ChemComm

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2025, **61**, 19028

Received 18th September 2025,
Accepted 9th October 2025

DOI: 10.1039/d5cc05032d

rsc.li/chemcomm

Hybrid AI/physics pipeline for miniprotein binder prioritization: application to the BRD3 ET domain

Jokent Gaza, [ab] Monica J. Roth, [c] Gaetano T. Montelione [de] and Alberto Perez *[ab]

AI-based protein design can rapidly generate thousands of candidate binders, but most fail to fold or bind productively, creating a critical need for robust prioritization. We present a generalizable hybrid pipeline that integrates deep-learning design and physics-based simulations to filter large libraries down to a handful of high-confidence candidates.

The bromodomain and extra-terminal (BET) family of proteins—BRD2, BRD3, BRD4, and BRDT—plays a critical role in regulating gene expression involved in immune response, cell cycle progression, inflammation, and cancer.[1] These proteins contain two bromodomains (BD1 and BD2) that act as epigenetic readers, recognizing acetylated chromatin, and an extraterminal (ET) domain that recruits transcription factors, elongation factors, and chromatin remodelers.[2,3] Therapeutic strategies have primarily focused on targeting the bromodomains,[4,5] but their structural similarity across BET paralogs and other bromodomain-containing proteins has limited selectivity, resulting in pan-inhibition and associated toxicities.[6–8] In contrast, the less studied ET domain functions as a selective interaction hub, binding short peptide epitopes from both host and viral proteins that typically adopt β-hairpin structures upon binding and bind with affinities ranging from millimolar to nanomolar.[9,10] Subtle sequence differences across ET domains offer a unique opportunity for designing selective inhibitors.

Peptide-based inhibitors often display high specificity for their targets, but their clinical translation is limited by poor stability, proteolytic susceptibility, and structural disorder in solution.[11] Miniproteins have emerged as promising alternatives: small, well-folded scaffolds that retain binding affinity while improving proteolytic stability and structural robustness.[12] Recent advances in AI-based tools have democratized miniprotein design, enabling the rapid generation of candidates that embed known peptide-binding motifs into stable protein frameworks. However, the vast majority of these *de novo* sequences are unlikely to fold correctly or bind with high affinity, making prioritization a critical challenge. In our previous work,[13] we used AlphaFold-based competitive binding assays (AF-CBA[14]) to identify peptide binders and binding sites from pulldown libraries. Building on these insights, we now explore whether miniproteins incorporating these peptide interaction motifs can be designed to bind with higher selectivity and remain robust to degradation. To do so, we developed a hybrid design and filtering pipeline (Fig. 1) capable of selecting high-quality binders from large AI-generated sequence libraries. Although the ET domains of BRD2, BRD3, and BRD4 all interact with murine leukemia virus integrase, we focused on BRD3-ET due to our extensive biochemical, biophysical, and structural characterization of this system.[10,13,15,16]

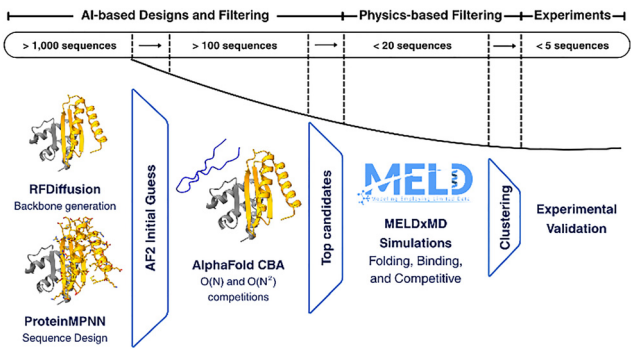

Fig. 1 Hybrid approach for designing novel miniproteins. Protein backbones are generated using RFDiffusion, then decorated with functional groups using ProteinMPNN, and validated using AlphaFold2 initial guess. The preliminary designs are filtered using AF-CBA to select the top ~20–30 designs. For the MELD simulations, we perform three tests: folding, binding, and competitive binding. Designs that pass all these tests are considered for further experimental validations.

