frontiers

Author's proof

Before checking your proof, **please read the instructions below**

- Carefully read the entire proof and mark all corrections in the appropriate place, using the commenting tools (**Adobe Help**). Do not use the Edit tool, as direct edits could be missed (the PDF was blocked for editing to prevent this); annotate your corrections instead.
- Provide your corrections in a single PDF file or post your comments in the Production Forum, making sure to reference the relevant query/line number. Upload or post all your corrections directly in the Production Forum to avoid any comments being missed.
- We do not accept corrections via email or in the form of edited manuscripts.
- Do not provide scanned or handwritten corrections.
- Before you submit your corrections, make sure that you have checked your proof carefully, as once you approve it you won't be able to make any further corrections.
- To ensure timely publication of your article, please submit your corrections within 48 hours. We will inform you if we need anything else; do not contact us to confirm receipt.
- Note that the column alignment at the bottom of each page is not ensured during this Author's Proof stage. The columns will be correctly aligned in the final PDF publication. You may therefore notice small differences in the structure of the Author's Proof PDF versus the final publication.

Do you need help? Visit our **Production Help Center** for more information. If you can't find an answer to your question, contact your Production team directly by posting in the Production Forum.

NOTE FOR CHINESE-SPEAKING AUTHORS: If you'd like to see a Chinese translation, click on the symbol next to each query. **Only respond in English** as non-English responses will not be considered. Translated instructions for providing corrections can be found **here**.

Quick checklist

- ☐ **Author names** - Complete, accurate and consistent with your previous publications.
- ☐ **Affiliations** - Complete and accurate. Follow this style when applicable: Department, Institute, University, City, Country.
- ☐ **Tables** - Make sure the meaning/alignment of your Tables is correct with the applied formatting style.
- ☐ **Figures** - Make sure we are using the latest versions.
- ☐ **Funding and Acknowledgments** - List all relevant funders and acknowledgments.
- ☐ **Conflict of interest** - Ensure any relevant conflicts are declared.
- ☐ **Supplementary files** - Ensure the latest files are published and that no line numbers and tracked changes are visible. Also, the supplementary files should be cited in the article body text.
- ☐ **Queries** - You must reply to all **typesetter's queries below** in order for production to proceed.
- ☐ **Content** - Read all content carefully and ensure any necessary corrections are made, then **upload them** to the Production Forum.

Author queries form

Query no.	Details required	Author's response
Q1	Confirm that the article title is correct and check that it makes sense. Note that titles/headings are formatted according to Frontiers' style.	
Q2	Provide a URL for the LOOP profile for the following authors if they wish this to be linked to the final published version. If they are not yet registered, ensure that they register with Frontiers at the provided link. Gary R. Abel Jacob Beal Samuel Curtis Leonard Foner Samuel P. Forry Corey M. Hudson Caitlin Jagla Rassin Lababidi Sheng Lin-Gibson Sebastian Rivera David J. Ross Bruce J. Wittmann	

Query no.	Details required	Author's response
	If a URL is not provided, the profile link will not be added to the article. Non-registered authors and authors with profiles set to "Private" will have the default profile image displayed. Note that we will not be able to add profile links after publication.	
Q3	The citation and surnames of all authors have been highlighted. Check that they are correct and consistent with your previous publications, and correct them if needed, noting that the format in the author list should be [First name] [Surname]. Please note that this may affect the indexing of your article in repositories such as PubMed.	
Q4	There is a discrepancy between the styling of the author names in the submission system and the manuscript. We have used [Brittany Rife Magalis] instead of [Brittany Magalis]. Please confirm that it is correct.	
Q5	Provide the department name for Affiliation [1, 2, 3, 4, 6, 8, 9, 12, 13, 14, 15] (if applicable).	
Q6	Confirm that all author affiliations are correctly listed. Per our style guidelines, affiliations are listed sequentially and follow author order. Requests for non-sequential affiliation listing or to add street addresses/postcodes will not be fulfilled. Note that affiliations should reflect those at the time during which the work was undertaken. If adding new affiliations, specify if these should be listed as a present address instead of a regular affiliation.	
Q7	Ensure you provide an active email address in the correspondence section. Confirm that the email address is correct and free of typos. Any changes to corresponding authors require individual confirmation from all original and added/removed corresponding authors. Please note, Authorship Change Forms are not required for amendments to the correspondence section.	
Q8	Confirm that the keywords are correct, and keep them to a maximum of eight and a minimum of five. (Note: a keyword can be made up of one or more words.)	
Q9	Check if the section headers (i.e., section leveling) have been correctly captured.	
Q10	Confirm that the Data Availability statement is accurate. Note that this statement may have been amended to adhere to our Publication Ethics guidelines.	
Q11	Confirm that the details in the "Author Contributions" section are correct. If any contributions need to be added/edited, choose the appropriate CRediT roles from the list available here and indicate which one(s) apply. Please be aware that writing roles ("Writing – original draft" and/or "Writing – review & editing") are a requirement for authorship.	
Q12	Check all grant numbers and funding information in the proof corresponds to your funding application. All funders should be credited, and all grant numbers should be correctly included in this section. Note that if you add any commercial funding, please ensure that the funders involvement/non-involvement in the manuscript is declared. If you provided a positive funding statement but don't provide funding details, then the statement will be updated to say no funding was received.	
Q13	A commercial affiliation is listed in your article and mentioned in the Conflict of Interest statement as per the guidelines. Check that the amendments to the Conflict of Interest are fine.	
Q14	Frontiers guidelines require listing the first 6 authors + et al. for articles with more than 6 authors. Please provide the names of the other jacak et al., 2021 authors.	
Q15	Provide the accessed date for the url (accessed [Month D, YYYY]) for reference Aclid, 2025, Carter and Butchello, 2026, CDC and USDA, 2025, IGSC, 2024, Jacak et al., 2021, The White House, 2025, UK DSIT, 2024, US NSTC, 2024, Vought and Kratsios, 2025, Williams et al., 2025, Wyschogrod et al., 2022.	
Q16	Provide the volume number and page range for the following references, if applicable. "Airas and Zhang, 2026; Beal and Bryan, 2024; Challacombe and Haas, 2024; Jacak et al., 2021; Mackelprang et al., 2025; UK DSIT, 2024; US NSTC, 2024; Wyschogrod et al., 2022."	

