Agentic Discovery of Cryomicroneedle Formulations

Hao Li, Lifu Du, Nurul Hameed, Shemonti Saha Authai,
Zlata Stefanovic, Chenjie Xu

Department of Biomedical Engineering, City University of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong SAR, China

Abstract

Cryomicroneedles offer a route to minimally invasive intradermal delivery of living cells, but their cryogenic formulations must reconcile cell protection with constraints on toxicity and device fabrication. Here we report an AI-assisted, closed-loop workflow for cryomicroneedle cryoprotectant discovery that combines literature curation, Gaussian-process surrogate modelling, Bayesian optimization, and sequential wet-lab validation. A curated dataset of 198 mesenchymal stem-cell cryopreservation formulations from 42 studies was converted into 21 ingredient features and used to train an uncertainty-aware literature prior. This model captured moderate structure in the literature data but failed prospectively, motivating iterative wet-lab correction. Across ten validation iterations and 106 wet-lab observations, the model progressively adapted to cryomicroneedle-specific outcomes: batch RMSE decreased from 41.21 to 6.86 percentage points, later-stage rank correlations became consistently positive, and the cumulative wet-lab predicted-versus-measured summary reached $R^2 = 0.942$. The best validated formulation achieved 95.15% post-thaw viability with low DMSO, ectoin, ethylene glycol, and fetal bovine serum. However, high viability alone did not ensure intact cryomicroneedle formation, highlighting the need for future multi-objective optimization. These results demonstrate that agent-assisted computational infrastructure can make data-efficient formulation discovery more accessible to labs with minimal data expertise in-house. Project code is available at `https://github.com/baitmeister/ML-for-CryoMN`.

Keywords: cryomicroneedle, cryopreservation medium, machine learning, artificial intelligence, mesenchymal stem cells

1 Introduction

Cryomicroneedle is an emerging platform for the intradermal delivery of living cells. In this approach, a cryogenic formulation containing living cells is molded into microneedle templates and frozen, embedding viable cells directly within the needle tips. These frozen microneedle patches can either be prepared in advance and shipped at low temperature to the end-user to be later applied with minimal preparation or be directly fabricated by the end-user on the spot. Thus, cryomicroneedle combines the minimally invasive advantages of microneedle delivery with the logistical benefits of living cell cryopreservation.[1, 2, 3]

A defining challenge in cryomicroneedle technology is the design of the cryogenic formulation. Unlike conventional cryopreservation media, which are optimized primarily to preserve

cells during freezing and thawing, cryomicroneedle formulations must provide not only effective cryoprotection to maintain cell viability but also a mechanically stable structure capable of penetrating skin and dissolving rapidly after insertion. The combined biological and mechanical constraints increase the complexity of formulation design and create a large and heterogeneous formulation design space. Numerous candidate ingredients have been reported in the cryobiology literature, including permeating cryoprotectants such as dimethyl sulfoxide (DMSO) and ethylene glycol; membrane-stabilizing sugars like sucrose; osmoprotective solutes such as ectoine; and complex additives like serum. However, the performance of a formulation depends on nonlinear interactions among multiple components. Using exhaustive experiments to explore this combinatorial space is impractical, and intuition-driven approaches tend to prioritize familiar formulations that may overlook uncertain yet promising regions of the design space. As a result, identifying effective formulations for cryomicroneedle becomes difficult through conventional experimental design.[4, 5]

Machine learning offers promising strategies for navigating such complex formulation spaces by guiding the experimental exploration more efficiently. Instead of relying on exhaustive screening or empirically guided design, data-driven predictive models can learn relationships between formulation composition and experimental outcomes, allowing researchers to prioritize experiments that most likely improve performance or reveal new information.[6, 7, 8] However, implementing such data-driven approaches in practice presents a substantial barrier. Constructing a usable computational pipeline typically requires integrating literature data extraction, dataset curation, model training, uncertainty-aware experiment selection, iterative experimental updating, and analysis tools. Many experimental laboratories recognize the potential value of machine learning for experimental optimization but lack the computational expertise or engineering time needed to build and maintain this infrastructure. Thankfully, recent democratization of artificial intelligence (AI) suggests a possible solution to this barrier. AI agents are increasingly capable of interacting with software tools, generating code, and coordinating multi-step computational workflows under human supervision. Without replacing human researchers in scientific reasoning and higher-level decision-making, such systems function as programmable collaborators that provide additional guidance and help researchers construct the computational infrastructure required for modern data-driven discovery.[9, 10, 11]

Here we present a proof-of-concept study demonstrating how an AI-assisted computational workflow can support the discovery of cryomicroneedle formulations (Fig. 1). We develop an iterative machine-learning framework that integrates literature-derived data, probabilistic surrogate modeling, Bayesian optimization, and sequential wet-lab validation. Much of the supporting computational infrastructure is implemented through AI-assisted coding under human oversight, thereby significantly lowering the practical barrier to deploying such methods in experimental laboratories. Using this framework, we demonstrate the discovery of cryomicroneedle cryoprotectant formulations that reduce reliance on DMSO while maintaining high post-thaw cell viability. Other important attributes, including microneedle mechanical performance, remain crucial but are reserved for future multi-objective studies. Therefore, the aim of the present work is not to claim a final optimal cryomicroneedle formulation, but to illustrate how AI-assisted workflow construction can enable data-efficient experimental optimization in research environments where specialized computational expertise is limited.

