nature portfolio

Peer Review File

Targeting protein-ligand neosurface using a generalizable deep learning approach

Corresponding Author: Professor Bruno Correia

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Referee #1

In the manuscript by Marchand et al., the authors present a computational pipeline for the de novo design of proteins targeting neosurfaces (surfaces arising from protein-ligand interfaces). The work constitutes an extension of their previously described computational methods for the design of protein-protein interactions using a geometric deep-learning approach. The ability to generalize their in silico tools to incorporate neosurfaces represents a significant advance and expands the applications accessible by their methods.

Historically, the discovery of small molecules able to induce protein–protein associations has been challenging and the small collection of compounds able to do so have most commonly been identified retrospectively and serendipitously. However, with the expanding applications of chemical inducers of proximity (CIPs), there has been an explosion of interest in the targeted discovery of such compounds. While experimental strategies for identifying CIPs have steadily grown, the ability to design such compounds in silico has not yet been realized. The manuscript by Marchand et al. shows for the first time that these intricate ternary complexes can be successfully computed using the power of deep-learning tools.

The authors first modified their original “protein-only” framework to allow for the incorporation of small-molecule surfaces. Then, three ligand-bound structures were selected from the PDB and the authors set out to identify ligand-dependent binders to these neosurfaces. Remarkably, the initial in silico screens identified moderate to weak affinity binders for each of the three targeted neosurfaces, which were further optimized into potent binders using established methods. When the optimized structures were profiled by BLI for binding affinity, they showed exquisite ligand-dependent binding (Fig. 4c), in some cases reaching shifts of several orders of magnitude between ligand-bound and apo targets. Further validation of the new MaSIF-neosurf pipeline is provided by a ternary complex crystal structure of one engineered complex (Fig. 4d), and a cryo-EM ternary structure of another (Fig. 4e), both showing reasonable agreement between the computationally modelled and experimentally determined complexes. Finally, CIP activity was further validated in a suite of biochemical and cell-based assays (Fig. 5).

While most CIP discovery efforts to date have focused on identifying the small-molecule component of a desired ternary complex (incuding for targeted protein degradation), the authors present a unique approach targeting the de novo design of an engineered protein component of a ternary complex. The resulting complexes could find applications in synthetic biology, where chemically controlled “ON-switches” have been proposed, for example, as safety switches for cell-based therapies. However, the breadth of applications for such systems is limited by the engineered protein components, and tools directed towards identifying small molecule CIDs for endogenous proteins will have a larger impact on therapeutic discovery. Nevertheless, the tools presented in this manuscript provide a valuable foundation for a nascent field where the intricacies of ternary complexes present significant challenges. As such, we support publication of the manuscript by Marchand et al. pending the authors' responses to the following comments:

1) In the initial benchmarking of the MaSIF-neosurf pipeline, only 14 ligand-induced protein complexes were selected for testing (albeit with two different splits to generate 28 “true-positives”). Furthermore, only 200 decoy proteins were included. Therefore, the benchmarking exercise amounts to selecting a true positive from 228 choices. This represents a poor proxy for the desired application of selecting a true binder from an enormous potential search space (i.e. all possible protein surface architectures). a) Why were only 200 decoys chosen? b) Are the 200 decoys chosen to be reasonably similar to the positive cases? This benchmark is only informative if the negative examples are non-trivially distinguishable from the

positive examples.

2) How were the three test cases selected for the de novo design of binders, considering the huge number of potential binary complexes available in the PDB? Were other binary complexes targeted unsuccessfully? It seems quite remarkable that all three test cases targeted by the authors ultimately delivered successful high-affinity binders. Are there particular aspects of these complexes that the authors believe contributed to this success rate? Would similar performance be expected for other cases as well?

3) On a similar note, is it expected that the patch database based on 640,000 proteins would be sufficiently large to contain compatible patches for virtually any protein/ligand target, as is suggested by the 100% success rate?

4) To understand how small molecules are incorporated into the MaSIF models, it would be useful to understand the importance of each of the 5 features used (i.e. shape index, distance-dependent curvature, Poisson-Boltzmann continuum electrostatics, hydrogen bond donor and acceptor potential, hydropathy). In particular, the choice of a hydropathy measure that extends to small molecules is (as the authors point out) not obvious. Given the large number of design choices involved in the author's model I would like to understand how sensitive the model is to these choices. At the most extreme - how does the model perform if this measure is removed (or randomized) during testing?