[a] *Department of Chemistry, University of Florida, Gainesville, Florida 32611, USA. E-mail: perez@chem.ufl.edu*
[b] *Quantum Theory Project, University of Florida, Gainesville, Florida 32611, USA*
[c] *Department of Pharmacology, Rutgers-Robert Wood Johnson Medical School, 675 Hoes Lane Rm 636, Piscataway, NJ 08854, USA*
[d] *Center for Biotechnology and Interdisciplinary Sciences, Rensselaer Polytechnic Institute, Troy, New York, 12180, USA*
[e] *Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, New York, 12180, USA*

Starting from five known ET-binding peptide motifs (Fig. S1), we used RFDiffusion[17] to design 3000 miniproteins per peptide with target length of 70–120 residues. The number of initial designs was based on the reported *in silico* success rate of the binder design pipeline using predicted aligned error (pAE) as the confidence metric.[18] We reasoned that with a pool large enough to statistically contain a reasonable number of successful designs, a prioritization pipeline should be able to recover the most promising ones. Conditional protein design preserved the peptide's hairpin interaction while scaffolding it with additional secondary structure to promote folding. ProteinMPNN[19] was then used to assign amino acid sequences compatible with these backbones, resulting in a library of 15 000 *de novo* sequences. As many of these candidates may be misfolded or non-binders, AlphaFold2 initial guess[18,20] offers a fast, orthogonal AI validation strategy to RosettaFold. As a first-pass filter, we applied the pAE to estimate structural confidence and binding mode quality. Designs with low pAE scores (pAE < 10) were retained, yielding 823 candidates. However, this set remains too large for most experimental efforts. While high-throughput groups may test hundreds of designs, most collaborative settings require prioritization of a small number of candidates with the highest likelihood of success.

To further prioritize candidates with high predicted binding affinity, we applied our previously developed AF-CBA,[14] which enables side-by-side structural prediction of multiple binders competing for a shared binding site. In this framework, sequences that consistently occupy the binding site more frequently than others are inferred to have higher relative binding affinity. We implemented this in a tiered manner to reduce computational cost:[13] the first O(N) stage filtered designs against five random miniproteins in the set, followed by an O(N^2) competition to rank the promising candidates. This filter reduced the pool from 823 to just 20 candidates.

However, structural inspection of these top-ranked designs revealed a systematic flaw. Many miniproteins exhibited elongated helical elements that disrupted globularity, resulting in high radius of gyration (RoG) values and poor packing around the binding domain (Fig. S2a). Retrospective analysis indicated that this was likely due to biases in the default RFDiffusion model weights we used which favored extended helices rather than compact folds. Despite this limitation, the designs preserved key features along the complex interface residues. All designs retained the canonical hairpin interface and a motif of alternating hydrophobic/charged residues in the binding epitope,[21–23] and several formed a conserved hydrogen bond between ET-domain residue Asp612 and a nearby basic residue on the binder (Fig. S2b). Aligning and analyzing these sequences led us to define two new structural motifs. Motif I combines features from CHD4:ET and NSD3:ET complexes, and Motif II uses elements from CHD4:ET and TP:ET complexes (Fig. S2d).

For the second round of miniprotein design, we surmised that using the Complex_beta weights in RFDiffusion would improve the likelihood that the new miniproteins would fold into globular structures. Using these weights, we generated 3000 backbones for each two motif (Motif I and Motif II), and proposed sequences *via* ProteinMPNN to produce a second design library of 6000 candidates. To determine whether improvements in design quality stemmed from the new motifs or from the updated generative weights, we also created a control set of 6000 sequences using the same motifs but the default RFDiffusion weights.

All designs were filtered using AlphaFold2 initial guess, retaining only those with high-confidence structures (pAE $<$ 10). We then applied AF-CBA to prioritize binders predicted to outcompete the viral TP peptide, and implemented two additional structure-based criteria: (1) an RoG $<$ 14 Å to exclude non-globular scaffolds, and (2) an RMSD $<$ 2 Å between bound and unbound conformations to identify candidates with minimal structural rearrangement upon binding. The latter serves as a proxy for minimizing the conformational free energy cost of binding, which can improve affinity and functional robustness.

Among the 6000 designs generated with Complex_beta weights, 31 (9 from Motif I and 22 from Motif II) passed all filters. In contrast, although some designs from the control set showed favorable AF2 predictions, none satisfied both structural criteria. This result underscores the importance of selecting the right model weights. With this manageable set of 31 compact, stable candidates in hand, we proceeded to physics-based validation using MELD simulations as an orthogonal test of folding and binding fidelity (Fig. 2).

MELD is an enhanced sampling approach that incorporates ambiguous and noisy information, such as generic heuristics about protein folding (*e.g.*, hydrophobic residues tend to form cores), to molecular dynamics and infers structures that are consistent with some subset of the data and the physics model using Bayesian inference.[24,25] During CASP evaluations, MELD has shown success in modeling designed proteins,[26] motivating its application to study miniproteins.

