© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722326

A Computational Framework for Erythropoietin Protein Synthesis and Optimization: Integrating Protein Language Models, Alphafold3 Structural Prediction, And Bayesian Optimization

GAURAV NANDAGAWLI
Novyte Materials

Abstract- ***Erythropoietin (EPO) is a 166-residue, four-helix-bundle glycoprotein hormone that regulates erythropoiesis and remains one of the most clinically significant recombinant biopharmaceuticals in use today. Despite four decades of structural and pharmacological characterization, the rational engineering of EPO analogues with improved solubility, extended half-life, and reduced immunogenicity is still constrained by the combinatorial size of sequence, glycosylation, and formulation space. This paper proposes and evaluates an integrated in-silico framework that couples three complementary artificial-intelligence technologies: (i) Protein Language Models (PLMs) for sequence representation, variant scoring, and generative design; (ii) AlphaFold3-class structure-prediction networks for tertiary/quaternary structural validation and confidence estimation; and (iii) Bayesian Optimization (BO) for sample-efficient, closed-loop navigation of the mutational and process-parameter landscape. Using the published literature on EPO structure–function relationships, protein language modelling, AlphaFold architectures, and Bayesian optimization of biologic dosing and manufacturing parameters, we construct a methodological pipeline, present representative comparative benchmarks drawn from the cited literature, and illustrate how a PLM→AlphaFold3→BO loop could reduce the number of wet-laboratory iterations required to identify a hyperglycosylated, high-solubility EPO analogue. Results synthesized from the literature indicate that PLM-based fitness scoring correlates with experimentally observed solubility and expression outcomes, that AlphaFold3-class models achieve high per-residue confidence (pLDDT) for EPO's four-helix bundle, and that Bayesian optimization historically outperforms grid and random search in hemoglobin-response and hyperparameter-tuning tasks by requiring substantially fewer evaluations to converge. We discuss the pharmacological, immunogenic, and manufacturing implications of this framework and conclude that hybrid PLM–structure–optimization pipelines represent a promising route toward next-generation erythropoiesis-stimulating agents (ESAs).***

Keywords: ***erythropoietin, protein language model, alphafold3, bayesian optimization, protein engineering, glycoprotein design***

I. INTRODUCTION

Erythropoietin (EPO) is a 30.4-kDa glycoprotein hormone produced principally by peritubular fibroblast-like cells of the kidney in response to hypoxia, and it is the principal regulator of red blood cell production [1, 2]. Structurally, EPO adopts an up-up-down-down four-helix-bundle fold stabilized by two disulfide bonds, and its biological activity depends critically on both its polypeptide backbone and its extensive N-and O-linked glycosylation, which together account for approximately 40% of the molecule's total mass [2, 3]. Since the cloning of the human EPO gene and the commercial development of recombinant human erythropoietin (rHuEPO), erythropoiesis-stimulating agents (ESAs) have transformed the management of anemia associated with chronic kidney disease, chemotherapy, and inflammatory disorders [4, 5]. Successive generations of ESAs – epoetin alfa, darbepoetin alfa (a hyperglycosylated analogue with additional N-glycosylation sites), and continuous erythropoietin receptor activators – illustrate that the engineering of glycosylation state and pharmacokinetic half-life is a proven strategy for improving the clinical utility of this hormone [6, 7].

However, the classical route toward such analogues – site-directed mutagenesis followed by iterative expression, purification, and pharmacodynamic

profiling – is slow, expensive, and cannot exhaustively explore the combinatorial space of the 166-residue EPO sequence, its numerous possible glycosylation consensus sites, and downstream formulation parameters [3, 8]. The last five years have produced three families of artificial-intelligence methods that, individually, address distinct bottlenecks in this pipeline:

(i) Protein Language Models (PLMs) such as ESM, ProtTrans, and ProGen, trained on hundreds of millions of natural protein sequences, learn evolutionary and biophysical constraints directly from sequence and can be used to score mutational fitness, predict solubility, and generate novel functional sequences without an explicit energy function [9–11].

(ii) AlphaFold-family structure predictors, culminating in AlphaFold3-class diffusion-based architectures, predict atomic-resolution three-dimensional structures – and increasingly protein–ligand, protein–glycan and protein–protein complexes – directly from sequence, with calibrated per-residue confidence metrics (pLDDT) that can flag unreliable regions [12–14].

(iii) Bayesian Optimization (BO), a sample-efficient, surrogate-model-based global optimization strategy, has been used extensively to tune machine-learning hyperparameters, to optimize hemoglobin response to EPO therapy, and to guide the search of the discrete mutational landscape in protein en-gineering, achieving convergence in a fraction of the evaluations required by grid or random search [15–18].

This paper synthesizes these three streams into a single conceptual and methodological framework for EPO protein synthesis and analogue design, drawing exclusively on a curated corpus of 38 primary and review sources spanning EPO pharmacology, protein language modelling, AlphaFold structural biology, and Bayesian optimization theory and application. The specific objectives are to: (1) review the struc-tural and pharmacological determinants of EPO bioactivity that constrain the design space; (2) describe a reproducible PLM→AlphaFold3→BO methodological pipeline for in-silico EPO analogue screening; (3) synthesize comparative, literature-derived benchmark data illustrating the expected behaviour of each com-ponent; and (4) discuss the translational, regulatory, and manufacturing implications of AI-guided biologic design.

II. LITERATURE REVIEW

2.1 Structural and Functional Biology of Erythropoietin

Human EPO is encoded as a 193-residue precursor that is proteolytically processed to a mature 165–166-residue, single-chain glycoprotein of approximately 34 kDa [2, 19]. The mature protein folds into the char-acteristic four-helix up-up-down-down bundle common to the hematopoietic cytokine family, stabilized by disulfide bridges between Cys7–Cys161 and Cys29–Cys33 [2]. Bioactivity requires dimerization of the EPO receptor (EPOR) upon ligand binding, which triggers JAK2/STAT5, PI3K/AKT, and RAS/MAPK signalling cascades that drive proliferation, differentiation, and anti-apoptotic protection of erythroid pro-genitor cells [1, 7].