5) Following the in silico screening of the three ligand-dependent protein binders, the authors selected ~2,000 top ranking seeds for each target and screened them by yeast display. Following two rounds of FACS enrichment, the isolated clones were sequenced and the authors selected one binder for each test case to characterize more deeply. What happened to the other 1,999 seeds? Were the selected seeds overwhelmingly enriched during yeast screening? Were other high-ranking seeds enriched? Based on these observations, how many seeds would need to be screened to identify binders for most targets? Furthermore, this approach assumes that 2k randomly selected seeds would provide no enrichment for these same targets. Given the modest enrichment values reported in Table S2 and the unique properties of the ligands used for these test cases, it would be valuable to show experimentally that this assumption is true. For most binary complexes, I would not expect that 2k random proteins would identify a neosurface binder; however, for highly lipophilic ligands like progesterone, I would not be surprised to see similar enrichments from a random protein sampling.

6) To validate functionally the identified CIDs, the authors use three different biochemical and cell-based assays, but only test one binder in each assay (Fig. 5). It would be much more informative if all three CIPs were tested in two functional assays, one extracellular and one intracellular (Fig. 5d,g). At present, the individual assays developed for each complex create uncertainty regarding the functional effects of the proteins in different systems.

7) Regarding the first functional assay (Fig. 5a-c), the authors report an IC50 of 1.2 μM; however, the ternary complex KD reported for this complex is 18 nM. Since this is a cell-free system, why does the functional readout show a shift of nearly two orders of magnitude?

Referee #2

This manuscript reports a computational method for designing protein surfaces for binding to neosurfaces. A neosurface is a protein-ligand interface where small molecules serve as 'molecular glue' that facilitates protein-protein binding. The computational methodology is closely related to that of a study published last year (ref. 9). The key difference is that in ref. 9, protein-protein interfaces are considered without ligand modification. The protein-ligand interface is featurized with properties including hydrophobicity and electrostatics. A fingerprint is built from these features using a neural network. The neosurface fingerprint is then input to the design/selection pipeline to generate a protein binding partner, where complementary interfaces are found by searching a library of protein surface fingerprints.

This technique is validated on a small number of known tertiary complexes and then used to design proteins that target three neosurfaces. The in-silico predictions are experimentally validated in a series of experiments including binding, mutagenesis and crystallography that characterize the affinity and structure of the binding complex and identify beneficial mutations. It is found that the binding is specific and targets the predicted neosurface interface. It is also shown that the complexes studied here can be employed in cellular systems and used as, for example, biosensors. The work is put in context with highly-visible protein structure prediction models in supplemental materials (see Figs S2 and S16).

The protein design predicted by the model is extensively validated through a series of wet-lab experiments and all presented data are robust and of high-quality. The presentation and conclusions are also clear and well-supported. However, while the computational approach is clearly valid and useful, it can be considered an added feature to the method presented by the authors in Refs. 8,9 for treating the case of neosurfaces. It is therefore unclear if it represents a significant advancement in the field for publication in Nature. The following suggestions may strengthen the manuscript in a future submission:

- Very recently, AlphaFold3 has been released which incorporates small molecules into their model [https://doi.org/10.1038/s41586-024-07487-w]. This represents an advancement over AlphaFold2, which is used in Fig. S16. How would AlphaFold3 perform or enhance the task of targeting protein neo-surfaces?

- The current computational method appears applicable to given protein-ligand interfaces, but leaves open the question of the design or selection of the ligand or 'molecular glue' itself. This would seem to introduce (another) large search space. How feasible is this as an extension of the present work?

- The computational approach of this work utilizes physical and chemical high-level ‘expertly chosen’ features such as Poisson-Boltzmann electrostatics and hydrophobicity scales when compared with recent approaches that may use only atoms/coordinates or sequences to represent the inputs. Does this choice allow the use of neural networks of lower complexity in the present work? Does the use of physical modeling such as electrostatics calculations limit the throughput of the technique?