Query no.	Details required	Author's response
Q17	Provide the volume number for the following references, if applicable. "Baum et al., 2026; Brixi et al., 2025; Passaro et al., 2025; Ruffolo et al., 2024; Tayouri et al., 2025; Wittmann et al., 2026."	
Q18	Provide a working DOI for "Jacak et al., 2021, UK DSIT, 2024, US NSTC, 2024, Wyschogrod et al., 2022", if applicable. Invalid DOIs will not be added to references. References without DOIs will still be included.	
Q19	Ensure that all the figures, tables, and captions are correct, and that all figures are of the highest quality/resolution. You may upload improved figures to the Production Forum. If so, please describe in visual terms the exact changes(s) made to help us confirm that the updated version has been used in the finalized proof. Please note that figures and tables must be cited sequentially, per the author guidelines.	
Q20	The image used in Figure 2 does not have any part labels; however, "Figures 2A–D" are cited in the article. Provide revised figure files containing these part labels or remove the part labels from the in-text citations.	
Q21	If you decide to use previously published and/or copyrighted figures in your article, please keep in mind that it is your responsibility as the author to obtain the appropriate permissions and licenses to reproduce them, and to follow any citation instructions requested by third-party rights holders. We require attribution for any elements where you are not the original copyright holder. If obtaining the reproduction rights involves the payment of a fee, these charges are to be paid by the authors. Please provide us with the following attribution details in each case: [Citation], [License], and [Source], and confirm whether you have reproduced or adapted the material. Third-party material includes any proprietary text, illustrations, figures, tables, images, artworks, datasets, databases, as well as trademarked logos or brand names, and any material taken from websites, such as: screenshots, photographs, illustrations, maps, icons, clipart, tables, or social media posts*. *Note: Social media posts require permission from the original user/poster before they can be reproduced.	
	Please confirm that the below mandatory Frontiers AI generated Alt-Text is an accurate visual description of your Figure(s). As part of our open science commitment, Alt-Text will enhance the accessibility of your manuscript. It is a short visual description that allows people using screen reading technology to clearly understand the contents of an image. These Figure Alt-text proposals won't replace your figure captions and will not be visible on your article. If you wish to make any changes, kindly provide the exact revised Alt-Text you would like to use, ensuring that the word-count remains at approximately 100 words for best accessibility results. **Further information on Alt-Text can be found here.** **Figure 1 Alt-Text** – Flowchart illustrating a screening pipeline using molecular representations, with three main sections: "Molecular representations" converts nucleic acid and protein sequences into embedded latent vectors; "Sequence-based screening" detects sequence variants based on similarity and features; "Function-based screening" predicts structural, binding, functional, and other properties, and detects variants based on predicted molecular properties. Output includes flag or clear, risk score, confidence, and reports. **Figure 2 Alt-Text** – Scientific illustration compares protein sequence clustering and detection methods. Left panel shows labeled protein families and a highlighted hazardous family. Right panels contrast sequence-based detection, which misses distant variants, with function-based detection that identifies more variants in a projected representation space.	

frontiers
in Bioengineering and
Biotechnology

TYPE Perspective
PUBLISHED XX XX 2026
DOI XXXX

OPEN ACCESS

EDITED BY
Clara Rubinstein,
University of Buenos Aires, Argentina

REVIEWED BY
Ranjit Ranbhor,
Odin Pharmaceuticals LLC, United States

*CORRESPONDENCE
Gary R. Abel,
gary@fourtheon.bio

RECEIVED 17 March 2026
REVISED 17 March 2026
ACCEPTED 07 April 2026
PUBLISHED XX XX 2026

CITATION
Abel GR, Alexanian T, Bartling C, Beal J, Curtis S, Flyangolts K, Foner L, Forry SP, Godbold GD, Horvitz E, Hu B, Hudson CM, Jagla C, Lababidi R, Lin-Gibson S, Magalis BR, Pannu J, Rivera S, Ross DJ, Wittmann BJ and Diggans J (2026) Beyond sequence similarity: toward function-based screening of nucleic acid synthesis.
Front. Bioeng. Biotechnol. 14:1832724.
doi: XXXX

Beyond sequence similarity: toward function-based screening of nucleic acid synthesis

Gary R. Abel[1,2*], Tessa Alexanian[3], Craig Bartling[4], Jacob Beal[5], Samuel Curtis[6], Kevin Flyangolts[7], Leonard Foner[8], Samuel P. Forry[9], Gene D. Godbold[10], Eric Horvitz[11], Bin Hu[12], Corey M. Hudson[13], Caitlin Jagla[5], Rassin Lababidi[3], Sheng Lin-Gibson[9], Brittany Rife Magalis[14], Jaspreet Pannu[2], Sebastian Rivera[15], David J. Ross[9], Bruce J. Wittmann[11] and James Diggans[16]

[1]Fourth Eon Bio, San Diego, CA, United States, [2]Johns Hopkins University Center for Health Security, Baltimore, MD, United States, [3]International Biosecurity and Biosafety Initiative for Science, Geneva, Switzerland, [4]Battelle Memorial Institute, Columbus, OH, United States, [5]RTX BBN Technologies, Cambridge, MA, United States, [6]Center for AI Standards and Innovation, Washington, DC, United States, [7]Aclid, New York, NY, United States, [8]SecureDNA, Basel, Switzerland, [9]National Institute of Standards and Technology, Gaithersburg, MD, United States, [10]Signature Science LLC, Charlottesville, VA, United States, [11]Microsoft, Office of the Chief Scientific Officer, Redmond, WA, United States, [12]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, United States, [13]The Align Foundation, Covina, CA, United States, [14]University of Louisville, Louisville, KY, United States, [15]Engineering Biology Research Consortium, Emeryville, CA, United States, [16]Twist Bioscience, South San Francisco, CA, United States

Synthetic nucleic acids are a key input to modern biotechnology, yet they represent dual-use materials that require robust screening to mitigate biosecurity risks. The prevailing screening paradigm, which identifies sequences of concern (SoCs) through sequence similarity to controlled pathogens and toxins, may not fully capture risks posed by AI tools that can decouple biomolecular function from reliance on known sequences. Rapidly advancing biodesign capabilities enable the generation of genes and proteins that might evade sequence-based detection. We highlight the critical need for function-based screening approaches that can detect sequences capable of hazardous biological functions, regardless of similarity to known SoCs. We examine the feasibility of function-based screening with an initial focus on proteins, arguing that, while protein sequence space is vast, biologically functional proteins are significantly constrained by biophysical and biochemical requirements that can be learned and modeled. We propose a concrete implementation framework organized along a continuum of complexity, starting with toxins as the most tractable targets before expanding to more complex pathogenic functions. We then discuss open challenges and describe a research and development strategy to address them.

KEYWORDS
biological foundation models, biosecurity, DNA synthesis screening, function-based screening, protein function prediction

Introduction

Q9 Synthetic nucleic acids are a key input for a wide range of applications in biomedicine, biotechnology, and synthetic biology. Because synthetic biology research is fundamentally dual-use, nucleic acid synthesis screening serves as a critical biosecurity safeguard against both accidents and deliberate misuse (Mackelprang et al., 2025). Screening systems are meant to operate as rule-out tests, seeking to confirm with high confidence that an ordered sequence does not pose a biosecurity risk. The prevailing screening paradigm approximates this by comparing sequence similarity against known sequences of concern (SoCs), which are defined primarily by taxonomic origin of the source organism. If an ordered sequence is a "Best Match" to a known SoC when compared to a comprehensive database, the sequence is flagged; if not, it can be cleared (IGSC, 2024; UK DSIT, 2024; US NSTC, 2024). In practice they operate as limited rule-out tests for known SoCs and close variants. Customer screening serves as an independent safeguard against deliberate misuse, but is not a substitute for sequence screening.