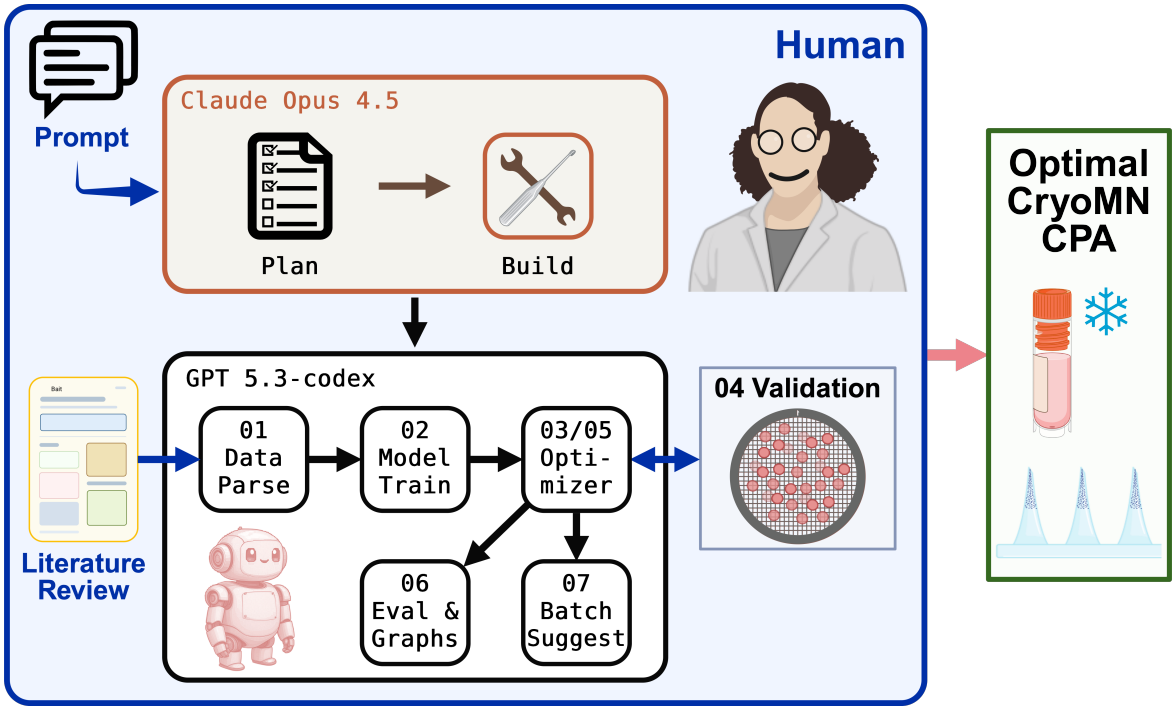

Fig. 1 — Schematic of the AI-assisted cryoprotectant formulation discovery workflow. Literature-derived formulation data are parsed into structured features, used to train Gaussian-process models, and iteratively updated with wet-lab validation results before new formulation batches are suggested.

2 Methods

2.1 AI-assisted workflow construction

The computational pipeline was developed using AI-assisted coding under direct human supervision. In the initial one-shot prompt (Supplementary Methods), the scientific task was defined as a low-data formulation-optimization problem with four constraints: to maximize post-thaw viability, to minimize DMSO usage, to restrict candidate formulations to no more than 10 active ingredients, and to enable iterative updating after each wet-lab round. The agent, powered by Claude Opus 4.5 in Antigravity, proposed the plan to implement the overall workflow structure, including data parsing, surrogate modeling, optimization methods, candidate generation, model updating, and evaluation. After human revision of said plan, the agent proceeded to generate code, repository structure, and documentation. The generated materials from this one-shot prompt were reviewed and executed by the authors using Codex on GPT-5.3-codex. Additional features in the final program that did not appear in the initial repository, including evaluation and explainability, stage-indexed coefficient-of-determination R^2 plots, and next-batch formulation design, were prompted by the authors, coded, and implemented using Claude Opus 4.5 in Antigravity and GPT-5.3-codex in Codex.

2.2 Literature dataset preparation

The initial dataset was assembled from 75 primary research articles reporting a total of 499 cryoprotective formulations and viability outcomes (Supplementary Methods; Supplementary Data). We focused on the cryopreservation of mesenchymal stem cells (MSCs) because it is one of the model cell types in cryopreservation research. The literature review was done manually using keywords related to MSCs, including naming variants such as mesenchymal stromal cells and source-specific variants such as umbilical cord-derived MSCs or adipose-derived stem/stromal cells (ASCs), together with viability. Reported cryoprotective formulations and immediate post-thaw viability values were manually extracted as free text. Types of viability assays and cooling protocols were not modeled separately in the present proof-of-concept study. Such free-text formulation and viability descriptions were then parsed into a structured matrix in which each canonical ingredient occupied one feature column and each row represented one unique formulation. Synonymous ingredient names were merged into shared canonical identities, whereas basal media and buffer systems were excluded as formulation variables. Concentrations were harmonized into a mixed unit representation: ingredients with defined molecular properties were converted to molar units, whereas complex mixtures and polymers that could not be meaningfully represented in molar form were retained as percentages. Molecular-weight-specific polyethylene glycol species were tracked separately because their cryoprotective behavior may differ across size classes. Any duplicate formulation with discordant reported viabilities, largely due to different cooling methods, was collapsed by averaging the viability values because cooling methods are out of scope. Any relative viability (i.e., Formulation B has an $x\%$ viability when normalized to Formulation A's viability) that might result in viability values greater than 100% was excluded from model training. To reduce feature space and instability from very sparse variables, ingredients that appeared in fewer than 3 formulations were deleted. The whole data-cleaning process yielded the final 198 unique literature formulations described by 21 canonical ingredient features.

2.3 Gaussian-process surrogate model

Formulation performance was modeled using Gaussian-process regression. This approach was chosen because it is suitable for sparse experimental datasets and provides both a predicted mean and a predictive uncertainty for each candidate formulation. The initial literature-only surrogate used a Matérn kernel and standardized ingredient features before fitting. Initial model performance was estimated using five-fold cross-validation, in which the literature dataset was split into five subsets and each subset was predicted once by a model trained on the remaining four subsets.

2.4 Candidate generation

The trained Gaussian-process surrogate was used to propose candidate cryoprotective formulations for wet-lab testing. In the first five iterations, candidates were generated by random sampling of the feasible formulation space and then ranked by the Gaussian-process predicted mean viability. This strategy provided broad initial exploration when little cryomicroneedle-specific wet-lab feedback was available, allowing the model to learn the assay-specific landscape

before being trusted to optimize it. In later iterations, candidate generation was replaced by Bayesian optimization. In this setting, the same Gaussian-process model served as the surrogate model, and an upper-confidence-bound (UCB) acquisition function detailed in Equation 1 was used to combine predicted viability and uncertainty into a single search objective. In Equation 1, x denotes a candidate formulation vector in feature space, $\mu(x)$ is the model-predicted mean viability for that candidate, $\sigma(x)$ is the model-predicted standard deviation for that candidate, and κ is the exploration weight controlling how strongly uncertainty is rewarded relative to predicted mean performance, which we set to $\kappa = 0.5$. The UCB score was computed in post-thaw viability percentage-point units. Model predictions were obtained as the sum of the literature-derived prior mean and the wet-lab residual-correction model, followed by the stage-specific additive mean correction and multiplicative uncertainty correction stored in the model metadata. Thus, both $\mu(x)$ and $\sigma(x)$ in Equation 1 are reported on the calibrated viability-percent scale.

$$a_{\mathrm{UCB}}(x) = \mu(x) + \kappa \cdot \sigma(x) \tag{1}$$

Then, differential evolution was used as the numerical optimizer to search for formulations with high acquisition scores under formulation constraints. Candidate formulations were required to satisfy practical constraints, including a maximum of 10 active ingredients and an upper DMSO bound. We set up two different DMSO upper limits, thus generating two different pools of candidate files: "general" at 5% DMSO and "DMSO-free" at 0.5% DMSO. More detailed reasoning on optimization methods, search bounds, acquisition parameters, and batch-selection rules are provided in Supplementary Methods.