- Note that Ref. 29 has been published [DOI: 10.1126/science.adl2528].

(Remarks on code availability)
Although I have not extensively reviewed the source, I have looked over the documentation. It is very well documented and gives the code/instructions to use the technique and reproduce the initial benchmark they perform. The work therefore appears reproducible for the code in the repo. It is usable by the community.

I have not tried to run the code. It would require many packages, typical for multi-disiplinary projects encompassing neural nets, cheminformatics, and protein modeling. Late stages in their pipeline would require the install of Rosetta as mentioned in the manuscript.

Referee #3

In the manuscript by Marchand et al., the authors build on their previously published model, MaSIF, to develop MaSIF-neosurf, expanding the target domain of protein binder design to protein-ligand complexes. The authors first demonstrate that the new method can enrich 14 ligand-dependent complexes over a library of decoy complexes (Fig. 1). They then use this method to design de novo protein binders targeting three neosurfaces (Fig. 2). Several initial and optimized designs have been characterized through binding, biochemical assays, and structural validations (Figs. 3 and 4). The optimized designs have also been tested in split complementation assays in cell-free and mammalian cell systems (Fig. 5).

This paper is well-done and represents a collaborative effort combining deep learning methods, yeast display, biochemical and cellular evaluations, and detailed structural studies to explore potential use cases for creating binders to protein-ligand complexes, an outstanding challenge in protein design. Although there is a question as to how efficient this method is compared to experimental methods, such as AbCIDs (Hill et al., Nat. Chem. Biol. (2018) 14:112-117), this study is a first step toward expanding deep learning methods to design selective binders for protein-ligand complexes. The implications of this study are particularly useful for developing molecular glues in synthetic biology, and, of course, protein design.

I could not detect any significant flaws in the work. Nevertheless, I have a few comments that I believe will improve the strength of the main takeaways for a broad audience.

Regarding comments and suggestions, I have four:

1. Comparison to experimental methods: Deep learning methods are undoubtedly improving, but how do they compare to existing experimental methods in terms of the success rate of both pipelines, the best potency straight out of initial design, preferences for certain targets, and whether deep learning methods have surpassed the efficiency of experimental methods in this task? What are the pros and cons of this method compared to experimental methods such as AbCIDs? This comparison and discussion would be informative for a general audience to better understand these methods and choose proper methods for their targets.

2. Generalizability and success rate discussion: How potent should ligand-protein complexes be? What are the lower and upper potency boundaries within which this method can be applied?

3. Public availability of site saturation mutagenesis data: The authors should consider making the site-saturation mutagenesis data publicly available in .csv or other machine-readable formats. This information could be useful as a benchmark in the field for similar design strategies and also for improvements to the current method.

4. Consistency in chemical drawings: The chemical drawing style is not consistent throughout the manuscript. The chemical drawing style in Fig. 3a is different from that in Fig. 3d for Bcl2. The same issue is present in Fig. S6a,b.

Version 2:

Reviewer comments:

Referee #1

In the resubmitted manuscript, the authors have done a nice job of addressing all our comments. We were especially happy to see the added discussion around a protein-ligand complex that was unsuccessful with their approach (JQ1-BRD4; SI Fig. 23 and Discussion), which we believe provides a more balanced presentation of their work. Additional improvements include the enlarged set of decoys used in their benchmarking experiments, which strengthens the MaSIF validation; the use

of LigandMPNN to improve the success of the seed grafting process; and the added functional validation experiments, in particular the CID-CAR experiment, to further validate the success and utility of MaSIF-neosurf. Given these modifications, we support publication of the manuscript in its revised form.

(Remarks on code availability)
The code deposited at https://github.com/LPDI-EPFL/masif-neosurf was reviewed and found to be complete and well documented, including instructions for running the main analyses provided in the manuscript

Referee #2

I thank the authors for their detailed response. They have addressed my concerns and I suggest no further changes.

I agree that the level of generalizability that this method displays is indeed uncommon. I look forward to future work which may more fully address comparisons between this method and emerging models in the literature, where generalizability serves as a key 'metric' for consideration.