While sequence-based screening has served as a foundation of biosecurity, its limitations are becoming increasingly relevant as both artificial intelligence (AI) and synthetic biology continue to advance. Challenges include unnecessary flagging of benign sequences from regulated organisms (Godbold et al., 2025b) and, more critically, failure to detect hazardous sequences that lack significant similarity to known SoCs, whether from unregulated organisms (Gemler et al., 2024; Williams et al., 2025), extensively modified or *de novo* designed proteins (Baker and Church, 2024; Wittmann et al., 2025; Hunter, 2024), or deliberate sequence obfuscation (Rose et al., 2024).

This is not a hypothetical future concern: AI-enabled protein design tools can already generate functional protein sequences that diverge substantially from natural sequences (Munsamy et al., 2024; Ruffolo et al., 2024; Sumida et al., 2024; Yeh et al., 2023). A redesigned toxin that binds the same cellular target as a natural toxin may have low sequence identity with any previously characterized protein. To current screening systems, such a sequence may appear novel[1] and unremarkable (Wittmann et al., 2025), eroding the efficacy of sequence-based screening.

There is growing recognition that synthesis screening must move beyond definitions of SoCs based on taxonomic origin or sequence similarity alone, toward detecting sequences that encode hazardous biological functions (Mackelprang et al., 2025; Godbold et al., 2025b). Recent policy guidance from the United States (The White House, 2025; Vought and Kratsios, 2025), United Kingdom (UK DSIT, 2024), and European Union (EU DG SANTE, 2025) further reinforce this necessity.

We focus on two objectives. First, we examine the theoretical feasibility of function-based screening, arguing that fundamental biophysical and biochemical requirements constrain functional proteins and lead to learnable patterns that enable prediction of biomolecular properties from sequence.[2] Second, we propose a concrete, near-term implementation framework that can begin to provide function-based screening capabilities while broader, more generalized predictive methods continue to mature. We contend that these two objectives represent points along a single developmental continuum, from targeted detection of specific known hazards toward increasingly general prediction of hazardous functions. Work on the nearer-term approach lays scientific and institutional foundations for the longer-term vision.

Sequence-based and function-based screening are complementary

We define sequence-based screening as detection based on significant sequence similarity to a known SoC, i.e., a regulated gene or protein sequence. Current sequence-based screening methods employ techniques such as sequence alignment (Altschul et al., 1990), exact matching of cryptographically hashed k-mers (Baum et al., 2026), Hidden Markov Models (HMM) (Finn et al., 2011), k-mer signatures (Wyschogrod et al., 2022), and combinations thereof (Laird et al., 2025; Balaji et al., 2022; Gemler et al., 2023; Wheeler, Carter, et al., 2024).

In contrast, we define function-based screening as detection of sequences whose predicted molecular properties indicate a capacity for biological *functions of concern*, i.e., functions that contribute substantially to host toxicity or pathogenesis. Existing capabilities can be leveraged to implement function-based screening (Figure 1), including functional annotation (Lin et al., 2024), structure prediction (Meng et al., 2025) and search (van Kempen et al., 2024), binding prediction (Passaro et al., 2025), functional signature detection (Beal and Bryan, 2024), embedding space search (Pantolini et al., 2024), and prediction of functional variants to proactively expand sequence databases (Baum et al., 2026).

These terms are imperfect, as sequence-based methods implicitly capture some functional information, while function-based methods typically take sequences as input. Importantly, sequence-based and function-based screening need not be mutually exclusive. Indeed, some existing screening tools already incorporate elements of function prediction (Aclid, 2025; Balaji et al., 2022; Baum et al., 2026; Beal and Bryan, 2024; Gemler et al., 2023), and many methods fall along a spectrum between the two. The most effective screening approaches will likely combine elements from both paradigms by using hybrid approaches that integrate new methods into existing screening pipelines, enabling a smooth transition as models mature.

1 The term 'novel' is often used loosely, but novelty is not a single axis: a novel protein may diverge from a known threat in sequence while conserving structure, or diverge in structure while conserving mechanism, or diverge in mechanism while targeting the same host pathway. Different screening methods address different axes of divergence.

2 The approaches outlined here focus on protein-coding sequences as the most tractable targets for function prediction. Other biopolymers such as functional RNAs and prion-forming proteins are important, but present distinct challenges and are therefore excluded from this discussion.

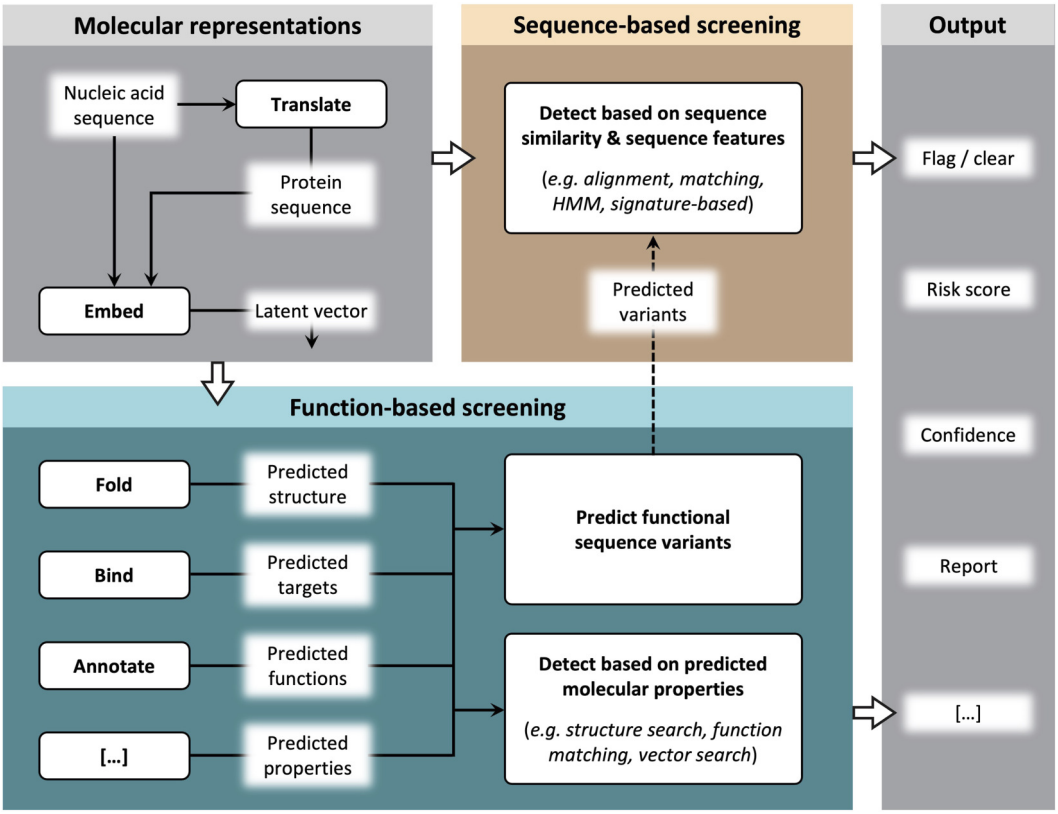

FIGURE 1
Conceptual overview of sequence-based and function-based screening elements. The molecular representations used as inputs to screening could include DNA and RNA sequences, translated protein sequences, and latent-space vectors from model embedding. Sequence-based screening detects sequences of concern through sequence similarity and sequence features (e.g., alignment, matching, HMM, or signature-based methods). Function-based screening utilizes computational methods to predict molecular properties such as structure, binding targets, and functional annotations, and detects functions of concern based on those predicted properties. New functional sequence variants can also be predicted and used to expand databases for sequence-based screening (dashed line). Both paradigms can produce a range of screening outputs that inform synthesis decisions. Note that elements shown are illustrative, not exhaustive.

