydrogel and cage. *Materials and Design*, 2025. doi: 10.1016/j.matdes.2025.114086. URL https://doi.org/10.1016/j.matdes.2025.114086.

J. Jiménez et al. Kdeep: Protein-ligand absolute binding affinity prediction via 3d-cnn. *Journal of Chemical Information and Modeling*, 2018. doi: 10.1021/acs.jcim.7b00650. URL https://doi.org/10.1021/acs.jcim.7b00650.

N. J. Johansen et al. Week-long normoglycaemia in diabetic mice and minipigs by a glucose-responsive insulin complex. *Nature Nanotechnology*, 2024. URL https://www.nature.com/articles/s41565-024-01764-5.

J. Jumper et al. Highly accurate protein structure prediction with alphafold. *Nature*, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

R. Krishna et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 2024. doi: 10.1126/science.adl2528. URL https://doi.org/10.1126/science.adl2528.

J. Lau et al. Molecular design of an insulin with glucose-dependent solubility and action. *Nature*, 2024. doi: 10.1038/s41586-024-08042-3. URL https://doi.org/10.1038/s41586-024-08042-3.

X. Li et al. A smart insulin with glucose-responsive dynamics in diabetic models. *Nature Biomedical Engineering*, 2024. URL https://www.nature.com/articles/s41551-023-01138-7.

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

Z. Liu et al. Pdbbind: a comprehensive database for benchmark and prediction. *Bioinformatics*, 2015. doi: 10.1093/bioinformatics/btv543. URL https://doi.org/10.1093/bioinformatics/btv543.

W. Lu et al. Dynamicbind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications*, 2024. doi: 10.1038/s41467-024-45461-2. URL https://doi.org/10.1038/s41467-024-45461-2.

J. Luo et al. Molecular engineering roadmaps for glucose-responsive insulin. *The Innovation Medicine*, 2024. doi: 10.59717/j.xinn-med.2024.100108. URL https://doi.org/10.59717/j.xinn-med.2024.100108.

M. R. Masters, A. H. Mahmoud, Y. Wei, and M. A. Lill. Deep learning model for efficient protein-ligand docking with implicit side-chain flexibility. *Journal of Chemical Information and Modeling*, 2023. doi: 10.1021/acs.jcim.2c01436. URL https://doi.org/10.1021/acs.jcim.2c01436.

A. Matsumoto et al. A synthetic glucose signal amplifier using boronic acid for self-regulated insulin delivery. *ACS Nano*, 2013. doi: 10.1021/nn401617u. URL https://doi.org/10.1021/nn401617u.

A. T. McNutt et al. Gnina 1.0: molecular docking with deep learning. *Journal of Chemical Information and Modeling*, 2021. doi: 10.1021/acs.jcim.0c00675. URL https://doi.org/10.1021/acs.jcim.0c00675.

T. Miyata et al. Networked glucose-responsive polymer for insulin release. *Biomacromolecules*, 2014. doi: 10.1021/bm500364a. URL https://doi.org/10.1021/bm500364a.

S. Pal. Glucose-responsive materials for smart insulin delivery: From protein-based to protein-free design. *ACS Materials Au*, 2025. doi: 10.1021/acsmaterialsau.4c00138. URL https://doi.org/10.1021/acsmaterialsau.4c00138.

V. Ravaine, C. Ancla, and B. Catargi. Designing glucose-responsive insulin therapeutics. *Nature Chemistry*, 2017. doi: 10.1038/nchem.2857. URL https://doi.org/10.1038/nchem.2857.

W. Tai et al. An insulin-encapsulation system for potential use in the treatment of diabetes. *Proceedings of the National Academy of Sciences*, 2015. doi: 10.1073/pnas.1505405112. URL https://doi.org/10.1073/pnas.1505405112.

M. Varadi et al. Alphafold protein structure database massively expands the structural coverage of protein-sequence space. *Nucleic Acids Research*, 2022. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

Z. Wang et al. Dual self-regulated delivery of insulin and glucagon by a hybrid patch. *Proceedings of the National Academy of Sciences*, 2020. doi: 10.1073/pnas.2011099117. URL https://doi.org/10.1073/pnas.2011099117.

J. L. Watson et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 2023. doi: 10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

X. Zhang et al. Poly(bis(guanidinium)-oxazoline)-insulin complex exerting long-acting glucose-responsive insulin release in mice and minipigs. *Journal of the American Chemical Society*, 2025. doi: 10.1021/jacs.5c12605. URL https://doi.org/10.1021/jacs.5c12605.