After the formulation candidates were generated, the final wet-lab slate was selected by a separate formulation recommendation step. In the first five iterations, where random search was used, the selection was done manually by picking candidates with high predicted viability yet not overly similar to other formulations. In later iterations, this step involved combining high-performing exploitation candidates with exploration candidates designed to expose model blind spots. Exploitation candidates were selected from the highest-ranked Bayesian-optimization outputs after removing already tested formulations, whereas exploration candidates were generated from residual patterns observed in previous wet-lab stages. Formulations that had performed better than predicted were used as local anchors, and historically underpredicted regions of formulation space were used to define blind-spot probes. Additional coverage probes were included to expand support in sparse regions of the observed formulation space. In this way, the formulation recommendation step translated model predictions and model failures into a balanced set of wet-lab validation candidates, intended to both identify strong formulations and improve future model performance.

2.5 Wet-lab validation

Selected formulations per recommendation were prepared. Umbilical cord-derived human mesenchymal stem cells at either the 7^{th} or 8^{th} passage were used to conduct the viability validation. These cells were cultured in low-glucose Dulbecco's modified Eagle medium (DMEM) with GlutaMAX and pyruvate (Gibco, Cat. 10567014) until 80% confluence before being trypsinized for

viability assay. Polydimethylsiloxane (PDMS) negative molds used to fabricate cryomicroneedles were affixed inside the wells in 6-well plates. Detailed specifications of the PDMS mold and the resulting cryomicroneedles can be found in Supplementary Methods. A 700 μL aliquot of the cell suspension solution prepared using the selected formulation at around 1×10^6 cells/mL was added to PDMS negative molds, and the plates containing the molds were then centrifuged once for 3 minutes at 400 g. Afterwards, no air bubbles or hollow cavities near the needle channels were visible. Excess liquid outside of the needle tips was wicked away with tissue paper. Needle tips were clearly filled with cell-containing candidate cryoprotectants. Another 700 μL of the respective cell-containing cryoprotectant solution was added to each mold. The 6-well plates containing molds were transferred to a 4°C refrigerator for 30 minutes, then to a -20°C freezer for 2.5 hours, before being transferred to a -80°C freezer for another 2.5 hours. Afterwards, these cryomicroneedle-containing plates were directly transferred to a 37°C floating water bath within 15 seconds, where the cryomicroneedles were thawed inside the wells of the 6-well plates. Thawed solutions were transferred to centrifuge tubes for centrifugation at 400 g for 4 minutes. The supernatant was then removed, and the cells were resuspended with 2 mL staining cocktail in a 12-well plate. This cocktail was prepared to have 1 μM CellTracker Green CMFDA and 1 drop of NucBlue Live ReadyProbes reagent containing Hoechst 33342 per mL in phosphate-buffered saline (PBS). Immediately after, the 12-well plates were transferred to the incubator (37°C, 5% CO_2, 95% relative humidity) for 45 minutes for staining. Afterwards, 18 μL of thawed cryomicroneedle solution was sampled from each well into a Countess Cell Counting Chamber Slide for viability assay. Post-thaw viability was defined as the number of CMFDA-positive (alive) cells divided by the number of Hoechst-positive (all) cells in the Countess image field. For each formulation, five image replicates were analyzed using ImageJ, and the average was used as the viability for that formulation.

2.6 Feedback and model update

Measured post-thaw viability values were incorporated into the next model update. The literature-only model generated the first validation batch, which is referred to as stage 0 or iteration 0. After this stage, the model was updated using a prior-mean correction strategy: the literature-trained Gaussian process was retained as the baseline predictor, and a second Gaussian process was fitted to the residuals between measured wet-lab viability and literature-model prediction. The final prediction was the sum of the literature-derived prediction and the wet-lab residual correction. After the model update, iteration 1 began. Later iterations followed the same closed-loop structure: update the surrogate, generate candidates, test selected formulations, and update the model again using wet-lab results.

This update strategy allowed the model to preserve broad formulation trends learned from the literature while adapting to cryomicroneedle-specific wet-lab outcomes. Wet-lab observations were given a default 50-fold greater local influence than literature data through source-specific noise assumptions. Wet-lab generalization was estimated by cross-validating with only the wet-lab rows while retaining all literature rows in the training set for every fold, and the residual diagnostics from this process were then converted into two post-hoc calibration terms applied downstream to all reported predictions: an additive mean correction and a multiplicative uncertainty correction. Details of the implementation, other parameters, and the cross-validation protocol can be found in Supplementary Methods.

2.7 Evaluation and explainability

After the wet-lab viability feedback, model evaluation was performed at the stage level using frozen checkpoints. Each saved model was evaluated against the wet-lab batch that had been selected from that stage, preserving the historical decision context. For example, the literature-only model was evaluated against the first wet-lab batch chosen from the literature-only candidate files (stage 0), whereas the frozen iteration 3 model was evaluated against the wet-lab formulations that were selected from the iteration 3 candidate outputs, not against formulations proposed later by iteration 4 or iteration 5 models. This prevents later model revisions from rewriting the evidence for earlier decisions and ensures that each stage is judged on the prospective experiment set it truly generated.

For each stage batch, we reported various metrics, including root mean squared error (RMSE), mean absolute error (MAE), ranking accuracy, and Spearman's and Kendall's rank correlation coefficients for that specific batch. Moreover, we reported the coefficient of determination using a prospective-cumulative R^2 protocol. At each stage k, predictions from the frozen stage model were evaluated against the cumulative wet-lab set available up to stage k. This design preserves temporal ordering while increasing effective sample size and outcome-range coverage, which makes the explained-variance estimate more stable and interpretable across iterations. We did not treat batch-specific R^2 as a primary endpoint because individual wet-lab batches were small, often with only a few formulations validated, and R^2 is highly sensitive to both outliers and within-batch response variance under those conditions. As a result, batch R^2 could fluctuate sharply, including sign reversals, without reflecting a meaningful change in practical predictive utility. In this setting, reporting only batch R^2 would therefore be noisy and potentially redundant relative to the broader prospective-cumulative trend.

To make the model's decisions and reasoning more interpretable, an explainability module was used to assess the active iteration-specific observed context. Graphs that showed feature importance, Shapley additive explanations (SHAP) summaries, support-aware marginal effects, pairwise interaction contours, acquisition landscapes, and uncertainty diagnostics were generated to identify dominant formulation variables, visualize where the model was interpolating versus extrapolating, and assess whether predictive uncertainty remained useful for experimental decision-making. Detailed definitions of the evaluation metrics, prospective-cumulative R^2 implementation, and explainability plots are provided in Supplementary Methods.

3 Results

The complete agent-assisted workflow connecting literature curation, model training, candidate generation, wet-lab validation, and batch recommendation is summarized in Fig. 1.