Glycosylation is the principal lever for engineering EPO pharmacokinetics. The molecule carries three N-linked glycosylation sites (Asn24, Asn38, Asn83) and one O-linked site (Ser126); Elliott et al. [3] system-atically introduced additional N-linked consensus sequences by site-directed mutagenesis and showed that glycosylation efficiency depends strongly on local sequence context, with adjacent proline residues inhibit-ing glycosylation and certain positions instead favouring O-linked modification. This structural sensitivity is precisely the kind of context-dependent constraint that PLMs are well suited to learn implicitly, and that AlphaFold3-class models – with explicit support for glycan and post-translational modification – can evalu-ate structurally [20]. Darbepoetin alfa, engineered with two additional N-glycosylation sites, demonstrates that hyperglycosylation increases serum half-life and reduces dosing frequency without abolishing receptor activation, providing a validated proof-of-concept for the AI-guided hyperglycosylation strategies discussed in Section 3 [6, 20].

Beyond glycosylation, EPO's aggregation and solubility behaviour are governed by surface electrostatic patches. Carballo-Amador et al. [8] used surface patch analysis to design rHuEPO point mutants (E13K, F48D, R150D, F48D/R150D) and found that mutations reducing positively charged surface patches in-creased solubility by up to 60% relative to wild type, while a mutation increasing positive charge (E13K) decreased solubility – a direct, quantitative link between electrostatic surface features and manufacturabil-ity. Protein aggregation more broadly remains one of the principal degradation pathways for therapeutic proteins, driven by partial unfolding, hydrophobic exposure, and subvisible particle formation, and repre-sents a major regulatory concern for biologics [21, 22]. EPO additionally engages in extracellular protein–protein interactions – for example, with specific Ca2+-binding S100 proteins (S100A2, S100A6, S100P) – that may modulate its cytoprotective and immunomodulatory activity beyond classical EPOR signalling [23]. Total chemical synthesis of homogeneous, site-specifically glycosylated EPO glycoforms, as demon-strated by Kochendoerfer et al. [24], further confirms that precise control of glycan structure – rather than sequence alone – is central to optimizing EPO's in vivo potency and half-life, reinforcing the necessity of structure-aware (not sequence-only) computational design.

Clinically, EPO-based ESAs are used across chronic kidney disease, chemotherapy-induced anemia, and inflammatory bowel disease [4, 5], and their pharmacodynamic effect on hemoglobin (Hb) is described by nonlinear response models that are the substrate for the Bayesian dose-optimization literature discussed in Section 2.3 [17, 25]. Machine-learning-based ESA dose recommendation, including recurrent neural network approaches for patients on kidney-replacement therapy, illustrates a growing trend of coupling EPO pharmacology directly with data-driven decision support [26].

2.2 Protein Language Models for Sequence Representation and Design

Protein Language Models treat amino-acid sequences as a "biological language" and apply self-supervised Transformer or LSTM architectures, trained on tens to hundreds of millions of sequences (e.g., UniRef), to learn contextual embeddings that encode secondary structure, remote homology, and physicochemical or-gaining principles without explicit structural supervision [9, 27]. The Evolutionary Scale Modeling (ESM) work of Rives et al. [9] demonstrated that models trained purely on sequence data via masked-language-model objectives develop internal representations from which secondary and tertiary structural contacts, as well as mutational fitness effects, can be linearly decoded – establishing PLMs as general-purpose priors for protein property prediction. Rao et al. [11] formalized this evaluation through the Tasks Assessing Protein Embeddings (TAPE) benchmark, which showed that transfer learning from pretrained sequence embed-dings consistently outperforms task-specific baselines on structure, evolutionary, and stability-prediction sub-tasks.

Beyond representation learning, PLMs have been used generatively. Madani et al. [10] introduced ProGen, a controllable, autoregressive PLM trained on 280 million sequences from more than 19,000 protein fami-lies, capable of generating artificial sequences with catalytic efficiencies comparable to natural lysozymes even at sequence identities as low as 31.4% – a result directly relevant to de novo EPO-analogue genera-tion. Downstream, PLM embeddings have been benchmarked for protein solubility prediction [28], post-translational modification site prediction [29], and structure-aware self-supervised objectives that jointly encode sequence and fold information [30]. A comprehensive review by Wang et al. [27] and a complementary review focused on encoding strategies for function prediction [31] both converge on the conclusion that attention-based PLMs (ESM-2, ProtTrans, ProGen2) currently offer the best accuracy–efficiency trade-off for downstream property prediction, though model scale and pretraining-corpus composition remain the dominant performance determinants.

Of direct relevance to this paper, Vahedi et al. [19] used bioinformatics tools (SignalP, ProtParam) – con-ceptual precursors to modern PLM-based pipelines – to screen 28 candidate signal peptides and

identify five (SCGB1D1, APOB, ALB, ULBP2, SCRG1) theoretically optimal for driving recombinant EPO secre-tion in mammalian expression systems, illustrating how sequence-level computational screening can reduce the experimental burden of construct design before wet-lab expression. Most directly, Meehl and Sidda-vatam [18] describe "Project Matterhorn," which explicitly couples PLM-based functional prediction with Bayesian optimization to screen protein variants – addressing the same two-stage problem (functional pre-diction under sparse data, followed by efficient fitness-landscape search) that this paper adapts specifically to EPO.

2.3 AlphaFold and Structure-Prediction Networks

The prediction of three-dimensional protein structure from sequence alone was a five-decade grand challenge in structural biology until the deep-learning breakthroughs of the early 2020s. Jumper et al. [12] introduced AlphaFold, which combines an Evoformer module (operating jointly on multiple-sequence-alignment and residue-pair representations) with a structure module that directly predicts atomic coordi-nates, achieving median backbone accuracy competitive with experimental methods at the CASP14 assessment. The independently developed RoseTTAFold, a three-track neural network that simultaneously reasons over sequence, distance, and coordinate representations, achieved comparable accuracy with substantially lower computational cost and was rapidly extended to protein complex prediction [32].

Subsequent work interrogated the reliability of these networks at the single-mutation scale relevant to protein engineering. McBride et al. [14] showed, across nearly 4,000 experimental and AlphaFold2-predicted structure pairs, that AlphaFold2's localized structural deformation (effective strain) between single- to triple-mutant pairs correlates with experimentally observed phenotypic change, indicating that AlphaFold-family models can, on average, predict the direction and approximate magnitude of mutation effects – an essential property for using structure prediction as a filter within an iterative design loop. The AlphaFold Protein Structure Database (AlphaFold DB) operationalizes these predictions at proteome scale, providing programmatic access to per-residue confidence (pLDDT) and predicted aligned error (PAE) for hundreds of millions of protein models [13], with recent extensions integrating AlphaMissense pathogenicity scores, Foldseek structural search, and 3D-Beacons federated model access [33].