(Remarks on code availability)
Please see my comments on the original submission

Referee #3

The authors have addressed all my comments. I have no further comments and would recommend publishing the manuscript as it is.

Referee #4

The experiments with the CAR-T cells are well performed. There is some baseline killing in the absence of drug (leakiness) that should be commented on (Fig. 5k). In addition, while this provides proof of concept for the CAR dimerization, the data in Fig. 5i as well as Supp. 18 shows that the split CAR approach is likely less potent than a traditional CAR, which is not surprising and should be commented on. It should be noted that drug dimerization of CAR is not a novel approach and has already been in the clinic, but that does not appear to be the point of this manuscript so this should be acceptable as POC.

Authors' response to referees' comments (version 1):

Referee #1:

1.1 In the manuscript by Marchand et al., the authors present a computational pipeline for the de novo design of proteins targeting neosurfaces (surfaces arising from protein-ligand interfaces). The work constitutes an extension of their previously described computational methods for the design of protein-protein interactions using a geometric deep-learning approach. The ability to generalize their *in silico* tools to incorporate neosurfaces represents a significant advance and expands the applications accessible by their methods.

Historically, the discovery of small molecules able to induce protein–protein associations has been challenging and the small collection of compounds able to do so have most commonly been identified retrospectively and serendipitously. However, with the expanding applications of chemical inducers of proximity (CIPs), there has been an explosion of interest in the targeted discovery of such compounds. While experimental strategies for identifying CIPs have steadily grown, the ability to design such compounds *in silico* has not yet been realized. The manuscript by Marchand et al. shows for the first time that these intricate ternary complexes can be successfully computed using the power of deep-learning tools.

The authors first modified their original "protein-only" framework to allow for the incorporation of small-molecule surfaces. Then, three ligand-bound structures were selected from the PDB and the authors set out to identify ligand-dependent binders to these neosurfaces. Remarkably, the initial *in silico* screens identified moderate to weak affinity binders for each of the three targeted neosurfaces, which were further optimized into potent binders using established methods. When the optimized structures were profiled by BLI for binding affinity, they showed exquisite ligand-dependent binding (Fig. 4c), in some cases reaching shifts of several orders of magnitude between ligand-bound and apo targets. Further validation of the new MaSIF-neosurf pipeline is provided by a ternary complex crystal structure of one engineered complex (Fig. 4d), and a cryo-EM ternary structure of another (Fig. 4e), both showing reasonable agreement between the computationally modeled and experimentally determined complexes. Finally, CIP activity was further validated in a suite of biochemical and cell-based assays (Fig. 5).

While most CIP discovery efforts to date have focused on identifying the small-molecule component of a desired ternary complex (including for targeted protein degradation), the authors present a unique approach targeting the de novo design of an engineered protein component of a ternary complex. The resulting complexes could find applications in synthetic biology, where chemically controlled "ON-switches" have been proposed, for example, as safety switches for cell-based therapies. However, the breadth of applications for such systems is limited by the engineered protein components, and tools directed towards identifying small molecule CIDs for endogenous proteins will have a larger impact on therapeutic discovery. Nevertheless, the tools presented in this manuscript provide a valuable foundation for a nascent field where the intricacies of ternary complexes present significant challenges. As such, we support publication of the manuscript by Marchand et al. pending the authors' responses to the following comments.

R: We thank the reviewer for the positive assessment of our work and the constructive comments.

1.2 In the initial benchmarking of the MaSIF-neosurf pipeline, only 14 ligand-induced protein complexes were selected for testing (albeit with two different splits to generate 28 "true-positives"). Furthermore, only 200 decoy proteins were included. Therefore, the benchmarking exercise amounts to selecting a true positive from 228 choices. This represents a poor proxy for the desired application of selecting a true binder from an enormous potential search space (i.e. all possible protein surface architectures). a) Why were only 200 decoys chosen? b) Are the 200 decoys chosen to be reasonably similar to the positive cases? This benchmark is only informative if the negative examples are non-trivially distinguishable from the positive examples.

R: We initially chose 200 decoys to keep a similar proportion as in our last publications (1000 decoys for 114 dimeric complexes) (Gainza et al, Nature, 2023). As there are on average ~4000 patches per protein, this represents a large search space with more than 900'000 potential binding sites.