3.1 Initial baseline architecture

The one-shot prompt generated a baseline that was immediately organized as a modular closed-loop pipeline rather than as an isolated analysis notebook. The prompt defined the problem as a low-data cryoprotectant optimization, constrained candidate recipes to no more than 10

active ingredients, asked for DMSO minimization and viability maximization, and required the model to evolve after wet-lab validation. In response, the first substantive repository commit on January 24, 2026 added a four-stage architecture: free-text formulation parsing, Gaussian-process surrogate training, constrained candidate generation, and a validation/update loop. Each stage was placed in a separate source directory with a companion README, and the same commit added requirements, processed data, a trained baseline model, candidate tables, and a wet-lab validation template.

The data layer converted unstructured formulation descriptions into a machine-learning matrix by canonicalizing ingredient synonyms, excluding basal media and buffers, harmonizing concentrations where possible, retaining DMSO as both a feature and a design constraint, and averaging duplicate formulation entries with discordant viabilities. The modeling layer then fit a standardized Gaussian-process regressor with a Matérn kernel, which was selected because the literature dataset was small and because predictive uncertainty was needed for sequential experiment selection. The first baseline model tracked 21 ingredient features and already produced cross-validation diagnostics, feature-importance output, and serialized model artifacts.

The first optimization layer was already split into two candidate streams: a general low-DMSO pool with an upper DMSO bound and a near-DMSO-free pool. The first generated summaries ranked predicted formulations with uncertainty estimates and ingredient counts, while the validation module created the file interface needed to add measured post-thaw viability and retrain the model. The early commit history on GitHub showed the limitations of this baseline as well as its usefulness; yet within the next three days, follow-up commits corrected candidate formatting, added a dedicated differential-evolution Bayesian-optimization module, introduced explainability plots, and implemented a prior-mean wet-lab correction model. Therefore, the one-shot output did not produce the final scientific workflow, but it did generate the working scaffold from which the later iteration-aware, calibrated, and explainable system was built.

Table 1 — Top 15 ingredients sorted by appearance in unique formulations reported in the literature.

Ingredient	**Appearances**	**Highest viability reported (%)**
Dimethyl sulfoxide (DMSO)	89	100
Fetal bovine serum (FBS)	58	92
Sucrose	31	93.4
Trehalose	28	99
Ethylene glycol	20	96
Human serum albumin (HSA)	20	95.7
Glycerol	18	83
Hydroxyethyl starch (HES)	17	100
Human serum	13	95.7
Polyethylene glycol (PEG; ∼3350 Da)	13	94
Polyvinylpyrrolidone (PVP)	13	72.5
Methylcellulose	11	91
Propylene glycol	8	83.4
Ectoin	6	96
Betaine	5	84

3.2 Dataset cleanup and literature formulation landscape

Our dataset comprises 198 unique formulations from 42 independent studies retained after cleaning the original 393 formulations (Supplementary Data). The overall post-thaw viability across all formulations demonstrated significant variance, averaging 63.88% (median: 70.00%, range: 0.00–100.00%). Analysis of the distribution of standard cryoprotective agents reveals DMSO as the most prominent component, present in 89 formulations (44.9%). Other frequently utilized additives included fetal bovine serum (FBS) (29.3%), sucrose (15.7%), trehalose (14.1%), ethylene glycol (10.1%), and human serum albumin (HSA) (10.1%). The dataset contains a slight majority of DMSO-free formulations (n = 109, 55.1%). While popular, there remains an efficacy gap between the two categories: DMSO-containing formulations (n = 89) demonstrated a higher average post-thaw viability of 71.90%, whereas DMSO-free formulations (n = 109) yielded a notably lower average post-thaw viability of 57.33%. A summary of the literature data is presented in Table 1.

A critical subset of 59 formulations (29.8%) achieved excellent post-thaw cell viabilities of $\geq$80%. Within this high-performing threshold, the reliance on DMSO was pronounced, occurring in 42 (71.2%) of these top formulations. Other recurring components in this high-efficacy tier included FBS, trehalose, hydroxyethyl starch (HES), and ethylene glycol. These results established both the usefulness and the bias of the literature prior: the model could learn meaningful formulation structure, but a direct literature-trained optimum would likely over-favor DMSO-containing recipes.

3.3 Initial model and first wet-lab validation

To establish a pre-experimental baseline, we first trained a Gaussian process model using the literature-curated formulation dataset. On five-fold cross-validation, this model achieved an RMSE of 20.25 $\pm$ 1.99 with a mean R^2 of 0.247 $\pm$ 0.096, and when fit to the full literature set it reached a training RMSE of 14.55 and $R^2 = 0.618$ (Fig. 2a). These results indicated that the literature data contained recoverable predictive structure, but only with moderate fidelity. The literature-only model's SHAP importance showed that the initial model was strongly shaped by ingredients that were common or historically successful in the literature record (Fig. 2c).

That performance did not translate prospectively. In the first wet-lab validation on the literature-only model (iteration 0) using the iterative process described in Fig. 3, the model's error increased sharply on 18 candidate formulations, with RMSE rising to 41.21 while R^2 fell to -2.52, and rank-based agreement was poor (Fig. 2b, Fig. 4). We attribute this drop primarily to the heterogeneity and reporting bias of literature-derived data. The aggregated studies span different mesenchymal stem cell sources, cryopreservation workflows, assay endpoints, post-thaw handling conditions, and reporting standards, meaning that nominally similar formulations may reflect substantially different experimental contexts. The model update after this first batch changed the attribution pattern, as shown by the post-update SHAP importance (Fig. 2d), supporting the need for iterative wet-lab feedback rather than one-shot literature extrapolation.