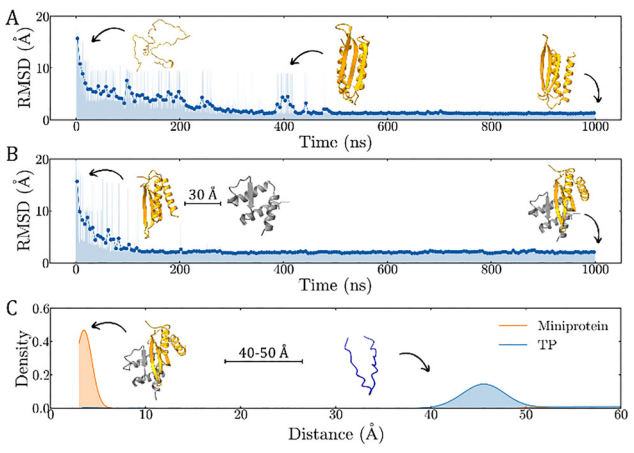

Fig. 2 Three MELD orthogonal tests for AI-designed miniproteins. (A) Folding simulation from the extended structure using coarse physical insights. (B) Binding simulation to validate the binding mode of the miniproteins. Both folding and binding simulations used AF2 structures as references for the RMSD calculations. (C) Competitive simulation in which both TP and the miniprotein compete for the same binding site. Relative binding affinity (calculated $\Delta\Delta G$) is estimated from the bound populations of each binder across the five lowest replicas.

For each of the 31 candidate designs, we carried out folding simulations using only sequence and secondary structure predictions as input. The resulting ensembles were analyzed *via* clustering to identify dominant metastable states. The higher the population of the top cluster and the more independent replicas (walkers) that sample it, the higher our confidence that the folded state is stable and accessible. Of the 31 designs, 21 exhibited top clusters with populations exceeding 50%, indicating a well-defined folding basin (Fig. S3 and Table S1). In 20 of these 21 cases, the representative structure of the dominant cluster was in excellent agreement with the AlphaFold-predicted model (RMSD < 5 Å), suggesting strong convergence between physics-based and AI-based predictions. Even among lower-population designs, several retained good structural agreement, indicating that MELD can recover the native fold even when sampling a more heterogeneous distribution.

After validating that our miniprotein designs could fold reliably, we next assessed whether they could bind the BRD3 ET domain in the expected manner. We used previous chemical shift perturbation data to define the possible binding sites in ET.[16] Given that both the ET domain and the designed miniproteins were predicted to be stably folded, these simulations focused exclusively on binding, without modeling folding upon association. MELD binding simulations apply restraints to preserve native-like flexibility while preventing global unfolding during enhanced sampling at elevated temperatures. These ambiguous restraints allow the proteins to explore multiple binding modes, enabling an ensemble-level view of binding specificity.

We analyzed the resulting complexes through clustering to identify dominant binding modes (Fig. S3 and Table S1). For 20 of the 31 designs, the top population cluster exceeded 50%, indicating a strong preference for a specific binding mode. Of these, 19 bound in a geometry consistent with the AlphaFold- or RoseTTAFold All-Atom[27] (RFAA)-predicted models. Selecting only those designs that showed agreement across MELD, AlphaFold2, and RFAA models yielded 12 high-confidence candidates: 4 from Motif I and 8 from Motif II (Fig. S4 and Table S1). Interestingly, the hairpin regions of these designs preserved distinct features of their source motifs beyond a known pattern of alternating hydrophobic/charged residues creating a zipper like interaction between the peptide and the receptor. Motif I designs based on NSD3 exhibited a flipped β-sheet orientation relative to Motif II designs derived from TP.

As a final filter, we applied the MELD Competitive Binding Assay (MELD-CBA) to evaluate whether each of the 12 surviving miniprotein designs could outcompete the TP peptide, a known high-affinity binder, for the ET domain binding site. In these simulations, both the miniprotein and TP were introduced simultaneously, and we monitored their occupancy of the ET domain across replica indices. While high-temperature replicas emphasize entropic flexibility, low-temperature replicas reflect enthalpic stabilization and shape complementarity.