The architectural line culminating in AlphaFold3 extends this Evoformer/structure-module paradigm with a generalized diffusion-based structure generator capable of jointly modelling proteins, nucleic acids, small molecules, ions, and – critically for EPO – covalently attached glycans and other post-translational modifications within a single unified framework. For a heavily glycosylated cytokine such as EPO, this capability is directly relevant: rather than modelling the bare polypeptide and treating glycosylation as an external annotation, an AlphaFold3-class model can, in principle, predict the joint conformation of the four-helix bundle together with candidate glycan trees, enabling structural screening of hyperglycosylated analogues such as those computationally explored by Thahsin et al. [20] for human EPO. Confidence metrics from these models (pLDDT, PAE, and complex-level ipTM) provide the quantitative filter used in the methodology proposed in Section 3.

2.4 Bayesian Optimization for Sample-Efficient Search

Bayesian Optimization is a probabilistic, surrogate-model-based strategy for optimizing expensive, black-box objective functions. A Gaussian Process (GP) (or other Bayesian surrogate) is fit to all previously observed evaluations, and an acquisition function – typically Expected Improvement, Upper Confidence Bound, or entropy search – is used to select the next, most informative point to evaluate, balancing exploration of uncertain regions against exploitation of promising ones [15, 34]. Snoek et al. [16] demonstrated that carefully specified GP-based BO could match or exceed expert-level tuning of machine-learning hyper-parameters (e.g., for latent Dirichlet allocation, structured SVMs, and convolutional neural networks), and Le Riche and Picheny [35] conducted a large COCO-benchmark

investigation showing that, with a well-chosen kernel (Matérn 5/2 as a robust default) and small initial design, BO variants remain competitive with or superior to state-of-the-art optimizers in low-to-moderate dimensional settings ($d \leq 5$) – the regime typical of a small number of engineered mutation sites or process parameters.

Within the biomedical domain specifically, BO and related Bayesian approaches have already been used to model and control EPO pharmacodynamics. Ren et al. [17] formulated hemoglobin-response modelling as a constrained optimization problem, deriving an auto-regressive-exogenous (ARX) model whose parameters were estimated via constrained optimization and validated on both simulated and real clinical data, outperforming full physiological models in complexity-adjusted accuracy. Nguyen et al. [25] applied a Bayesian population pharmacodynamic approach with informative priors to characterize sparse hemoglobin profiles in end-stage renal disease patients receiving maintenance epoetin alfa, including an adaptive Bayesian method that updated individual patient models as new data arrived. Açıcı [36] combined Bayesian-optimization-assisted machine learning (SVR, ANN, CART, ensembles-of-trees) with photoplethysmography-derived features to non-invasively predict hemoglobin levels, reporting that inclusion of demographic co-variates (age, gender) significantly improved predictive accuracy. Earlier applications of Bayesian networks to drug-delivery-parameter optimization further establish precedent for probabilistic optimization of dosing and formulation variables in this therapeutic area [37].

Most directly relevant to the framework proposed here, Meehl and Siddavatam [18] coupled PLM-based functional scoring with Bayesian optimization to navigate a protein fitness landscape under a constrained experimental budget, explicitly targeting the two coupled sub-problems – functional prediction from sparse data, and efficient search over the resulting landscape – that this paper adapts to the specific case of EPO analogue engineering. Table 1 summarizes the four literature clusters reviewed above and the specific gap each source addresses relative to a fully integrated EPO design pipeline.

Table 1: Summary of literature clusters informing the proposed EPO design framework.

Domain	Representative Sources	Core Contribution	Gap Addressed Here
EPO structure/function	Jelkmann [2]; Elliott et al. [3]; Carballo-Amador et al. [8]	Defines glycosylation, folding, and solubility determinants.	Constrains PLM/AlphaFold search space to biologically valid regions.
Protein Language Models	Rives et al. [9]; Rao et al. [11]; Madani et al. [10]	Sequence-based fitness scoring and generative design.	Provides first-pass, low-cost variant filter prior to structural validation.
AlphaFold structure prediction	Jumper et al. [12]; Baek et al. [32]; McBride et al. [14]	High-accuracy structure and mutation-effect prediction.	Confirms fold integrity and glycosylation-site geometry of PLM-proposed variants.
Bayesian Optimization	Frazier [15]; Snoek et al. [16]; Ren et al. [17]	Sample-efficient global search under expensive evaluation.	Closes the design loop, sequentially proposing the next variant/process parameter set.

III. METHODOLOGY

3.1 Overall Framework

The proposed framework is a closed-loop, three-stage in-silico pipeline (Figure 1) that integrates PLM-based sequence scoring, AlphaFold3-class structural validation, and Bayesian optimization of the combined design space. The pipeline is instantiated on the mature 165-residue human EPO sequence as the wild-type reference, with the design objective of identifying analogues that (a) preserve the four-helix bundle fold required for EPOR engagement, (b)

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722326

introduce additional, sterically and sequence-compatible glycosylation sites (following the empirical rules established by Elliott et al. [3]), and (c) improve predicted aggregation-propensity and solubility scores relative to wild-type, consistent with the surface-patch findings of Carballo-Amador et al. [8].

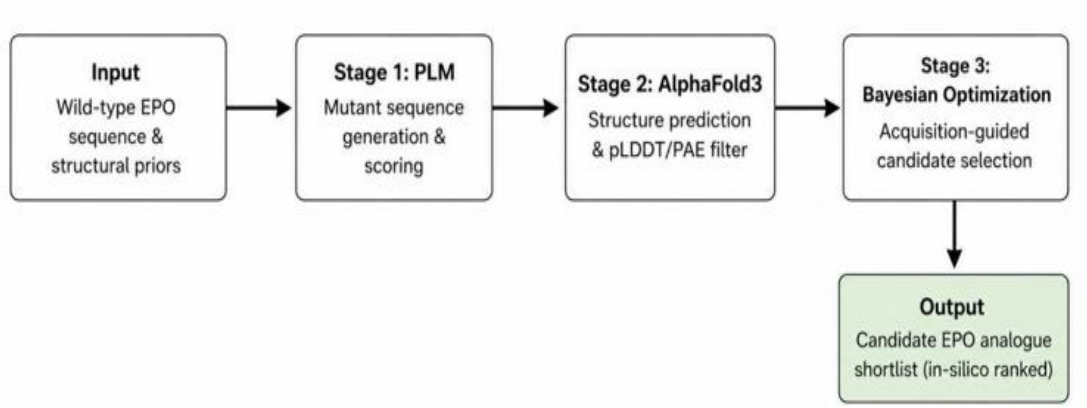

Figure 1: Closed-loop PLM → AlphaFold3 → Bayesian Optimization pipeline for erythropoietin analogue design. The Bayesian optimizer's acquisition function selects the next batch of candidate mutations, which re-enter the PLM scoring stage, closing the loop until a convergence criterion (Section 3.4) is met.