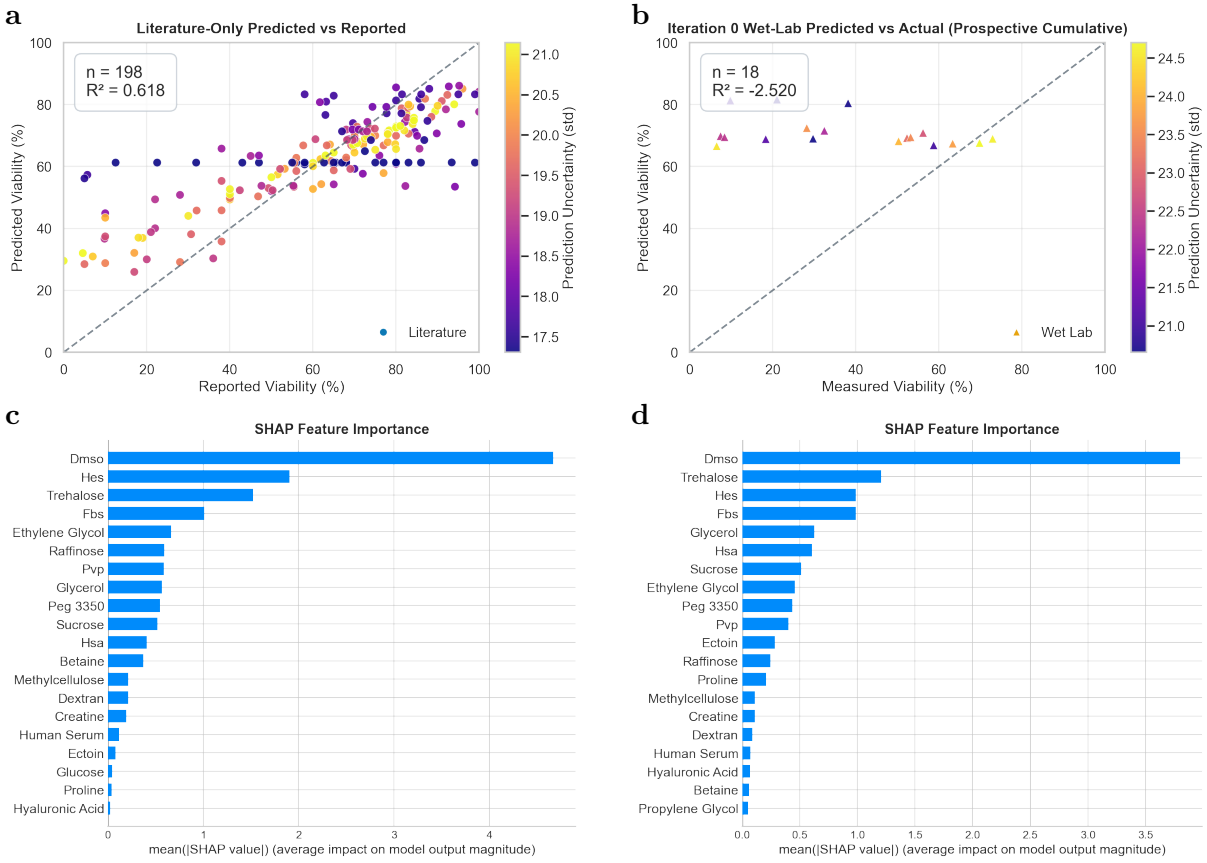

Fig. 2 — Initial model and first wet-lab validation: a, literature-only predicted versus reported viability; b, wet-lab predicted versus measured viability for the first validation batch; c, SHAP importance for the literature-only model; d, SHAP importance after model update using the first wet-lab validation batch results. In a and b, each dot represents a formulation's viability.

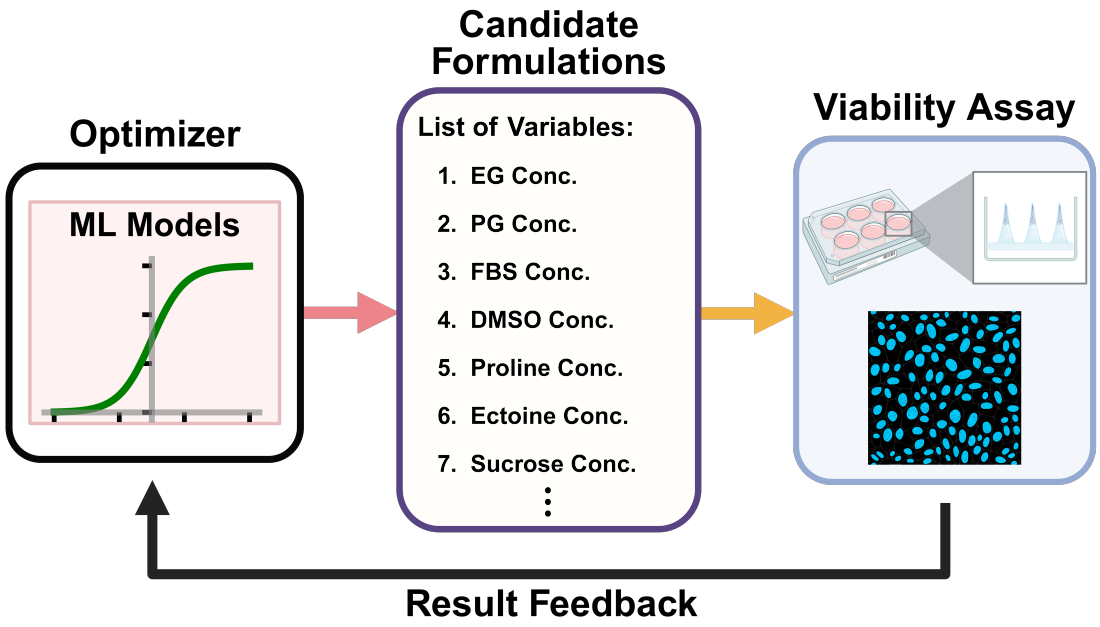

Fig. 3 — Schematic of the iterative optimization process. An optimizer based on the trained model generates candidate formulations for viability testing. The measured results are then used to update the trained models.

3.4 Model after ten iterations of wet-lab validation

Across the closed-loop validation campaign, the model was updated after each wet-lab batch and progressively shifted from a literature-only predictor toward a wet-lab-corrected surrogate. In total, the validation table contains 106 wet-lab observations after the literature-derived training set. The practical outcome improved relative to the literature-only batch: the best formulation in the literature-only batch reached 72.93% viability, iteration 1 increased the best-so-far value to 79.09%, and iteration 5 identified the highest observed formulation, EXP5110 with 21.0 mM DMSO, 291.1 mM ectoin, 1.79 M ethylene glycol, and 5.4% FBS, achieving 95.15% viability (Supplementary Data). We would like to note that iteration 5's elevated viability should be interpreted cautiously because it may reflect a combination of model-guided enrichment and batch-level experimental variation. The high-performing formulations clustered around a related low-DMSO ectoin/ethylene glycol/FBS region, but they were tested within a narrow wet-lab window in which subtle differences in cell condition, handling, and measurement could systematically shift viability. Thus, this result is useful evidence for a promising local formulation region, but should not be taken as a standalone demonstration of global model improvement. Over the course of the validation loop, model error also improved. Batch RMSE fell from 41.21 in the literature-only prospective test to 21.67 after iteration 1, 14.74 after iteration 2, 10.10 after iteration 8, and 6.86 after iteration 9, while later-stage rank agreement became consistently positive (Fig. 4).