Among the 12 designs, none of the four Motif I candidates were able to consistently outcompete TP. In contrast, 5 of the eight Motif II designs exhibited dominant binding at the lowest temperature replicas, indicating stronger enthalpic interactions

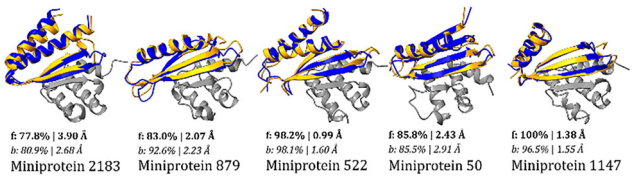

Fig. 3 Top five designs. MELD structures (orange) and AF2 structures (blue) are shown. For clarity purposes, only ET MELD structures are displayed. Cluster population and centroid RMSD are reported for folding (f) and binding (b) simulations.

(Fig. S5 and Fig. 3). Interestingly, the replica-dependent binding profiles revealed diverse thermodynamic behaviors. For one design (Miniprotein 2183), the miniprotein outcompeted TP consistently across all replicas, thus suggesting both favorable entropy and enthalpy. Four miniproteins (Miniproteins 879, 522, 50, and 1147), on the other hand, only dominated at low temperatures, implying a higher entropic cost compensated by stronger binding interactions.

These results highlight the nuanced balance between conformational flexibility and binding strength. The disordered TP peptide can rapidly sample orientations and form initial contacts but ultimately incurs a higher penalty as it folds upon binding. In contrast, pre-folded miniproteins may be slower to sample binding-compatible conformations at high temperatures but exhibit stronger, more specific interactions at lower temperatures.

In terms of protein–protein interactions, a known hotspot in the BRD3 ET domain is a hydrophobic pocket (Fig. 4) near VAL596, where high-affinity binders such as the viral TP peptide and host protein NSD3 typically contribute a tryptophan or phenylalanine residue.[10] Notably, 12 of our top 31 MELD-validated designs (5 from Motif I and 7 from Motif II) incorporated a tryptophan at this position, despite no explicit biasing during design. This convergence reinforces the biological relevance of the selected binders and highlights the pipeline's ability to recover key molecular recognition features. Furthermore, the presence of a surface-exposed tryptophan in the unbound state that becomes buried upon binding also provides a convenient feature for future fluorescence-based binding assays. To assess experimental feasibility, we evaluated common N-terminal tags using MELD binding simulations (Fig. S6–S9 and Tables S2, S3). All constructs retained the expected binding mode, with AviTag variants showing the lowest perturbation. *In silico* assessment also showed high expected stability at 65 °C, low aggregation propensity and high solubility (Tables S4 and S5).

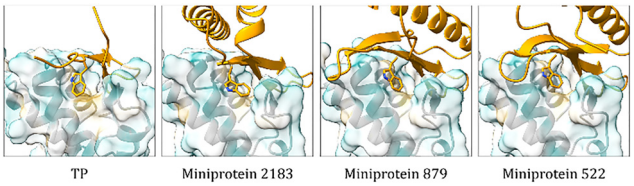

Fig. 4 Interactions at a key hydrophobic pocket for TP and the top three designs. The ET domain is shown as a surface, with hydrophobic regions colored orange and hydrophilic regions colored cyan.

Our results demonstrate how an integrated AI/physics pipeline can reduce thousands of designs to a handful of compact and stable candidate binders. Based on these predictions, our team is acquiring these constructs for experimental validation, which will be published in future work, comparing to our deposited predictions. While applied here to BRD3-ET, the strategy is broadly applicable to peptide-derived motifs and other flexible protein–protein interactions. By releasing all predictions as blind benchmarks, we aim to promote transparency, reproducibility, and community-wide validation, in line with FAIR principles and the spirit of CASP and CAPRI style challenges. This work provides a practical resource for BET targeting while also serving as a blueprint for prioritizing designs in binder discovery pipelines.

Conflicts of interest

GTM is a founder and advisor to Nexomics Biosciences, Inc., which does not constitute a conflict of interest for this study.

Data availability

Supporting data are included in the supplementary information (SI). Supplementary information is available. See DOI: **https://doi.org/10.1039/d5cc05032d**.

All sequences and results are uploaded in our GitHub repository (**https://github.com/PDNALab/Miniprotein_Design**) and in Zenodo DOI: **https://doi.org/10.5281/zenodo.16755842**.