Constraints on functional proteins enable prediction

The sequence space of all possible proteins is vast. For a protein of even a modest length of 75 residues, the number of possible sequences is far greater than the estimated number of atoms in the observable universe (Zhang et al., 2025). Yet the subregions corresponding to *biologically functional* proteins are constrained, with a much smaller subset representing hazardous functions (Figure 2). These constraints reflect fundamental requirements that a protein must satisfy to function successfully within one or more biological contexts, and further, to contribute to pathogenicity or toxicity. Biophysical constraints govern whether a sequence can fold into a stable, functional conformation or adopt a functional disordered ensemble, while biochemical constraints further restrict viable sequences, as specific molecular functions require precise spatial arrangement of catalytic residues and geometric complementarity at binding interfaces. These constraints have measurable consequences, and even a simple binding function may have fewer than one in 10^{11} functional sequences (Keefe and Szostak, 2001). The functional regions of sequence space are thus *many* orders of magnitude smaller and, crucially, are structured by a common set of biophysical and biochemical principles. The same constraints that make functional sequences rare also make them predictable.

The constraints on functional proteins create statistical regularities in how sequence maps to structure and function, which can be learned empirically by AI models and exploited to detect specific functions of concern. In particular, biological foundation models have demonstrated a capacity for extracting such patterns and using them to predict protein properties (Bjerregaard et al., 2025; Li et al., 2024). There is growing evidence that foundation models can implicitly capture abstract representations of the underlying constraints on proteins (Ahdritz et al., 2024; Hayes et al., 2025; Brixi et al., 2025; Airas and Zhang, 2026), suggesting they might be able to recognize functional patterns in non-natural sequences through their latent space geometry. In the long term, such generalizable property prediction could unlock a more resilient approach to function-based screening, for example, by integrating across scales and modalities as an AI virtual cell (Bunne et al., 2024) to enable prediction of whether a novel protein would disrupt critical host cell processes.

It is not clear how far in the future generalized screening can be achieved, as the extent of such model capability generalization

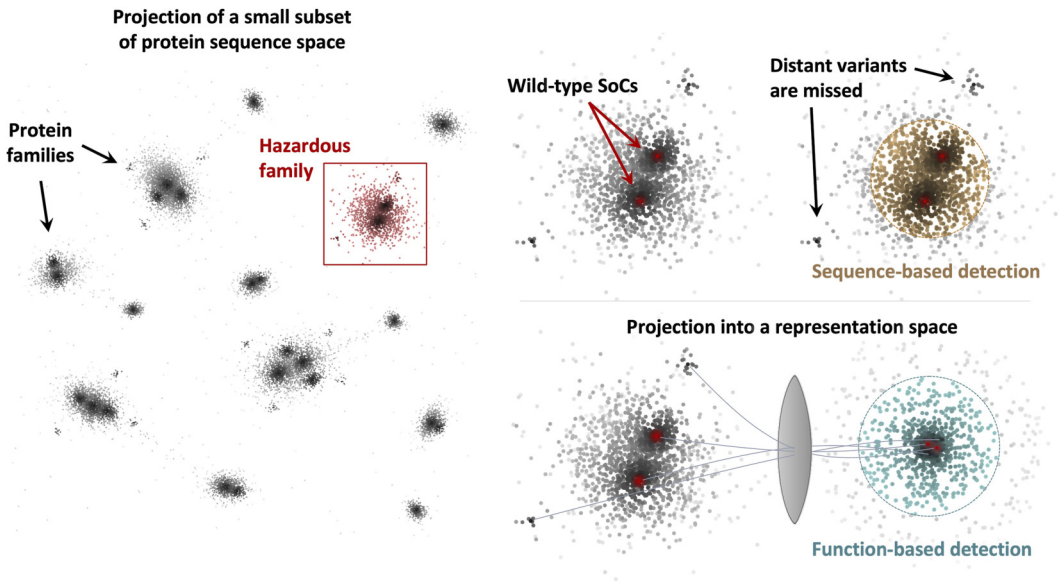

FIGURE 2
Illustrative representation of a projection of a small subset of protein sequence space onto a two-dimensional grid. Each point represents a possible protein, with the point intensity corresponding to its activity for a particular function. **(A)** Several isolated clusters corresponding to individual protein functional families are shown, with a specific hazardous family highlighted. **(B)** A close-up view of the hazardous cluster from **(A)**, with two wild-type reference proteins shown as red dots surrounded by mutants of varying activity. **(C)** Current Sequence-Based Screening flags a subset of adjacent proteins, but misses any functional proteins that are sufficiently far away in sequence space (as indicated by arrows). **(D)** Function-based screening aims to project proteins into a different representation space where they cluster based on similarity in structure, function, and other molecular properties, to enable detection of functional analogs that may be distant in sequence space.

remains an open empirical question. In the near term, however, the goal is more targeted: to identify sequences whose properties indicate specific harmful functions. This is a narrower target that can be approached incrementally, starting from targeted detection of specific known hazards and progressing towards broader function prediction as models mature.

From theoretical feasibility to practical implementation

Function-based screening is both necessary and theoretically feasible. The practical challenge is to move toward operational deployment with limited data and imperfect models.

A useful starting point is to observe that synthesis screening does not require comprehensive sequence-to-function prediction. Instead, it only needs to prevent the acquisition of sequences that could do harm in the hands of a malicious or careless actor. Comprehensive function prediction asks "What does this protein do?"—a classification problem across a vast and poorly defined label space (Bileschi et al., 2022). Screening asks a narrower question: "Can we confidently exclude that this protein performs specific harmful function X?" where X is drawn from a known set of harmful functions. This is a binary exclusion problem for a set of narrow, well-defined targets.