While we did not specifically select decoys with high similarity to the real binders, we did choose exclusively proteins that are engaged in protein-protein interactions (PPIs) as these are more challenging decoys than monomeric proteins due to the presence of their existing binding interfaces.

To provide stronger evidence, we repeated the benchmark with a much larger decoy set of 8879 proteins involved in PPIs from the PDBbind database. By increasing the dataset size, we also increase the probability of including non-trivially distinguishable examples in the benchmark. However, we still see similar performance and success in the recovery of the known binders. We replaced the original benchmark with this improved version in the revised manuscript (see Figure 1, panel b).

1.3 How were the three test cases selected for the *de novo* design of binders, considering the huge number of potential binary complexes available in the PDB?

R: To clarify this point we added the following section about target selection to the material and methods: "The target proteins were selected based on multiple factors such as: the reported protein-ligand affinity [69]; the resolution of the structural data; the interface propensity; and the solvent-accessible surface area of the small molecule when bound to the receptor to ensure a measurable interface with the designed binders. More practical considerations such as small molecule purchase availability or feasibility of the target protein expression were also considered."

[69] Liu, Z. et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics 31, 405–412 (2015).

1.4 Were other binary complexes targeted unsuccessfully? It seems quite remarkable that all three test cases targeted by the authors ultimately delivered successful high-affinity binders. Are there particular aspects of these complexes that the authors believe contributed to this success rate? Would similar performance be expected for other cases as well?

R: We thank the reviewer for bringing up this point and we also acknowledge the importance of discussing unsuccessful design cases. In the revised manuscript, we have reported an unsuccessful case with JQ1-bound BRD4. We observed that some computational metrics of the designs targeting this complex (computed binding energy, interchain hydrogen bonds and interface contact area) were inferior to those of the other pool of designs. We also observed a potential issue with the flexibility of the ligand which is not modeled in our system (see newly added Supplementary Fig. 23). Accounting for flexible and hard-to-target regions is a known limitation of our approach and will be investigated in future works. To further clarify the point we added Supplementary Fig. 23 and the following statement in the discussion: "We also observed an unsuccessful example with BRD4:JQ1 complex, most probably due to ligand flexibility and computational metrics of the designs inferior to those of other test cases. Most deep learning methods, including ours, exhibit superior performance on hydrophobic patches, while significant challenges persist in accurately modeling polar interfaces.[9,43]"

[9] Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. Nature 617, 176–184 (2023).

[43] Cao, L. et al. Design of protein-binding proteins from the target structure alone. Nature 605, 551–560 (2022).

1.5 On a similar note, is it expected that the patch database based on 640,000 proteins would be sufficiently large to contain compatible patches for virtually any protein/ligand target, as is suggested by the 100% success rate?

R: Although the patch database is numerically large and presents a high sequence diversity (402 million surface patches/fingerprints) it may not represent a universal answer to all target sites as the seeds are mostly using helical and beta sheet conformations. It is possible that some target sites might require different shapes or more flexible conformations like loops which are harder to graft on a scaffold protein. Of note, in these design problems with several components the success rate is not only dependent on the motif used to target one site, but also on the target site and its features (accessibility, hydrophobicity, conformational rigidity of the small molecule, etc.).

1.6 To understand how small molecules are incorporated into the MaSIF models, it would be useful to understand the importance of each of the 5 features used (i.e. shape index, distance-dependent curvature, Poisson-Boltzmann continuum electrostatics, hydrogen bond donor and acceptor potential, hydropathy). In particular, the choice of a hydropathy measure that extends to small molecules is (as the authors point out) not obvious. Given the large number of design choices involved in the author's model I would like to understand how sensitive the model is to these choices. At the most extreme - how does the model perform if this measure is removed (or randomized) during testing?

R: Thank you for this suggestion. Ablation studies investigating the importance of the input features for predicting PPIs have been conducted in the publication that originally described the MaSIF method (Ref. [8], in particular, Fig. 3d and Fig. 5c). It was observed that different sets of geometric and chemical features contribute to a varying extent to the successful prediction of binding sites and complementary fingerprints. The best performance, however, was consistently achieved with a network trained on all five input features.