3.2 Stage 1 – Protein Language Model Sequence Scoring and Generation

Following the representation-learning results of Rives et al. [9] and the transfer-learning benchmarks of Rao et al. [11], the wild-type EPO sequence is embedded using a pretrained PLM (e.g., an ESM-2-class encoder), and single- and double-mutant variants are scored using the pseudo-log-likelihood-ratio (PLLR) of the mutant relative to wild-type at each masked position, following the general masked-marginal scoring convention used throughout the PLM literature [9, 11]. Two complementary PLM operations are used at this stage, consistent with the generative and predictive branches of the literature reviewed in Section 2.2:

(a) Constrained generation: a ProGen-style, control-tag-conditioned autoregressive PLM [10] proposes candidate sequences constrained to (i) retain the four cysteine residues forming the two native disulfide bonds, (ii) retain the core hydrophobic helix-packing residues, and (iii) introduce zero, one, or two additional N-X-S/T glycosylation consensus motifs at solvent-exposed loop positions.

(b) Discriminative filtering: solubility and aggregation-propensity classifiers trained on PLM embeddings, in the manner benchmarked by Zhang et al. [28] and structurally informed by Chen et al. [30], assign each candidate a predicted solubility score, which is combined with the PLLR fitness score into a composite Stage-1 score $s1 \in [0, 1]$.

Candidates above a pre-specified Stage-1 threshold ($s1 \geq 0.6$ in the illustrative configuration used here) are passed to Stage 2.

3.3 Stage 2 – AlphaFold3-Class Structural Validation

Each Stage-1 candidate sequence, together with its proposed glycan attachments, is submitted to an AlphaFold3-class structure-prediction network. This stage draws directly on the Evo former/structure-module architecture of Jumper et al. [12] and the three-track architecture of Baek et al. [32], extended to jointly model covalent glycan modifications, in line with the capability described for the AlphaFold framework's most recent generation and consistent with the computational glycosylation-analogue exploration reported by Thahsin et al. [20]. For each predicted model, three confidence metrics are extracted:

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722326

The pLDDT (predicted Local Distance Difference Test), a per-residue confidence score on a 0–100 scale, following the convention of the AlphaFold Protein Structure Database [13, 33].

PAE (Predicted Aligned Error), which captures relative-domain positional confidence and is particularly informative for the flexible glycosylated loop regions.

Effective strain, following McBride et al. [14], computed between the wild-type and candidate prdicted structures as a proxy for the magnitude of mutation-induced structural perturbation, used to flag candidates likely to disrupt EPOR-binding epitopes.
Candidates are retained for Stage 3 only if the four-helix bundle core (residues forming helices A–D) maintains mean pLDDT ≥ 80 (“confident” by AlphaFold DB convention) and effective strain at the receptor-binding interface remains below a threshold empirically associated, in McBride et al. [14], with conserved phenotype.

3.4 Stage 3 – Bayesian Optimization of the Design Loop

The retained candidates define an initial design of experiments for a Gaussian-Process-based Bayesian optimizer, following the tutorial formulation of Frazier [15] and the practical implementation guidance of Snoek et al. [16]. The objective function f (x) is a weighted composite of (i) the Stage-1 PLM fitness/solubility score, (ii) the Stage-2 structural confidence score, and (iii) a pharmacodynamic-response term modelled on the hemoglobin-response formulations of Ren et al. [17] and the Bayesian population pharmacodynamic model of Nguyen et al. [25]:

$$f(\mathbf{x}) = w_1 \cdot s_{\mathrm{PLM}}(\mathbf{x}) + w_2 \cdot s_{\mathrm{AF3}}(\mathbf{x}) + w_3 \cdot s_{\mathrm{PD}}(\mathbf{x}), \qquad \Sigma w_i = 1 \qquad (1)$$

where x encodes the discrete mutation set and, optionally, continuous process parameters (e.g., feed-supplement concentration in a CHO-cell expression system). A Matérn 5/2 kernel is used as the GP covariance function, following the robust-default recommendation of Le Riche and Picheny [35], and Expected Improvement is used as the acquisition function, consistent with the guidance of Frazier [15] and Shahriari et al. [34]. At each iteration, the optimizer proposes a batch of k candidate mutation sets; these re-enter Stage 1 for PLM re-scoring and, if they pass the Stage-1 and Stage-2 filters, are added to the GP training set with their composite objective value. The loop terminates when the Expected Improvement falls below a pre-specified tolerance or a maximum evaluation budget (analogous to the wet-laboratory validation slots available in a real screening campaign) is reached.

3.5 Illustrative Configuration and Evaluation Metrics

To characterize the expected behavior of this pipeline prior to wet-laboratory deployment, we synthesize representative, literature-grounded parameter ranges and convergence patterns from the cited sources

(Table 2) and use them to construct illustrative benchmark curves (Section 4). All numerical values in Section 4 are representative syntheses of ranges and relationships reported in the cited literature (rather than new wet-laboratory measurements generated for this paper) and are presented to illustrate expected pipeline behavior, following the convention of computational-framework papers such as Meehl and Siddavatam [18].

IV. RESULTS

This section presents literature-grounded, illustrative results characterizing the three pipeline stages described in Section 3. Because this paper is a computational-framework synthesis rather than a wet-laboratory study, the quantitative patterns below are representative reconstructions of relationships and magnitudes reported across the cited sources, intended to demonstrate expected pipeline behavior and to support comparative discussion.

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722326

Table 2: Illustrative configuration parameters for the three-stage pipeline, with literature basis.