This framing suggests a pragmatic near-term approach: before developing generalist models for broad function prediction, specialist models can be trained to detect one (or a few) specific functions of concern. For example, a specialist model for detecting N-glycosidase ribosome-inactivating toxins asks only "Could this sequence encode a protein capable of depurinating ribosomal RNA?" It needs only to output whether the sequence can be confidently excluded from performing the target function, whether it likely encodes the target function, or whether there is insufficient confidence to rule it out, with the latter two cases triggering review. In the near term, this classification or rule-out decision can potentially be served by small, lightweight classifiers trained on positive examples (known sequences with the target function, plus computationally generated variants) and negative examples (diverse sequences known not to have the target function), using sequence features or model embeddings as inputs. Small, specialized models can be rigorously validated against ground truth backed by experimental data and, importantly, their failure modes can be more readily characterized and understood.

The progression from specialized to generalized models is also motivated by practical considerations, as the sensitivity-specificity tradeoff may scale poorly across a large collection of independent models. An intermediate approach could use an ensemble of models that predict different molecular properties, producing a profile that can identify harmful functions. The resulting signal serves as an indicator of biosecurity risk that must be integrated into existing screening workflows where flagged sequences require expert review, making it critical to minimize false positive rates while maintaining high sensitivity.

Tractable targets should be prioritized first

Given that "function" is a broad and ill-defined concept, we propose approaching function-based screening by prioritizing a few narrow, well-defined and highly tractable functional categories, and bootstrapping into a more generalized screening paradigm. Protein cytotoxins represent the clearest starting point: the relationship between structure and function is comparatively well-understood, decades of toxicology research provide structure-activity relationships and characterized variants, the mechanistic space is bounded, and detection aligns with current regulation of controlled toxins (CDC & USDA, 2025). Viral entry proteins, particularly receptor-binding proteins for pandemic-capable viruses, would be a natural next step. Work here can leverage advances in structure and binding affinity prediction.

More complex and context-dependent functions, such as innate immune subverting sequences and elements of fungal and protozoan pathogenesis, should be deferred due to greater challenges in data availability, context dependence, identification of host-exploiting functions, and ontological definition (Godbold et al., 2022; Godbold and Scholz, 2024). Starting with the most tractable targets and demonstrating operational feasibility builds the methodology, data pipelines, validation methods, and institutional capacity needed to expand towards more generalized function-based screening.

Moving from concept through development to deployment

Developing function-based screening models for reliably detecting functions of concern requires several types of data: (a) positive examples, including experimentally measured natural sequences or computationally generated variants that encode the target function; (b) negative examples, including diverse sequences from organisms without the target function; and (c) held-out validation sets drawn from different taxonomic groups with little sequence or structural similarity and including experimentally validated synthetic sequences. Generating adequate high-quality training data is a significant undertaking, likely requiring several iterative rounds of variant design, data curation, model building, and validation.

It is important to acknowledge that data and modeling relating to hazardous functions are inherently sensitive, and that pursuit of such work outside of appropriately secured institutions could itself pose biosecurity risks. However, not all targets present equal sensitivity concerns. Initial development efforts should prioritize well-characterized functions of concern (e.g., well-known protein toxins) whose sequences, structures, and functional properties are already extensively documented in the open literature. For these targets, the marginal information hazard from generating additional functional variants is minimal, as the underlying biology is already widely accessible. Beginning with such targets (and employing benign proxies when possible) allows the research community to validate the full model-development pipeline while producing models with immediate defensive value. The focus should be on collecting data that accelerates defensive capabilities without generating new functional insights beyond what is necessary to advance screening.

As development progresses to less-characterized, less-public, or higher-risk functions of concern, training data becomes increasingly sensitive, as detailed information about which sequence modifications preserve toxic or pathogenic functions could itself pose a biosecurity risk. Therefore, sensitive data and trained models should be carefully controlled and distributed only through a tiered access framework (Bloomfield et al., 2026; Carter and Butchello, 2026; Wittmann et al., 2025), wherein model developers securely access the data, trusted providers and screening tool developers receive model weights for deployment, and others access screening via software-as-a-service to limit data and weight proliferation. A provider deploying a toxin-detection model can screen incoming orders without ever seeing the specific variants in the training set, keeping the information hazard contained. This framework should be implemented early on while the stakes are lower, in a graduated approach that allows operational security measures to mature and scale with the actual information hazard.

This motivates an ecosystem organized around complementary roles, in which no single entity needs access to all sensitive components. Secure research institutions generate training data, train and validate models, and conduct red-team evaluations. Trusted national and international bodies manage controlled access to models and sensitive test sets. Screening tool developers integrate validated models into production screening software, while synthesis providers deploy them and report anonymized hit patterns. And government agencies provide coordination, oversight, and threat-informed prioritization. The ecosystem should support continual improvement through operational feedback. Hit pattern reporting, expert review of flagged sequences, and emerging threat intelligence can drive rapid retraining and redeployment of individual models, creating a defense posture whose decision boundaries shift as models are updated, making them difficult to evade (Wyschogrod et al., 2022).

Open challenges and research priorities

We believe that the approach described above is achievable with current methods and institutional capacity. Several challenges, however, will shape how quickly and successfully function-based screening can be implemented.

First, operationalizing function-based screening at any level requires clear rules for determining which biological functions warrant flagging. Several biosecurity-relevant annotation frameworks have been developed, including the Virulence Factor Database (VFDB) (Liu et al., 2022), the Pathogen–Host Interactions database (PHI-base) (Urban et al., 2022), the Functional Hazards Database (Gemler et al., 2022), Functions of Sequences of Concern (FunSoCs) (Godbold et al., 2022), PathGO (Jacak et al., 2021), and a recent formal extension of the Gene Ontology framework to pathogenic biological process terms (Godbold et al., 2025a). A key priority is to define consensus rules for determining which individual functions or combinations pose sufficient risk to warrant flagging during screening. In this regard, the recently established Sequence Biosecurity Risk Consortium (SBRC) is well positioned to

develop function-based screening rubrics through careful and systematic assessment of biosecurity risk from different functions by subject matter experts (Beal and Alexanian, 2025).

Second, significant gaps remain in our understanding of where and how biological AI model predictions fail, due in part to a lack of tools and datasets to systematically evaluate their performance out of distribution. It will be crucial to develop evaluation methods and benchmarks that assess prediction accuracy for functional properties (Notin et al., 2023) for both natural and non-natural sequences across a range of protein types (Challacombe and Haas, 2024). Uncertainty quantification deserves particular attention: For any prediction used in screening, it will be important to estimate confidence, as this directly influences interpretation and determines the sensitivity and cost tradeoffs between false negatives and false positives. Robustness under adversarial conditions must also be systematically tested using structured red-teaming exercises, including whether models can maintain performance when sequences are deliberately designed to evade detection or significantly deviate from model training data (Batalis et al., 2024; Wittmann et al., 2025). Short sequence fragments carry less information and thus pose a notable challenge, as do multi-element constructs that combine coding sequences with regulatory and translational components. These challenges motivate use of mitigations such as analyzing order pools to predict plausible assembly products (Tayouri et al., 2025; Wittmann et al., 2026).