To answer the question about sensitivity to removal of individual input features in the context of ternary complexes as described in this study, we performed two additional sets of experiments.

First, as requested by the Reviewer, we experimented with two strategies to remove features during testing, namely: setting them to zero and randomization. To analyze their effects, we compared the average descriptor distances of interacting surface vertices for each benchmark case as this reflects how complementary the surfaces are according to MaSIF. If we zero out each feature at test time (Figure R1), the effects seem to be negligible for the chemical input features, whereas the removal of the geometric features seemingly leads to better (lower) descriptor distances. This is a result of the fact that zero curvature (flat surfaces on both sides) implies perfect shape complementarity. Presumably, the effect is less pronounced for the chemical features as zero charges can be considered less beneficial than a positive/negative pair, for example.

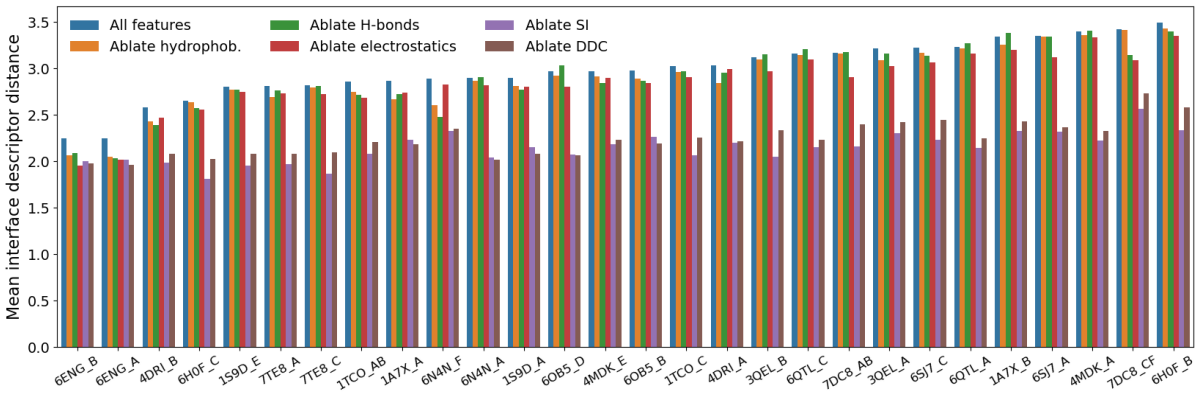

Figure R1: Feature ablation with zeros.

To randomize features, we drew random values uniformly from their input domain (Figure R2). We repeated this procedure three times for each target and its known binder with different random seeds, resulting in nine combinations. Here, we plot the same quantity as above but visualized as boxplots to show the variability across the nine repetitions. Again, we noticed that feature removal at test time has different effects on different feature types. For instance, randomization of distance-dependent curvature (DDC) leads to large variance because this feature usually tends to be rather uniform, which makes the model output very sensitive to random fluctuations it has not been trained on.

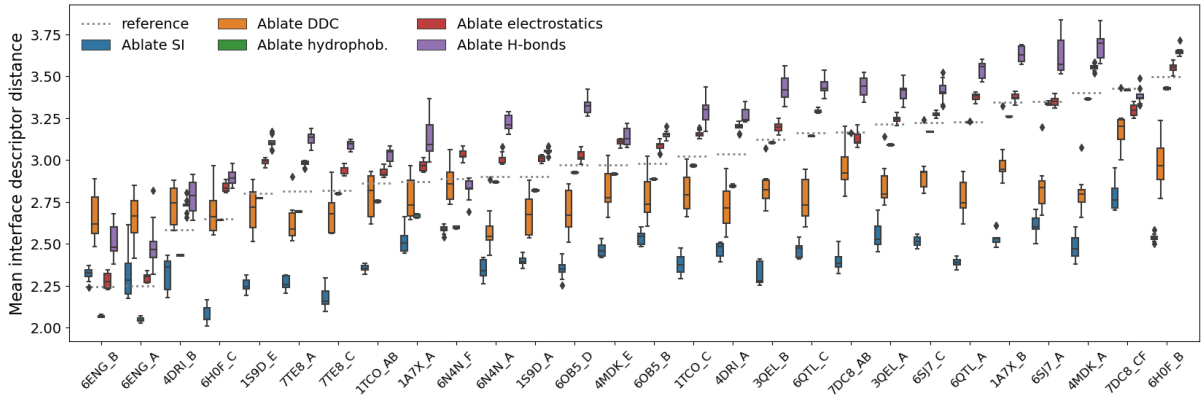

Figure R2: Feature ablation with randomization.