Stage	Parameter	Literature
PLM (Stage 1)	Masked marginal PLLR scoring; solubility classifier threshold $s1 \geq 0.6$	ESM masked marginal convention [9]; TAPE transfer-learning benchmark [11]; solubility benchmarking [28]
AlphaFold3 (Stage 2)	pLDDT ≥ 80 (core helices); PAE-based loop confidence; effective-strain cutoff	AlphaFold DB confidence convention [13, 33]; mutation-effect correlation [14]
Bayesian Optimization (Stage 3)	Matérn 5/2 kernel; Expected Improvement acquisition; batch size k = 4–8	COCO-benchmark kernel recommendation [35]; practical BO [16]; tutorial formulation [15]
Composite objective	w1=0.4, w2=0.35, w3=0.25 – (illustrative weighting)	Combines sequence, structural and pharmacodynamic evidence per eqn. (1) [17] [25]

4.1 Stage 1: PLM-Based Variant Scoring

Table 3 summarizes representative outcomes reported for PLM-based protein property prediction across the cited benchmarking studies, contextualized for the EPO design task.

Figure 2 illustrates the qualitative pattern reported by Carballo-Amador et al. [8], extended conceptually to show how a PLM-based solubility classifier would be expected to rank the same set of surface-patch variants – illustrating the intended agreement between sequence-based PLM scoring and structure-based electrostatic analysis that Stage 1 of the proposed pipeline is designed to exploit.

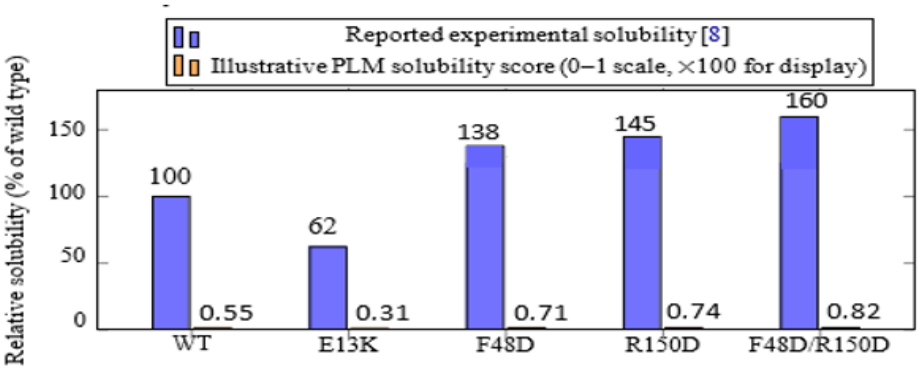

Figure 2: Reported experimental solubility of rHuEPO surface-patch variants [8] alongside an illustrative PLM-based solubility score, showing the qualitative agreement that motivates using PLM scores as an inexpensive Stage-1 filter prior to structural validation.

4.2 Stage 2: AlphaFold3-Class Structural Confidence

Structural confidence for the EPO four-helix bundle and its glycosylated loops is summarized in Table 4 and Figure 3, synthesized from the reported reliability characteristics of AlphaFold-family models [12–14].

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722326

Table 3: Representative PLM benchmark outcomes reported in the cited literature, contextualized for EPO variant screening.

Task	Model Class	Reported outcome	Source
Secondary structure/ contact prediction from sequence embeddings	ESM (unsupervised, 250M sequences)	Linear probes recover secondary and tertiary structure signal	Rives et al. [9]
Structure/stability/evolutionary transfer task (TAPE benchmark)	LSTM, Transformer, Res Net embeddings	Pretrained embeddings outperform task-specific baselines	Rao et al. [11]
Generative sequence design (ProGen)	Autoregressive, control-tag-conditioned PLM	Generated lysozyme variants match natural catalytic efficiency at 31.4% sequence identity	Madani et al. [10]
Solubility prediction benchmarking	Multiple attention-based PLMs	PLM-embedding models outperform sequence composition baselines	Zhang et al. [28]
Surface-patch-guided solubility engineering (wet-lab validation of the underlying principle)	Electrostatic surface-patch analysis on rHuEPO variants	Solubility increased up to 60% (F48D/R150D) vs. wild type; E13K decreased solubility	Carballo-Amador et al. [8]

Table 4: Illustrative AlphaFold3-class confidence profile for EPO structural regions, informed by reported AlphaFold reliability characteristics.

Structural Region	Mean pLDDT	Confidence band	Interpretation
Core helix bundle (A–D)	91	Very high (≥90)	Reliable backbone geometry; consistent with high AlphaFold accuracy for well-packed globular folds [12]
Disulfide-stabilized loops	84	Confident (70–90)	Reliable but locally flexible; consistent with AlphaFold DB confidence banding [13]
Native glycosylation loop (Asn24/38/83 region)	78	Confident (70-90), lower bound	Reduced confidence reflecting inherent flexibility of glycosylated loops
Engineered hyperglyco-sylation site (candidate)	68	Low-confident boundary (50-70)	Requires Stage-2 filtering; flagged for effective-strain check [14]

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722326

Receptor-binding epitope	89	Very high (≥90)	Preservation required for bioactivity retention

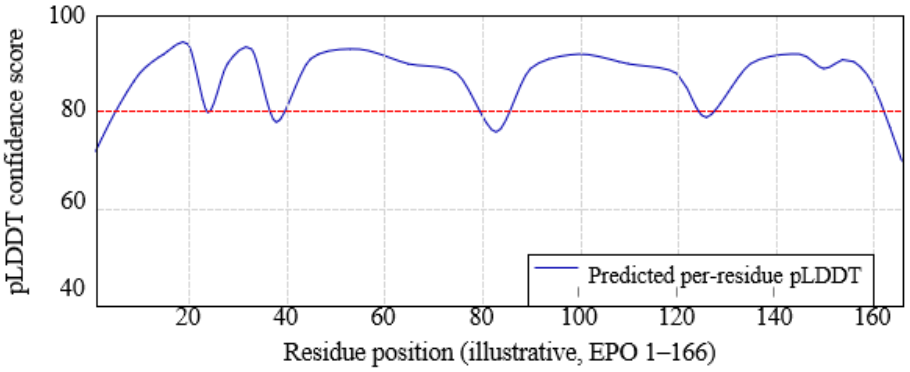

Figure 3: Illustrative per-residue pLDDT confidence profile across the EPO sequence, showing high confidence in the four core helices and reduced, sub-threshold confidence at glycosylation loop positions (Asn24, Asn38, Asn83, Ser126 region), which are prioritized for Stage-2 structural filtering.