Third, there is a need to expand data collection efforts to enable training of screening-relevant models. Existing experimental data on protein function is overwhelmingly from naturally evolved sequences or close mutants, biased by common measurement techniques and functions that are easily measured in high throughput (Schnoes et al., 2013), and concentrated on a small number of model organisms (Kustatscher et al., 2022; Rocha et al., 2023). How much of functional protein space has been explored remains unclear, given evolution's reliance on incremental sampling through mutation under selection (Hoarfrost et al., 2022; Mahlich et al., 2023). This potentially leaves vast regions uncharacterized, and hinders model training and validation. Continued progress will require scaling up experimental data collection across a wide range of protein functions (Cortade et al., 2024), with particular emphasis on characterizing more distant regions of sequence space that have not been explored by nature but are becoming accessible to biodesign. Investigating pathogenic functions poses additional challenges, as the complexity of pathogen-host interactions limits the utility of reductionist approaches (Moxon and Tang, 2000), while testing modified pathogens can pose biosecurity risks. These risks should be mitigated by using non-replicating or non-infectious models (Godbold et al., 2023) and, when possible, safe proxy functions (Ikonomova et al., 2025). Finally, functional assays have been notoriously difficult to standardize across laboratories (Hirsch and Schildknecht, 2019), and ongoing standardization efforts (Sansone et al., 2019) will be essential for reliable model development.

Conclusion

Function-based screening provides key advantages over the current sequence-based paradigm, and is both theoretically feasible and practically achievable. Near-term priorities include defining function-based screening criteria for an initial set of targets, collecting training data on natural and computationally predicted variants, developing detection-focused models within appropriately secured research institutions, integrating function-based methods into screening tools, and piloting deployment with synthesis providers. In parallel, continued work on ontological frameworks, evaluation benchmarks, and experimental data collection across functional space will lay the groundwork for broader function-based screening. Progress along this continuum need not wait for any single challenge to be fully resolved; the near-term strategy builds the data, methods, and institutional capacity that more ambitious approaches will require. As the threat landscape evolves, so must the defenses. Screening alone cannot address all biosecurity risks, but it remains one of the most scalable, tractable, and effective points of intervention. Advancing function-based screening will require coordination among research institutions, synthesis providers, screening tool developers, and government agencies; targeted investment in training data and evaluation infrastructure; and sustained momentum from development through deployment. Together, these efforts can materialize a defensive posture that anticipates the threat landscape rather than perpetually reacts to it.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

GA: Writing – original draft, Visualization, Project administration, Conceptualization, Writing – review and editing. TA: Visualization, Writing – review and editing. CB: Writing – review and editing. JB: Writing – review and editing. SC: Writing – review and editing. KF: Writing – review and editing. LF: Writing – review and editing. SF: Writing – review and editing. GG: Writing – review and editing. EH: Writing – review and editing. BH: Writing – review and editing. CH: Writing – review and editing, Visualization. CJ: Writing – review and editing. RL: Writing – review and editing. SL-G: Writing – review and editing. BM: Writing – review and editing. JP: Writing – review and editing. SR: Writing – review and editing. DR: Writing – review and editing. BW: Writing – review and editing. JD: Writing – review and editing, Conceptualization, Writing – original draft.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work was supported by funding from Sentinel Bio. BH acknowledges the Center for National Security and International Studies at Los Alamos National Laboratory for its support of this work.

Acknowledgements

The authors thank Janika Schmitt for early feedback on the concept, Joshua Monrad and Hanna Pálya for feedback on the manuscript draft, Jim Gibson for assistance with drafting early figure versions, and Ian Beatty and Svetlana Ikonomova for feedback on the figures.

Conflict of interest

Q13 Authors JB, CJ were employed by RTX BBN Technologies.

Author KF was employed by Aclid.

Author GG was employed by Signature Science LLC.

Authors EH, BW were employed by Microsoft.

Author JD was employed by Twist Bioscience.

The remaining author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was used in the creation of this manuscript. The authors acknowledge use of generative AI tools to assist with background literature exploration, editorial suggestions, and drafting code used to generate the plots in Figure 2. Models used include Claude Opus 4.5/4.6 (Anthropic), Gemini 3 (Google), and GPT-5.2 (OpenAI). All outputs were carefully reviewed, verified, and edited by the authors, who take full responsibility for the content of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aclid (2025). Aclid sequence screening. Available online at: https://www.aclid.bio/product/sequence-screening.

Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., et al. (2024). OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat. Methods* 21 (8), 1514–1524. doi:10.1038/s41592-024-02272-z

Airas, J., and Zhang, B. (2026). Knowledge distillation of a protein language model yields a foundational implicit solvent model. *arXiv:2601.05388*. doi:10.48550/arXiv.2601.05388

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi:10.1016/S0022-2836(05)80360-2

Baker, D., and Church, G. (2024). Protein design meets biosecurity. *Science* 383 (6681), 349. doi:10.1126/science.ado1671

Balaji, A., Kille, B., Kappell, A. D., Godbold, G. D., Diep, M., Elworth, R. A. L., et al. (2022). SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning. *Genome Biol.* 23 (1), 133. doi:10.1186/s13059-022-02695-x

Batalis, S., Schuerger, C., Gronvall, G. K., and Walsh, M. E. (2024). Safeguarding mail-order DNA synthesis in the age of artificial intelligence. *Appl. Biosaf.* 29 (2), 79–84. doi:10.1089/apb.2023.0020

Baum, C., Berlips, J., Chen, W., Cozzarini, H., Cui, H., Damgård, I., et al. (2026). A system capable of verifiably and privately screening global DNA synthesis. *Natl. Sci. Rev.*, nwag103. doi:10.1093/nsr/nwag103

Beal, J., and Alexanian, T. (2025). Creating enforceable biosecurity standards for nucleic acid providers. *Eng. Biol.* 9 (1), e70003. doi:10.1049/enb2.70003

Beal, J., and Bryan, C. (2024). A conserved residue knowledge (CoRK) approach to developing AI-Proof function of concern signatures. *Synthesis screening workshop. Synth. Screen. Workshop*. doi:10.5281/zenodo.18751560

Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., et al. (2022). Using deep learning to annotate the protein universe. *Nat. Biotechnol.* 40 (6), 932–937. doi:10.1038/s41587-021-01179-w

Bjerregaard, A., Groth, P. M., Hauberg, S., Krogh, A., and Boomsma, W. (2025). Foundation models of protein sequences: a brief overview. *Curr. Opin. Struct. Biol.* 91, 103004. doi:10.1016/j.sbi.2025.103004

Bloomfield, D., Black, J. R. M., Crook, O., Brandes, N., Hanke, M. S., Inglesby, T. V., et al. (2026). Biological data governance in an age of AI. *Science* 391 (6785), 558–561. doi:10.1126/science.aeb2689

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., et al. (2025). Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025.02.18.638918. doi:10.1101/2025.02.18.638918

Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., et al. (2024). How to build the virtual cell with artificial intelligence: priorities and opportunities. *Cell* 187 (25), 7045–7063. doi:10.1016/j.cell.2024.11.015

Carter, S. R., and Butchello, G. (2026). A framework for managed access to biological AI tools. Available online at: https://www.nti.org/analysis/articles/a-framework-for-managed-access-to-biological-ai-tools/.