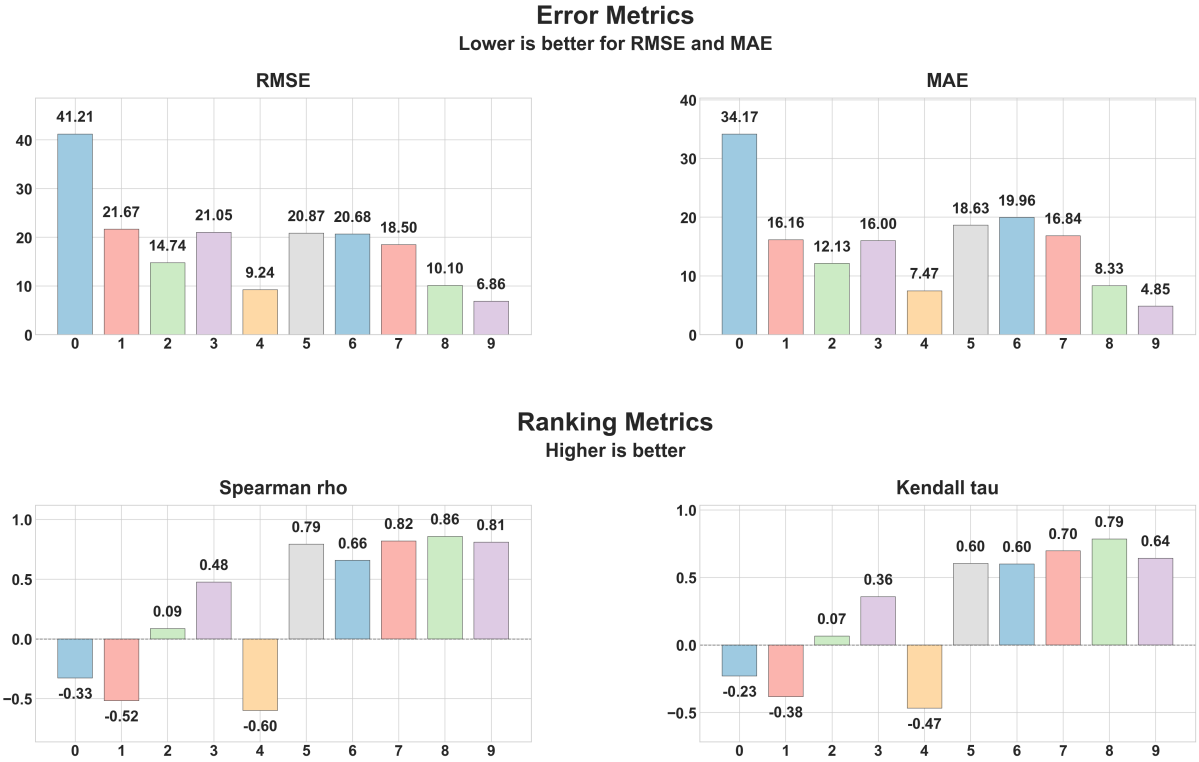

Fig. 4 — Batch-level model evaluation across validation stages. Frozen model checkpoints were evaluated against the wet-lab batches generated at the corresponding stage. RMSE and MAE are reported in viability percentage points; lower values indicate smaller prediction error. Spearman's ρ and Kendall's τ measure rank agreement within each batch; higher values indicate better preservation of formulation ordering. Because individual wet-lab batches were small, these metrics should be interpreted as stage-level diagnostics rather than progressive performance curves.

A comprehensive look across all batches showed the same overall trend while retaining the expected noise of small experimental batches (Fig. 4). The earliest stages (iteration 0-4) using random search were unstable: models had erratic ranking behavior (Spearman's ρ and Kendall's τ); even at iteration 4, the model still had a negative rank correlation despite iteration 4's low RMSE due to the narrow viability range of formulations in that batch. In contrast, later Bayesian optimization stages (iteration 5-9) showed exploratory behavior that improved ranking significantly: Spearman's ρ reached 0.79, 0.66, 0.82, 0.86, and 0.81 across iterations 5-9. Although the exploration formulations drove up error in the early Bayesian optimization stages, error eventually dropped with more iterations. These metrics show that the loop became practically better at understanding the viability range and at nominating testable formulations with good viability. In the end, iteration 9 had a batch RMSE of 6.86, and the cumulative wet-lab predicted-versus-measured summary containing 106 wet-lab observations reached an R^2 of 0.942 (Fig. 5a). After iteration 9's model was updated with wet-lab results, the iteration 10 SHAP feature importance was plotted to further demonstrate the shift from the initial literature model, with ethylene glycol and ectoin emerging as major contributors alongside the original DMSO-rich literature signal rather than remaining subordinate to it (Fig. 5b).

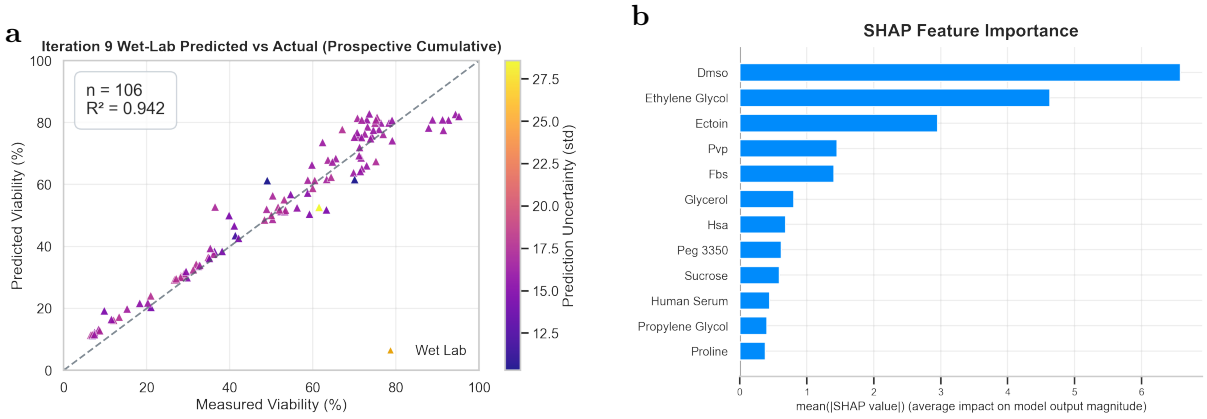

Fig. 5 — Latest wet-lab-corrected model summary after iterative validation: a, cumulative predicted-versus-measured wet-lab viability summary using the iteration-9 checkpoint across 106 wet-lab observations; b, iteration-10 SHAP importance after updating the iteration-9 model with wet-lab validation results. DMSO remained an influential feature, while ethylene glycol and ectoin emerged as major wet-lab-adapted contributors.