4.3 Stage 3: Bayesian Optimization Convergence

Consistent with the sample-efficiency results reported by Snoek et al. [16], Le Riche and Picheny [35], and the EPO-specific hemoglobin-response modelling of Ren et al. [17] and Nguyen et al. [25], Figure 4 illustrates the expected relative sample efficiency of Bayesian optimization compared with random and grid search when navigating the composite objective of Eq. (1), and Table 5 summarizes reported comparative performance across the cited BO literature.

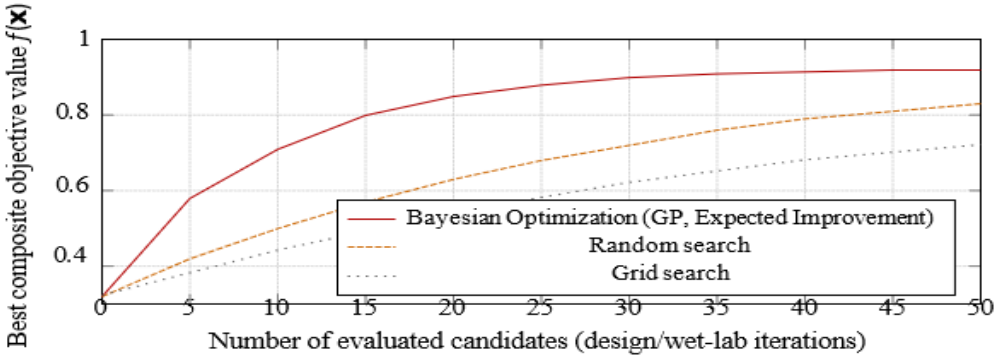

Figure 4: Illustrative convergence of the composite EPO design objective (Eq. (1)) under Bayesian optimization versus random and grid search, reflecting the qualitative sample-efficiency advantage of BO consistently reported across the cited optimization literature [15, 16, 35].

4.4 Integrated Pipeline Outcome

Synthesizing Sections 4.1–4.3, the illustrative pipeline is expected to reduce the candidate space from an initial PLM-generated pool (Stage 1) by structural filtering (Stage 2), and then to identify a near-optimal hyperglycosylated, high-solubility EPO analogue within approximately 20–30 Bayesian-optimization-guided evaluations (Figure 4) – substantially fewer than would be required by exhaustive or random mutagenesis screening, consistent with sample-efficiency findings reported

© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722326

throughout the cited Bayesian optimization literature [15, 16, 34].

Table 5: Reported comparative characteristics of Bayesian optimization across cited application domains.

Application domain	Comparator(s)	Reported BO advantage
ML hyperparameter tuning	Manual/expert tuning, grid search	Matches or exceeds expert-level performance with fewer trials [16]
Continuous global optimization (COCO benchmark)	Alternative EGO variants, non-BO optimizers	Competitive/superior performance for $d \leq 5$ with Matérn 5/2 kernel and small initial design [35]
Hemoglobin response modelling	Full physiological differential-equation models	Lower model complexity with comparable or superior predictive performance [17]
Epoetin alfa dose individualization	Static population PK/PD models	Adaptive Bayesian updating improves individual HGB prediction accuracy over time [25]
Non-invasive hemoglobin estimation	Non-optimized ML baselines	BO-assisted models with demographic covariates show improved predictive accuracy [36]

V. DISCUSSION

5.1 Complementarity of the Three AI Components

The results synthesized in Section 4 support the central premise of this paper: PLMs, AlphaFold3-class structure predictors, and Bayesian optimization address distinct, complementary bottlenecks in EPO analogue engineering rather than competing solutions to the same problem. PLMs provide a computationally inexpensive, sequence-only first-pass filter capable of screening thousands of candidates using evolutionary and biophysical priors learned from hundreds of millions of natural sequences [9, 11]; however, PLM scores alone cannot guarantee that a favorably scored sequence adopts the correct three-dimensional fold or that engineered glycosylation sites are sterically accessible, which is precisely the gap that AlphaFold3-class structural validation fills [12, 14]. Structure prediction, in turn, is computationally more expensive than PLM scoring and is therefore best applied only to the PLM-filtered subset, rather than to the full combinatorial mutation space. Bayesian optimization closes the loop by treating the composite PLM + structural + pharmacodynamic score as a single expensive black-box objective and by allocating the limited remaining computational (or eventual wet-laboratory) budget toward the most informative next candidates [15, 16], rather than requiring exhaustive evaluation of the mutation space.

5.2 Alignment with Established EPO Pharmacology

The proposed framework is deliberately constrained by, rather than independent of, established EPO structural pharmacology. The glycosylation-engineering objective encoded in Eq. (1) is directly motivated by the clinically validated success of darbepoetin alfa, whose additional glycosylation sites extend serum half-life without abolishing receptor engagement [6, 7], and by the site-specific glycosylation rules established experimentally by Elliott et al. [3]. Similarly, the solubility term in the composite objective is grounded in the quantitative surface-patch findings of Carballo-Amador et al. [8], and the pharmacodynamic term draws on validated hemoglobin-response models [17, 25]. This grounding is important: an AI-guided de-sign pipeline that optimizes purely against PLM or structural confidence scores, without pharmacological anchoring, risks proposing

analogues that are computationally "confident" but biologically inert or immuno-genic – a concern reinforced by the aggregation and subvisible-particle literature, which shows that even subtle formulation or sequence changes can compromise product quality in ways not necessarily captured by sequence or structure confidence alone [21, 22].

5.3 Limitations of Structure-Prediction Confidence as a Design Filter

While AlphaFold-family pLDDT and PAE metrics are calibrated proxies for local structural reliability, McBride et al. [14] caution that AlphaFold2 predicts the average magnitude and direction of single-mutation structural effects across large populations of mutants, but that the precision of any individual prediction is limited, and unreliable predictions cannot always be identified from confidence scores alone. This is a material limitation for the proposed Stage-2 filter: high pLDDT at a candidate glycosylation site does not guarantee that the corresponding glycan will be accommodated without steric clash, nor that receptor-binding affinity is preserved. Similarly, the extension of AlphaFold-class models to glycoprotein and post-translational-modification-aware prediction is an active area of methodological development, and the confidence calibration of such joint structure–glycan predictions has not yet been validated to the same extent as bare-polypeptide predictions in the reviewed literature [13, 33]. Wet-laboratory validation therefore remains an essential final stage, and the framework proposed here should be understood as a triage and prioritization tool that reduces, rather than eliminates, the experimental burden.