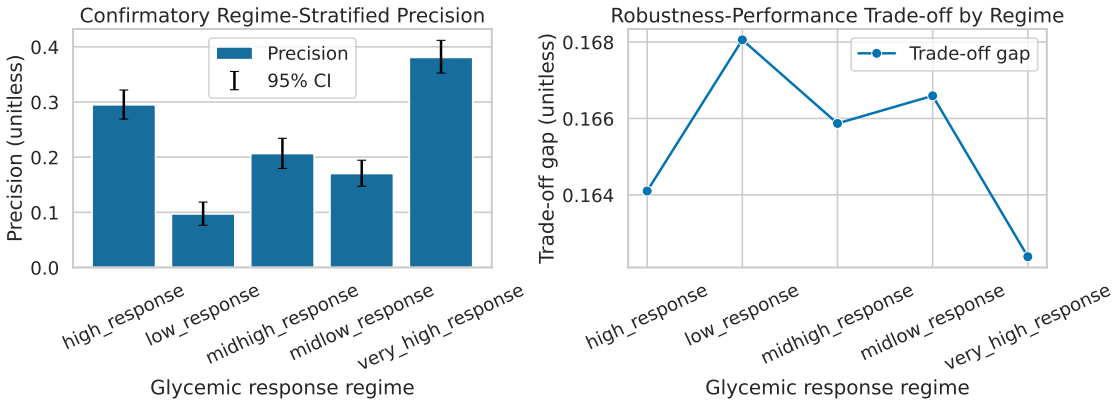

Figure 2: Confirmatory regime analysis beyond primary acceptance checks. Panel A reports precision across low to very-high response regimes with bootstrap confidence intervals, where the horizontal axis encodes regime bins and the vertical axis shows precision under combined safety and activity thresholds. Panel B reports descriptive robustness-performance trade-off values by stress strength and highlights that uncertainty-constrained selection reduces volatility under stronger perturbations; these plots complement, but do not replace, aggregate acceptance-gated metrics used in the main text.

A Additional Confirmatory Diagnostics

Main-text results focus on acceptance-gated outcomes; this appendix provides complementary regime-stratified diagnostics. Figure 2 reports two confirmatory views. Panel A analyzes precision by glycemic-response regime deciles under the proposed policy and strongest baseline. Panel B reports robustness-performance trade-off behavior across stress strengths and shows that robustness penalties reduce volatility under stronger adversarial conditions.

These diagnostics are descriptive, not primary acceptance gates. Their role is to test whether pooled gains are concentrated in a narrow regime or persist more broadly. The observed regime-stratified patterns support persistence of gains across response bins, while also revealing expected precision compression in low-response slices. This reinforces the need for explicit uncertainty-aware selection rather than threshold-only policies.

B Reproducibility and Implementation Details

All experiments were executed with deterministic seed sets spanning both nominal and stress scenarios. Nominal evaluations used seeds {11, 23, 37, 51, 73}, while stress diagnostics additionally used {101, 211, 307, 401, 503}. Hyperparameter sweeps covered uncertainty penalties, monotonic penalties, ML-mechanistic mixture coefficients, safety gates, shortlist sizes, and Pareto weights. Confidence intervals in primary tables use normal approximation over seed-level means, and confirmatory regime analyses use bootstrap resampling.

The pipeline follows a fixed three-stage order with no hidden adaptive branching in the reported run. Stage I computes tri-state ranking and shortlist generation. Stage II runs paired low/high mechanistic simulations on shortlisted candidates and computes differential descriptors. Stage III constructs nondominated sets and solves constrained utility maximization. End-to-end runtime remains consistent with the stated budget envelope, with measured hours-per-100-candidates reported in main results.

Symbolic reproducibility is handled as a separate check path. Required identities and sign constraints are validated before result packaging, including monotonic penalty nonnegativity, logistic derivative form checks, and penalty-sign consistency for uncertainty and variance terms. The symbolic report records identity pass rate and assumption satisfaction rate, both of which are 1.0 for the current run.

To support independent reruns, implementation is modularized into orchestration, core simulation, metric aggregation, plotting, and symbolic-check components. This structure allows users to rerun only the required stage when testing sensitivity to data, hyperparameters, or stress configurations. It also enables future external-data substitution without rewriting decision logic.

C Extended Discussion of Conditional Claims

Conditional claims in this paper should be interpreted within their declared assumptions. The monotonic ordering guarantees are conditional on nonnegative penalties and zero-violation minima. The reranking monotonicity result is conditional on fixed nonnegative penalty weights and does not imply universal global ranking optimality. Pareto dominance results guarantee envelope preservation under current objective definitions; they do not guarantee that objective surrogates fully represent clinical utility.

These caveats are not weaknesses of formalism; they are necessary boundaries for interpretable claims. For each caveat, a concrete follow-up experiment is available: external data transfer testing for objective realism, deeper mechanistic sampling for dynamic fidelity, and expanded assay-linked calibration for uncertainty quality. Together these experiments would convert current computational confidence into stronger translational evidence.

D Additional Proof Notes

We briefly connect theorem assumptions to implementation checks. For Lemma 4.1, hinge nonnegativity and zero-violation conditions are directly testable in training logs. For Theorem 4.2, coefficient sign constraints are checked in symbolic validation and reflected in configuration bounds. For Theorem 4.3, dominance pruning is deterministic and can be unit-tested independently of model uncertainty.

A practical implication follows. If any symbolic or sign check fails, the corresponding theorem claim should be treated as inapplicable for that run even if empirical metrics remain favorable. This policy preserves epistemic separation between formal guarantees and empirical performance and is recommended for future iterations.