Notes and references

1 N. Wang, R. Wu, D. Tang and R. Kang, *Signal Transduction Targeted Ther.*, 2021, **6**, 23.
2 C. Dhalluin, J. E. Carlson, L. Zeng, C. He, A. K. Aggarwal, M.-M. Zhou and M.-M. Zhou, *Nature*, 1999, **399**, 491–496.
3 K. Fujinaga, F. Huang and B. M. Peterlin, *Mol. Cell*, 2023, **83**, 393–403.
4 L. R. Vidler, N. Brown, S. Knapp and S. Hoelder, *J. Med. Chem.*, 2012, **55**, 7346–7359.
5 J. Shi and C. R. Vakoc, *Mol. Cell*, 2014, **54**, 728–736.
6 M. Petretich, E. H. Demont and P. Grandi, *Curr. Opin. Chem. Biol.*, 2020, **57**, 184–193.
7 D. B. Doroshow, J. P. Eder and P. M. LoRusso, *Ann. Oncol.*, 2017, **28**, 1776–1787.
8 A. Alqahtani, K. Choucair, M. Ashraf, D. M. Hammouda, A. Alloghbi, T. Khan, N. Senzer and J. Nemunaitis, *Future Sci. OA*, 2019, **5**, FSO372.
9 S. Rahman, M. E. Sowa, M. Ottinger, J. A. Smith, Y. Shi, J. W. Harper and P. M. Howley, *Mol. Cell. Biol.*, 2011, **31**, 2641–2652.
10 S. Aiyer, G. V. T. Swapna, L.-C. Ma, G. Liu, J. Hao, G. Chalmers, B. C. Jacobs, G. T. Montelione and M. J. Roth, *Structure*, 2021, **29**, 886–898.
11 L. Di, *AAPS J.*, 2015, **17**, 134–143.
12 Z. R. Crook, N. W. Nairn and J. M. Olson, *Trends Biochem. Sci.*, 2020, **45**, 332–346.
13 A. Mondal, B. Singh, R. H. Felkner, A. D. Falco, G. Swapna, G. T. Montelione, M. J. Roth and A. Perez, *Angew. Chem., Int. Ed.*, 2024, **63**, e202405767.
14 L. Chang and A. Perez, *Angew. Chem., Int. Ed.*, 2023, **62**, e202213362.
15 A. Sharma, R. C. Larue, M. R. Plumb, N. Malani, F. Male, A. Slaughter, J. J. Kessl, N. Shkriabai, E. Coward, S. S. Aiyer, P. L. Green, L. Wu, M. J. Roth, F. D. Bushman and M. Kvaratskhelia, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 12036–12041.
16 A. Mondal, G. V. T. Swapna, M. M. Lopez, L. Klang, J. Hao, L. Ma, M. J. Roth, G. T. Montelione and A. Perez, *J. Chem. Inf. Model.*, 2023, **63**, 2058–2072.
17 J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. D. Bortoli, E. Mathieu, S. Ovchinnikov, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek and D. Baker, *Nature*, 2023, **620**, 1089–1100.
18 N. R. Bennett, B. Coventry, I. Goreshnik, B. Huang, A. Allen, D. Vafeados, Y. P. Peng, J. Dauparas, M. Baek, L. Stewart, F. DiMaio, S. D. Munck, S. N. Savvides and D. Baker, *Nat. Commun.*, 2023, **14**, 2625.
19 J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King and D. Baker, *Science*, 2022, **378**, 49–56.
20 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.
21 D. C. C. Wai, T. N. Szyszka, A. E. Campbell, C. Kwong, L. E. Wilkinson-White, A. P. G. Silva, J. K. K. Low, A. H. Kwan, R. Gamsjaeger, J. D. Chalmers, W. M. Patrick, B. Lu, C. R. Vakoc, G. A. Blobel and J. P. Mackay, *J. Biol. Chem.*, 2018, **293**, 7160–7175.
22 Q. Zhang, L. Zeng, C. Shen, Y. Ju, T. Konuma, C. Zhao, C. R. Vakoc and M.-M. Zhou, *Structure*, 2016, **24**, 1201–1208.
23 B. L. Crowe, R. C. Larue, C. Yuan, S. Hess, M. Kvaratskhelia and M. P. Foster, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 2086–2091.
24 J. L. MacCallum, A. Perez and K. A. Dill, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 6985–6990.
25 A. Perez, J. L. MacCallum and K. A. Dill, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 11846–11851.
26 A. Perez, J. A. Morrone, E. Brini, J. L. MacCallum and K. A. Dill, *Sci. Adv.*, 2016, **2**, e1601274.
27 R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, R. McHugh, D. Vafeados, X. Li, G. A. Sutherland, A. Hitchcock, C. N. Hunter, A. Kang, E. Brackenbrough, A. K. Bera, M. Baek, F. DiMaio and D. Baker, *Science*, 2024, **384**, eadl2528.