Apart from regular performance metrics, in each batch, we generated the acquisition landscape and interaction contours that are central to quickly understanding the potentially unexplored areas and ingredient interactions. These plots translate a high-dimensional Gaussian-process model into decision maps that an experimentalist can act on. The acquisition landscape shows where the model predicts high viability, where it is uncertain, and where the combination of the two gives high upper-confidence-bound utility. The script automatically displays the acquisition landscape of the two most important cell-viability determinants. In the latest model's acquisition landscape (Fig. 6a), the highest utility region remains concentrated around the ectoin/ethylene glycol neighborhood that had already produced strong wet-lab results, but the uncertainty panel also identifies nearby areas where additional tests would be most informative. This makes the next experimental decision less arbitrary: one can choose candidates trending toward the high-prediction region to confirm reproducibility, or choose candidates near the high-uncertainty/high-UCB boundary to test whether the optimum can be extended.

The interaction contours provide a complementary view. The script automatically selects the most important cell-viability determinant and compares its interactions with the next three most important determinants. In the latest model's interaction contour (Fig. 6b), the ethylene glycol/ectoin panel is the most directly actionable because it maps the main chemistry repeatedly selected by the model and validated in the wet lab; the high-response region suggests that future batches should refine this pair locally rather than scan the full formulation space again. The ethylene glycol/DMSO contour is also important because it helps define whether DMSO is acting as a necessary co-protectant or whether low-DMSO formulations can remain viable when ethylene glycol and ectoin are tuned appropriately. Similarly, the ethylene glycol/FBS contour highlights a potentially useful serum-dependent effect, suggesting that FBS may not simply be a background supplement but may modify the concentration range in which ethylene glycol performs well.

Together, these two types of figures support a practical next-step strategy. The acquisition landscape identifies where to test next, while the interaction contours suggest why those tests are biologically and formulation-wise interesting. The most compelling follow-up experiments would therefore not be isolated single recipes, but small local grids around ectoin/ethylene glycol/FBS with controlled low-DMSO levels, plus boundary probes in regions where the model has high uncertainty but plausible acquisition value. Such experiments would directly test whether the apparent optimum is robust, whether DMSO can be further reduced, and whether the serum/ethylene glycol interaction represents a reproducible formulation effect rather than a batch-specific artifact.

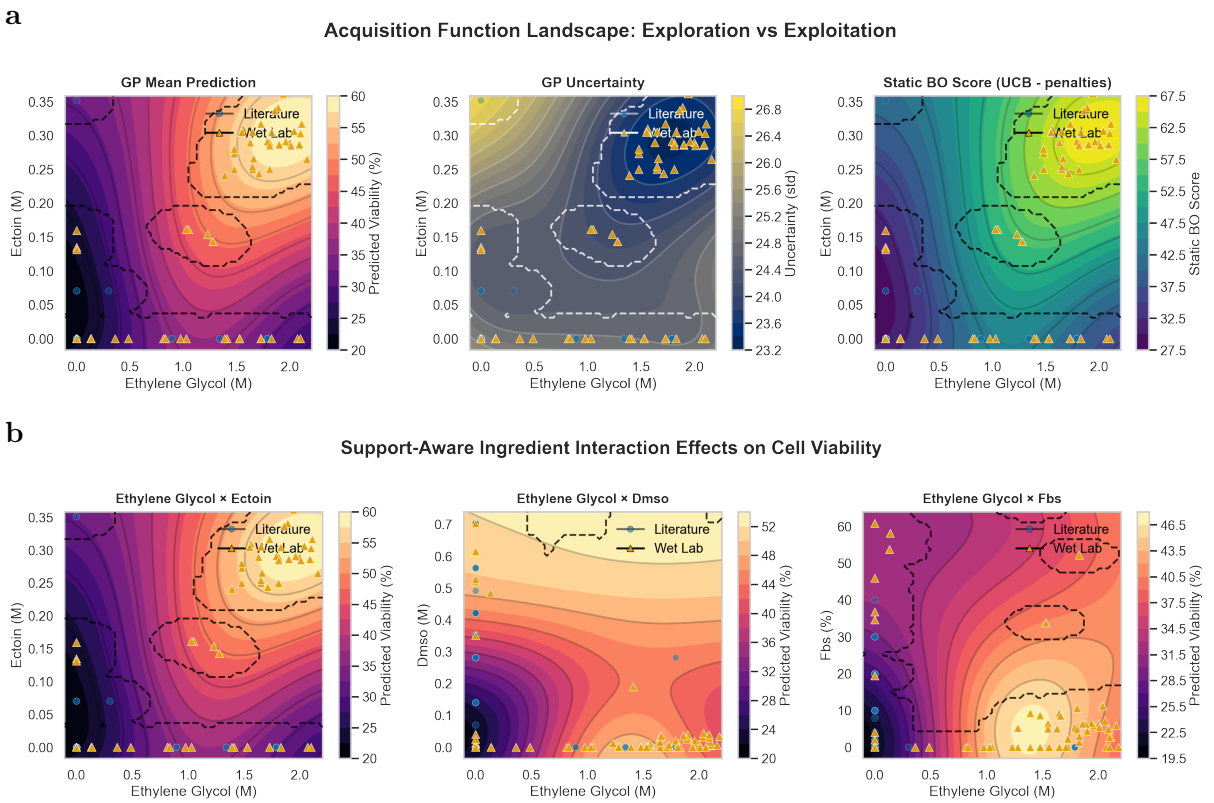

Fig. 6 — Acquisition and interaction landscapes after updating the iteration-9 model: a, predicted mean viability, predictive uncertainty, and static Bayesian-optimization score across the ectoin/ethylene glycol slice. The static Bayesian-optimization score combines UCB with implementation penalties and is therefore not identical to pure UCB (Supplementary Methods); b, support-aware interaction contours for selected ingredient pairs. Filled contour colors indicate model-predicted viability on the displayed two-feature slice while other ingredients are held at the plotting baseline. Dashed boundaries indicate the support-aware region used to distinguish interpolation from extrapolation. Circles denote literature observations and triangles denote wet-lab observations.

4 Discussion

Owing to the rapid development of large language models (LLMs), this study demonstrates that the lack of in-house computational specialization is no longer an absolute barrier to using machine learning in experimental formulation research. With a single initial prompt, Claude Opus 4.5 generated a runnable baseline repository that translated a scientific objective into data parsing, Gaussian-process modeling, Bayesian optimization with constrained candidate generation, validation templates, and written documentation. Subsequent work with GPT-5.3-codex and human review then converted that first scaffold into a more rigorous closed-loop workflow with iteration-aware model checkpoints, wet-lab-weighted residual correction, explainability plots, and batch recommendation. The important point is that the first answer was useful enough as an executable starting point for a real experimental campaign.[9, 10, 11]

This mode of AI assistance is particularly valuable in projects like cryomicroneedle formulation discovery because the computational burden is broad rather than defined by one algorithm. The workflow required many small but connected engineering tasks: synonym handling, unit conversion, duplicate management, model serialization, candidate filtering, uncertainty reporting, wet-lab data ingestion, stage-level evaluation, and figure generation. These are tasks that can consume substantial researcher time even when the underlying scientific question is clear. In our case, the AI-generated infrastructure helped turn a manually curated literature table into a sequential optimization system that ultimately identified simple two- to three-component formulations with high post-thaw viability. These simple formulation results were not what we initially expected given the 10-ingredient limit, but it was a natural optimization outcome; in this small-data setting, high-order combinations are poorly supported and mostly extrapolative, so the model naturally favors sparse formulations where the viability signal is better constrained. In the end, we are happy to see that the assistance provided by LLM agents was not only faster coding, but also a faster conversion of biological intent into a functional and reproducible computational workflow.