CDC and USDA (2025). Select agents and toxins list, federal select agent program. Available online at: https://www.selectagents.gov/sat/list.htm.

Challacombe, C. A., and Haas, N. S. (2024). Towards a dataset for state of the art protein toxin classification. *Synth. Biol.* doi:10.1101/2024.04.14.589430

Cortade, D., d'Oelsnitz, S., Chadha, A., Hayes, O., Taghon, G., Doerr, M., et al. (2024). Design of a generalized platform for gathering protein sequence → function datasets at scale. doi:10.5281/zenodo.13909104

EU DG SANTE (2025). Proposal for a regulation to establish measures to strengthen the Union's biotechnology and biomanufacturing sectors (european biotech act) [government report]. The european Commission's directorate-General for Health and food safety (DG SANTE). Available online at: https://health.ec.europa.eu/publications/proposal-regulation-establish-measures-strengthen-unions-biotechnology-and-biomanufacturing-sectors_en.

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39 (Suppl. l_2), W29–W37. doi:10.1093/nar/gkr367

Gemler, B. T., Mukherjee, C., Howland, C. A., Huk, D., Shank, Z., Harbo, L. J., et al. (2022). Function-based classification of hazardous biological sequences: demonstration of a new paradigm for biohazard assessments. *Front. Bioeng. Biotechnol.* 10, 979497. doi:10.3389/fbioe.2022.979497

Gemler, B. T., Mukherjee, C., Howland, C., Fullerton, P. A., Spurbeck, R. R., Catlin, L. A., et al. (2023). UltraSEQ, a universal bioinformatic platform for information-based clinical metagenomics and beyond. *Microbiol. Spectr.* 11 (3), e04160-22. doi:10.1128/spectrum.04160-22

Gemler, B. T., Mukherjee, C., Fullerton, P. A., Diggans, J., and Bartling, C. (2024). A sensitivity study for interpreting nucleic acid sequence screening regulatory and guidance documentation: toward a foundational synthetic nucleic acid sequence screening framework. *Appl. Biosaf.* 29 (3), 150–158. doi:10.1089/apb.2023.0026

Godbold, G. D., and Scholz, M. B. (2024). Annotation of functions of sequences of concern and its relevance to the new biosecurity regulatory framework in the United States. *Appl. Biosaf.* 29 (3), 142–149. doi:10.1089/apb.2023.0030

Godbold, G. D., Kappell, A. D., LeSassier, D. S., Treangen, T. J., and Ternus, K. L. (2022). Categorizing sequences of concern by function to better assess mechanisms of microbial pathogenesis. *Infect. Immun.* 90 (5), e0033421. doi:10.1128/IAI.00334-21

Godbold, G. D., Hewitt, F. C., Kappell, A. D., Scholz, M. B., Agar, S. L., Treangen, T. J., et al. (2023). Improved understanding of biorisk for research involving microbial modification using annotated sequences of concern. *Front. Bioeng. Biotechnol.* 11, 1124100. doi:10.3389/fbioe.2023.1124100

Godbold, G. D., Proescher, J., and Gaudet, P. (2025a). New and revised gene ontology biological process terms describe multiorganism interactions critical for understanding microbial pathogenesis and sequences of concern. *J. Biomed. Semant.* 16 (1), 4. doi:10.1186/s13326-025-00323-8

Godbold, G. D., Ternus, K. L., Flyangolts, K., Wheeler, N., Parker, M., Beal, J., et al. (2025b). The case for limiting "Sequences of Concern" to those with demonstrated pathogenic function. *Appl. Biosaf.* 30 (3), 206–210. doi:10.1089/apb.2025.0015

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., et al. (2025). Simulating 500 million years of evolution with a language model. *Science* 387 (6736), 850–858. doi:10.1126/science.ads0018

Hirsch, C., and Schildknecht, S. (2019). *In vitro* research reproducibility: keeping up high standards. *Front. Pharmacol.* 10, 1484. doi:10.3389/fphar.2019.01484

Hoarfrost, A., Aptekmann, A., Farfañuk, G., and Bromberg, Y. (2022). Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nat. Commun.* 13 (1), 2606. doi:10.1038/s41467-022-30070-8

Hunter, P. (2024). Security challenges by AI-assisted protein design. *EMBO Rep. Lond.* 25, 2168–2171. doi:10.1038/s44319-024-00124-7

IGSC (2024). IGSC harmonized screening protocol v3.0. Available online at: https://genesynthesisconsortium.org/wp-content/uploads/IGSC-Harmonized-Screening-Protocol-v3.0-1.pdf.

Ikonomova, S. P., Wittmann, B. J., Piorino, F., Ross, D. J., Schaffter, S. W., Vasilyeva, O., et al. (2025). Experimental evaluation of AI-Driven protein design risks using safe biological proxies. *bioRxiv*. doi:10.1101/2025.05.15.654077

Jacak, R., Godbold, G. D., Ernlund, A., et al. (2021). *PathGO: the pathogenesis gene ontology* [C++]. *JHU Appl. Phys. Lab. Biol. Sci.* Available online at: https://github.com/jhuapl-bio/pathogenesis-gene-ontology.

Keefe, A. D., and Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature* 410 (6829), 715–718. doi:10.1038/35070613

Kustatscher, G., Collins, T., Gingras, A.-C., Guo, T., Hermjakob, H., Ideker, T., et al. (2022). Understudied proteins: opportunities and challenges for functional proteomics. *Nat. Methods* 19 (7), 774–779. doi:10.1038/s41592-022-01454-x

Laird, T. S., Flyangolts, K., Bartling, C., Gemler, B. T., Beal, J., Mitchell, T., et al. (2025). *Inter-tool analysis of a NIST dataset for assessing baseline nucleic acid sequence screening* (p. 2025.05.30.655379). *bioRxiv*. doi:10.1101/2025.05.30.655379

Li, Q., Hu, Z., Wang, Y., Li, L., Fan, Y., King, I., et al. (2024). Progress and opportunities of foundation models in bioinformatics. *Briefings Bioinforma.* 25 (6), bbae548. doi:10.1093/bib/bbae548

Lin, B., Luo, X., Liu, Y., and Jin, X. (2024). A comprehensive review and comparison of existing computational methods for protein function prediction. *Briefings Bioinforma.* 25 (4), bbae289. doi:10.1093/bib/bbae289

Liu, B., Zheng, D., Zhou, S., Chen, L., and Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res.* 50 (D1), D912–D917. doi:10.1093/nar/gkab1107

Mackelprang, R., Rivera, S., Klonowski, J., Smith, L., and Hook-Barnard, I. (2025). Strengthening a safe and secure nucleic acid synthesis ecosystem: Outcomes of EBRC stakeholder engagement. *Eng. Biol. Res. Consortium*. doi:10.25498/E4311B