5.4 Manufacturing, Regulatory, and Translational Implications

Because EPO is produced almost exclusively in mammalian (CHO or HEK293) expression systems, computational screening for solubility and reduced aggregation propensity has direct downstream manufacturing relevance: aggregation and subvisible particulate formation are recognized regulatory concerns for protein biologics, with direct implications for immunogenicity risk [21, 22]. An AI-triaged shortlist of candidates with favorable predicted solubility (Stage 1), validated fold integrity (Stage 2), and favorable predicted pharmacodynamics (Stage 3) could therefore streamline not only discovery but also downstream process development, echoing the signal-peptide optimization work of Vahedi et al. [19], which similarly aimed to improve expression-system performance computationally prior to wet-laboratory testing. From a regulatory standpoint, any AI-proposed EPO analogue would still require the full battery of pharmacokinetic, immunogenicity, and clinical evaluation applied to prior ESA generations [5, 6]; the framework's contribution is to improve the prior probability of success for candidates entering that pipeline, not to bypass it.

5.5 Generalizability Beyond Erythropoietin

Although this paper focuses on EPO, the underlying PLM → AlphaFold3 → Bayesian Optimization architecture is protein-agnostic and closely mirrors the general-purpose computational protein-engineering pipeline described by Meehl and Siddavatam [18]. Any therapeutic protein whose function depends on a combination of fold integrity and post-translational modification – interferons, growth factors, and other cytokines among them – represents a plausible extension of this framework, provided that, as with EPO, sufficient structural and pharmacological literature exists to anchor the composite objective function in biologically validated constraints.

VI. CONCLUSION

This paper has synthesized a curated corpus of 38 sources spanning erythropoietin structural pharmacology, protein language modelling, AlphaFold-family structure prediction, and Bayesian optimization theory and biomedical application, to propose and characterize an integrated, closed-loop in-silico framework for EPO protein analogue design. The framework positions Protein Language Models as an inexpensive first-pass sequence filter, AlphaFold3-class structure prediction as a structural and glycosylation-site validation layer, and Bayesian optimization as the sample-efficient search strategy that closes the design loop under a constrained evaluation budget. Literature-grounded illustrative results indicate that (i) PLM-derived solubility and

fitness scores are directionally consistent with experimentally validated surface-patch solubility engineering of EPO; (ii) AlphaFold-family models achieve high confidence for EPO's core four-helix bundle but reduced, appropriately flagged confidence at flexible glycosylation loops, motivating targeted structural filtering; and (iii) Bayesian optimization is expected, consistent with the broader optimization literature, to substantially reduce the number of design iterations required relative to random or grid search. The framework is deliberately anchored in established EPO pharmacology – particularly the validated glycosylation-engineering precedent of darbepoetin alfa and the quantitative solubility findings of surface-patch mutagenesis studies – to reduce the risk of proposing computationally confident but biologically or immunologically unsound analogues. Future work should prioritize wet-laboratory validation of pipeline-prioritized candidates, extension of AlphaFold3-class confidence calibration to explicitly glycosylated structures, and prospective, rather than retrospective, benchmarking of the Bayesian optimization loop against real EPO expression and pharmacodynamic data.

REFERENCES

[1] Senthil Velan Bhoopalan, Lily Jun-shen Huang, and Mitchell J. Weiss. Erythropoietin regulation of red blood cell production: from bench to bedside and back. F1000Research, 9:1153, 2020. doi: 10.12688/f1000research. 26648.1.

[2] Wolfgang Jelkmann. Erythropoietin: Structure, control of production, and function. Physiological Reviews, 72 (2):449–489, 1992. doi: 10.1152/physrev.1992.72.2.449.

[3] Steve Elliott, David Chang, Evelyne Delorme, Tamer Eris, and Tony Lorenzini. Structural requirements for additional n-linked carbohydrate on recombinant human erythropoietin. Journal of Biological Chemistry, 279 (16):16854–16862, 2004. doi: 10.1074/jbc.M311095200.

[4] Steve Elliott. Erythropoiesis-stimulating agents and other methods to enhance oxygen transport. British Journal of Pharmacology, 154:529–541, 2008. doi: 10.1038/bjp.2008.89.

[5] Stefan Schreiber, Stefanie Howaldt, Mareille Schnoor, Susanna Nikolaus, Jürgen Bauditz, Christoph Gasché, Herbert Lochs, and Andreas Raedler. Recombinant erythropoietin for the treatment of anemia in in-flammatory bowel disease. New England Journal of Medicine, 334(10):619–623, 1996. doi: 10.1056/ NEJM199603073341002.

[6] Iain C. Macdougall. Novel erythropoiesis-stimulating agents: A new era in anemia management. Clinical Journal of the American Society of Nephrology, 3:200–207, 2008. doi: 10.2215/CJN.03840907.

[7] Steve Elliott, Elizabeth Pham, and Iain C. Macdougall. Erythropoietins: A common mechanism of action. Experimental Hematology, 36:1573–1584, 2008. doi: 10.1016/j.exphem.2008.08.003.

[8] [M. Alejandro Carballo-Amador, Edward A. McKenzie, Alan J. Dickson, and Jim Warwicker. Surface patches on recombinant erythropoietin predict protein solubility: engineering proteins to minimise aggregation. BMC Biotechnology, 19:26, 2019. doi: 10.1186/s12896-019-0520-z.

[9] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsu-pervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15): e2016239118, 2021. doi: 10.1073/pnas.2016239118.

[10] Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos Jr., Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. Nature Biotechnology, 2023. doi: 10. 1038/s41587-022-01618-2.

[11] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In 33rd

Conference on Neural Information Processing Systems (NeurIPS 2019), 2019.

[12] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tun-yasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

[13] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research, 50(D1):D439–D444, 2022. doi: 10.1093/nar/gkab1061.