Still, the present work is not a demonstration of full scientific autonomy. Literature search and extraction persisted as major human bottlenecks. Search application programming interfaces (APIs) such as Semantic Scholar and OpenAlex lower the cost of finding papers, but they do not guarantee that the relevant corpus is complete, that full text is accessible, or that experimental details are preserved in a machine-readable form.[12, 13] Many useful cryopreservation papers reported formulations in tables, scanned PDFs, supplemental files, or inconsistent prose that make rule-based automation error-prone. Literature search and extraction, even with LLM agents, can confuse absolute viability with viability normalized to a control, miss units, merge chemically distinct additives, or treat basal media as active formulation variables. These errors change the training data and therefore change the experiments suggested by the model. For this reason, manual curation and domain-aware checking remained necessary.

Apart from wet-lab validation, which remained fundamentally human and physical, manuscript writing continued to be a human responsibility. LLM agents were helpful for drafting, restructuring, checking code history, and linking computational outputs to figures, but the authors still had to decide what the study could honestly claim. That included distinguishing a proof-of-concept result from a clinic-ready formulation, arguing how useful the results are and how they should be interpreted, as well as describing limitations without

overstating the autonomy of the system. In scientific writing, the hardest task may not be producing sentences, but deciding which claims are justified and presentable.

On the experimental front, the most important limitation is that post-thaw viability did not by itself predict cryomicroneedle usability, partly because the present optimization was single-objective. In our validations, most of the formulations that exceeded 70% cell viability did not form an intact cryomicroneedle patch. A plausible reason is that the formulation chemistry that protects cells during freezing can conflict with the formulation chemistry needed to form a mechanically stable frozen microneedle. High concentrations of permeating cryoprotectants, serum, polymers, and osmolytes can depress the freezing point, increase the unfrozen liquid fraction, alter ice-crystal structure, change viscosity, and/or weaken the frozen matrix. These effects may preserve cells but prevent formation of a rigid, continuous needle array capable of demolding, handling, and inserting into skin. As a result, the high viability we observed should be interpreted as a biological success rather than a complete cryomicroneedle formulation success. The model explicitly optimized post-thaw viability while using DMSO limits and ingredient-count constraints as guardrails, but it did not optimize mechanical strength, patch integrity, and insertion efficiency. This workflow was appropriate for a first proof-of-concept, but it also means that the model could have favored formulations that are biologically protective yet physically unsuitable for cryomicroneedle delivery.[4, 5, 1]

Future work should therefore move from a viability-only formulation search to a multi-objective active-learning workflow. The next model should treat post-thaw viability, intact patch formation, fracture force, and successful skin insertion as coupled outcomes.[6, 8] Some endpoints could be modeled as continuous objectives, whereas others, such as intact patch formation, may be very useful as a hard feasibility filter. Safety constraints should also be added explicitly, including stricter bounds on ethylene glycol and DMSO, dose-normalized exposure estimates, and searches for chemically safer substitutes that preserve the beneficial ectoin/ethylene-glycol region without carrying the same translational burden.[5] We hypothesize that a diversification of ingredients can have a cooperative effect in reducing the concentration of each ingredient, and that a reduced concentration across the board will render forming intact cryomicroneedles tenable.

References

[1] Hao Chang, Sharon W. T. Chew, Mengjia Zheng, Daniel Chin Shiuan Lio, Christian Wiraja, Yu Mei, Xiaoyu Ning, Mingyue Cui, Aung Than, Peng Shi, Dongan Wang, Kanyi Pu, Peng Chen, Haiyan Liu, and Chenjie Xu. Cryomicroneedles for transdermal cell delivery. *Nature Biomedical Engineering*, 5(9):1008–1018, May 2021. doi: 10.1038/s41551-021-00720-1.

[2] Mengjia Zheng, Tianli Hu, Yating Yang, Xuan Qie, Huaxin Yang, Yuyue Zhang, Qizheng Zhang, Ken-Tye Yong, Wei Liu, and Chenjie Xu. In situ-formed cryomicroneedles for intradermal cell delivery. *NPG Asia Materials*, 16, February 2024. doi: 10.1038/s41427-024-00531-1.

[3] Shubhmita Bhatnagar, Kaushalkumar Dave, and Venkata Vamsi Krishna Venuganti. Microneedles in the clinic. *Journal of Controlled Release*, 260:164–182, August 2017. doi: 10.1016/j.jconrel.2017.05.029.

[4] David Whaley, Kimia Damyar, Rafal P. Witek, Alan Mendoza, Michael Alexander, and Jonathan R. T. Lakey. Cryopreservation: An overview of principles and cell-specific considerations. *Cell Transplantation*, 30, 2021. doi: 10.1177/0963689721999617.

[5] Kathryn A. Murray and Matthew I. Gibson. Chemical approaches to cryopreservation. *Nature Reviews Chemistry*, 6(8):579–593, August 2022. doi: 10.1038/s41570-022-00407-4.

[6] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, January 2016. doi: 10.1109/JPROC.2015.2494218.

[7] Peter I. Frazier. A tutorial on bayesian optimization. *CoRR*, abs/1807.02811, 2018. doi: 10.48550/arXiv.1807.02811.

[8] Gary Tom, Stefan P. Schmid, Sterling G. Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M. Rajaonson, Marta Skreta, Naruki Yoshikawa, Samantha Corapi, Gun Deniz Akkoc, Felix Strieth-Kalthoff, Martin Seifrid, and Alán Aspuru-Guzik. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024. doi: 10.1021/acs.chemrev.4c00055.

[9] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents. *CoRR*, abs/2402.01030, 2024. doi: 10.48550/arXiv.2402.01030.

[10] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37, 2024. doi: 10.48550/arXiv.2405.15793.

[11] Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651(8107):914–919, March 2026. doi: 10.1038/s41586-026-10265-5.

[12] Suzanne Fricke. Semantic scholar. *Journal of the Medical Library Association*, 106(1):145–147, 2018. doi: 10.5195/jmla.2018.280.

[13] Jason Priem, Heather A. Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *CoRR*, abs/2205.01833, 2022. doi: 10.48550/arXiv.2205.01833.