Mahlich, Y., Zhu, C., Chung, H., Velaga, P. K., De Paolis Kaluza, M. C., Radivojac, P., et al. (2023). Learning from the unknown: exploring the range of bacterial functionality. *Nucleic Acids Res.* 51 (19), 10162–10175. doi:10.1093/nar/gkad757

Meng, Y., Zhang, Z., Zhou, C., Tang, X., Hu, X., Tian, G., et al. (2025). Protein structure prediction *via* deep learning: an in-depth review. *Front. Pharmacol.* 16, 1498662. doi:10.3389/fphar.2025.1498662

Moxon, R., and Tang, C. (2000). Challenge of investigating biologically relevant functions of virulence factors in bacterial pathogens. *Philosophical Trans. R. Soc. Lond. Ser. B Biol. Sci.* 355, 643–656. doi:10.1098/rstb.2000.0605

Munsamy, G., Illanes-Vicioso, R., Funcillo, S., Nakou, I. T., Lindner, S., Ayres, G., et al. (2024). Conditional language models enable the efficient design of proficient enzymes. *bioRxiv*. doi:10.1101/2024.05.03.592223

Notin, P., Kollasch, A. W., Ritter, D., Niekerk, L. V., Paul, S., Spinner, H., et al. (2023). ProteinGym: large-scale benchmarks for protein design and fitness prediction. *bioRxiv*. doi:10.1101/2023.12.07.570727

Pantolini, L., Studer, G., Pereira, J., Durairaj, J., Tauriello, G., and Schwede, T. (2024). Embedding-based alignment: combining protein language models with dynamic programming alignment to detect structural similarities in the twilight-zone. *Bioinformatics* 40 (1), btad786. doi:10.1093/bioinformatics/btad786

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., et al. (2025). Boltz-2: towards accurate and efficient binding affinity prediction. *bioRxiv*, 659707. doi:10.1101/2025.06.14.659707

Rocha, J. J., Jayaram, S. A., Stevens, T. J., Muschalik, N., Shah, R. D., Emran, S., et al. (2023). Functional unknomics: systematic screening of conserved genes of unknown function. *PLOS Biol.* 21 (8), e3002222. doi:10.1371/journal.pbio.3002222

Rose, S., Alexanian, T., Langenkamp, M., Cozzarini, H., and Diggans, J. (2024). Practical questions for securing nucleic acid synthesis. *Appl. Biosaf.* 29 (3), 159–171. doi:10.1089/apb.2023.0028

Ruffolo, J. A., Nayfach, S., Gallagher, J., Bhatnagar, A., Beazer, J., Hussain, R., et al. (2024). Design of highly functional genome editors by modeling the universe of CRISPR-cas sequences. *bioRxiv*, 2024.04.22.590591. doi:10.1101/2024.04.22.590591

Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L., et al. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* 37 (4), 358–367. doi:10.1038/s41587-019-0080-8

Schnoes, A. M., Ream, D. C., Thorman, A. W., Babbitt, P. C., and Friedberg, I. (2013). Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLOS Comput. Biol.* 9 (5), e1003063. doi:10.1371/journal.pcbi.1003063

Sumida, K. H., Núñez-Franco, R., Kalvet, I., Pellock, S. J., Wicky, B. I. M., Milles, L. F., et al. (2024). Improving protein expression, stability, and function with ProteinMPNN. *J. Am. Chem. Soc.* 146 (3), 2054–2061. doi:10.1021/jacs.3c10941

Tayouri, S., Kogan, V., Beal, J., Levy, T., Farbiash, D., Flyangolts, K., et al. (2025). Defending synthetic DNA orders against splitting-based obfuscation. *bioRxiv*, 2025.03.12.642526. doi:10.1101/2025.03.12.642526

The White House (2025). *Improving the safety and security of biological research* (executive order no. 14292). Available online at: https://www.whitehouse.gov/presidential-actions/2025/05/improving-the-safety-and-security-of-biological-research/.

UK DSIT (2024). UK screening guidance on synthetic nucleic acids for users and providers. *Dep. Sci. Innovation, and Technol. United Kingdom. GOV.UK*. Available online at: https://www.gov.uk/government/publications/uk-screening-guidance-on-synthetic-nucleic-acids/uk-screening-guidance-on-synthetic-nucleic-acids-for-users-and-providers.

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., et al. (2022). PHI-Base in 2022: a multi-species phenotype database for pathogen–host interactions. *Nucleic Acids Res.* 50 (D1), D837–D847. doi:10.1093/nar/gkab1037

US NSTC (2024). *Framework for nucleic acid Synthesis screening* [government report]. *Exec. Office Pres. Office Sci. Technol. Policy (OSTP)*. Available online at: https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/04/Nucleic-Acid_Synthesis_Screening_Framework.pdf.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., et al. (2024). Fast and accurate protein structure search with foldseek. *Nat. Biotechnol.* 42 (2), 243–246. doi:10.1038/s41587-023-01773-0

Vought, R. T., and Kratsios, M. J. (2025). National science and technology Memorandum-2: fiscal year (FY) 2027 administration research and development budget priorities and cross-cutting actions. Available online at: https://www.whitehouse.gov/wp-content/uploads/2025/09/M-25-34-NSTM-2-Fiscal-Year-FY-2027-Administration-Research-and-Development-Budget-Priorities-and-Cross-Cutting-Actions.pdf.

Wheeler, N. E., Carter, S. R., Alexanian, T., Isaac, C., Yassif, J., and Millet, P. (2024). Developing a common global baseline for nucleic acid synthesis screening. *Appl. Biosaf.* 29 (2), 71–78. doi:10.1089/apb.2023.0034

Williams, A., Popescu, S., Berke, A., Vazquez, E., and Nevo, S. (2025). *Identifying and closing gaps in the federal select agent program: opportunities for improvement in an era of emerging biotechnologies*. RAND. Available online at: https://www.rand.org/pubs/research_reports/RRA3628-1.html.

Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., et al. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science* 390 (6768), 82–87. doi:10.1126/science.adu8578

Wittmann, B. J., Wheeler, N. E., Murphy, S. T., Mitchell, T., Magalis, B., Gemler, B. T., et al. (2026). The limits of sequence-based biosecurity screening tools in the age of AI-Assisted protein design. *bioRxiv*, 2026.03.04.709671. doi:10.64898/2026.03.04.709671

Wyschogrod, D., Manthey, J., Mitchell, T., Murphy, S., Clore, A., and Beal, J. (2022). Adapting malware detection to DNA screening. *GitHub*. Available online at: https://jakebeal.github.io/Publications/IWBDA2022-FASTNA.pdf.

Yeh, A. H.-W., Norn, C., Kipnis, Y., Tischer, D., Pellock, S. J., Evans, D., et al. (2023). *De novo* design of luciferases using deep learning. *Nature* 614 (7949), 774–780. doi:10.1038/s41586-023-05696-3

Zhang, G., Liu, C., Lu, J., Zhang, S., and Zhu, L. (2025). The Role of AI-Driven *de novo* Protein Design in the Exploration of the Protein Functional Universe. *Biology* 14 (9), 1268. doi:10.3390/biology14091268