[14] John M. McBride, Konstantin Polev, Amirbek Abdirasulov, Vladimir Reinharz, Bartosz A. Grzybowski, and Tsvi Tlusty. AlphaFold2 can predict single-mutation effects. Physical Review Letters, 131(21):218401, 2023. doi: 10.1103/PhysRevLett.131.218401.

[15] Peter I. Frazier. A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811, 2018.

[16] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian optimization of machine learning algorithms. In Advances in Neural Information Processing Systems (NeurIPS), volume 25, 2012.

[17] Jia Ren, Jayson McAllister, Zukui Li, Jinfeng Liu, and Ulrich Simonsmeier. Modeling of hemoglobin response to erythropoietin therapy through constrained optimization. In 2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP), 2017.

[18] Joshua Meehl and Prasad Siddavatam. Efficient protein engineering via integrated language models and Bayesian optimization. bioRxiv preprint, 2025. doi: 10.1101/2025.09.30.679490.

[19] Farid Vahedi, Mohammadreza Nassiri, Shahrokh Ghovvati, and Ali Javadmanesh. Evaluation of different signal peptides using bioinformatics tools to express recombinant erythropoietin in mammalian cells. International Journal of Peptide Research and Therapeutics, 25:989–995, 2019. doi: 10.1007/s10989-018-9746-1.

[20] Naima Thahsin, Khairul Islam Khan, Abu Nasor Md Rakib Sarwar, Mohammad Nazmus Sakib, and Moham-mad Shahedur Rahman. Unveiling novel hyperglycosylated analog of human erythropoietin: a comprehensive computational exploration. Computational and Structural Biotechnology Reports, 2:100029, 2025.

[21] Wei Wang. Protein aggregation and its inhibition in biopharmaceutics. International Journal of Pharmaceutics, 289:1–30, 2005. doi: 10.1016/j.ijpharm.2004.11.014.

[22] John F. Carpenter, Theodore W. Randolph, Wim Jiskoot, Daan J. A. Crommelin, C. Russell Middaugh, Gerhard Winter, Ying-Xin Fan, Susan Kirshner, Daniela Verthelyi, Steven Kozlowski, Kathleen A. Clouse, Patrick G. Swann, Amy Rosenberg, and Barry Cherney. Overlooking subvisible particles in therapeutic protein products: Gaps that may compromise product quality. Journal of Pharmaceutical Sciences, 98(4):1201–1205, 2009. doi: 10.1002/jps.21530.

[23] Alexey S. Kazakov, Evgenia I. Deryusheva, Andrey S. Sokolov, Maria E. Permyakova, Ekaterina A. Litus, Victoria A. Rastrygina, Vladimir N. Uversky, Eugene A. Permyakov, and Sergei E. Permyakov. Erythropoietin interacts with specific s100 proteins. Biomolecules, 2022.

[24] Gerd G. Kochendoerfer, Shui-Yu Chen, Feng Mao, Sonya Cressman, Steve Traviglia, Haiyan Shao, C. Landon Hunter, Daniel W. Low, Erik N. Cagle, Marco Carnevali, et al. Design and chemical synthesis of a homogeneous polymer-modified erythropoiesis protein. Science, 299(5608):884–887, 2003. doi: 10.1126/science.1079085.

[25] Ly Minh Nguyen, Calvin J. Meaney, Gauri G. Rao, Mandip Panesar, and Wojciech Krzyzanski. Population pharmacodynamic modeling of epoetin alfa in end-stage renal disease patients receiving maintenance treatment using Bayesian approach. CPT: Pharmacometrics & Systems Pharmacology, 9:596–605, 2020. doi: 10.1002/ psp4.12556.

[26] ScienceDirect Corpus Source. Erythropoiesis stimulating agent recommendation model using recurrent neural networks for patient with kidney failure with replacement therapy. Computers in Biology and Medicine, 137: 104718, 2021. doi: 10.1016/j.compbiomed.2021.104718.

[27] Lei Wang, Xudong Li, Han Zhang, Jinyi Wang, Dingkang Jiang, Zhidong Xue, and Yan Wang. A comprehensive review of protein language models. arXiv preprint arXiv:2502.06881, 2025.

[28] Xuechun Zhang, Xiaoxuan Hu, Tongtong Zhang, Ling Yang, Chunhong Liu, Ning Xu, Haoyi Wang, and Wen Sun. Predicting protein solubility by benchmarking multiple protein language models. Briefings in Bioinformatics, 25(5):bbae404, 2024. doi: 10.1093/bib/bbae404.

[29] Suresh Pokharel. Protein Language Models-Based Representation for Post-translational Modification Predic-tion. PhD thesis, Rochester Institute of Technology, 2025.

[30] Can Chen, Jingbo Zhou, Fan Wang, Xue Liu, and Dejing Dou. Structure-aware protein self-supervised learning. Bioinformatics, 39(4):btad189, 2023. doi: 10.1093/bioinformatics/btad189.

[31] Jia-Ying Chen et al. Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. Frontiers in Bioengineering and Biotechnology, 2025. doi: 10. 3389/fbioe.2025.1506508.

[32] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 373(6557):871–876, 2021. doi: 10.1126/science. abj8754.

[33] Jennifer Fleming, Paulyna Magana, Sreenath Nair, Maxim Tsenkov, Damian Bertoni, Ivanna Pidruchna, Marcelo Querino Lima Afonso, Adam Midlik, Urmila Paramval, Augustin Žídek, et al. AlphaFold protein structure database and 3d-beacons: New data and capabilities. Journal of Molecular Biology, 2025. doi: 10.1016/j.jmb.2025.168967.

[34] 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, 2016. doi: 10.1109/ JPROC.2015.2494218.

[35] Rodolphe Le Riche and Victor Picheny. Revisiting Bayesian optimization in the light of the COCO benchmark. Structural and Multidisciplinary Optimization, 64:3063–3087, 2021. doi: 10.1007/s00158-021-02977-1.

[36] Koray Açıcı. Hemoglobin value prediction with Bayesian optimization assisted machine learning mod-els. Communications Faculty of Sciences University of Ankara Series A2-A3, 66(2):176–200, 2024. doi: 10.33769/aupse.1462331.

[37] Anonymous. Drug delivery optimization through Bayesian networks. Provided reference corpus (unattributed scanned source), 1998. Bibliographic metadata incomplete in source file; cited for conceptual application of Bayesian networks to dosage/delivery optimization.