Overall, in both attempts of removing features at test time, it was not possible to compare output effects in an unbiased way because we cannot guarantee equal magnitudes of the input perturbations.

Based on this observation, we decided that a more fair comparison requires model retraining with each feature ablated (Figure R3), which is more in line with how such experiments are conducted in the machine learning literature. Specifically, we retrained five models on the original training set in each of which one of the input features was ablated. We then identified interacting patches at the interface of our 28 benchmark cases (center points within 1 Å of each other) and computed their descriptor distances. We also selected a random binder patch away from the interface for each interacting patch on the target surface as negative examples. Given the descriptor distances of these sets of positive (interacting) and negative (non-interacting) pairs, we computed the area under the receiver operating characteristic (ROC-AUC) to compare different models (see newly added Supplementary Fig. 4 that is also reported below). Our results indicate that MaSIF relies slightly less on the shape index feature to discern native interactions in the 28 benchmark cases, whereas the median ROC-AUC for the remaining four features is consistently lower than the baseline model with all features. However, the absolute effect of removing single features is substantially lower than the combined effect of removing all geometric or chemical features as performed in Ref. [8] as there is a certain degree of redundancy: shape index and distance-dependent curvature measure surface curvature, while hydrophobicity, electrostatics, and hydrogen bonding potential all describe the chemical complementarity.

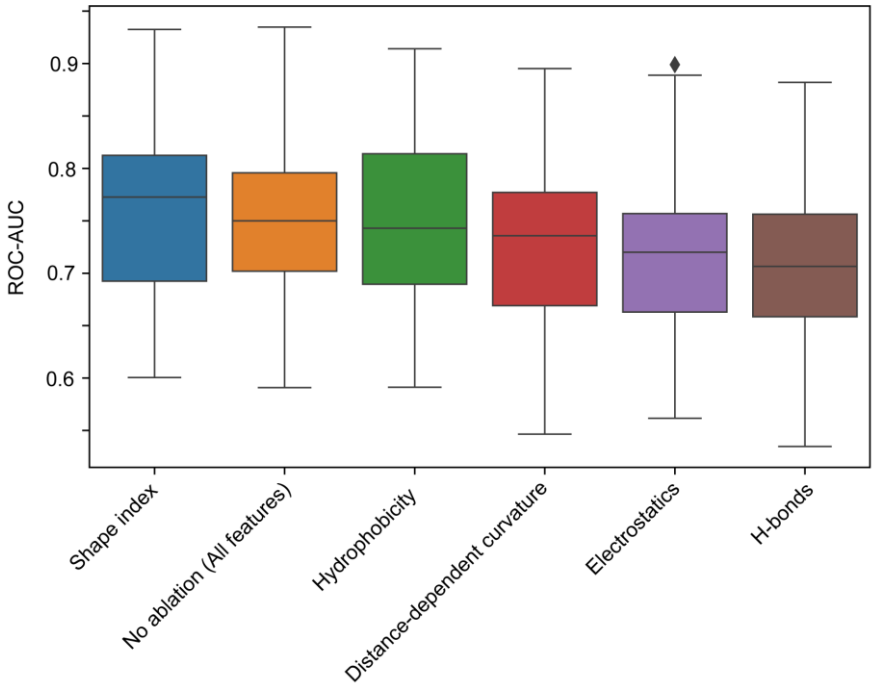

Figure R3: Feature ablation with model retraining.

[8] Gainza, Pablo, et al. "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning." Nature Methods 17.2 (2020): 184-192.

1.7 Following the *in silico* screening of the three ligand-dependent protein binders, the authors selected ~2,000 top ranking seeds for each target and screened them by yeast display. Following two rounds of FACS enrichment, the isolated clones were sequenced and the authors selected one binder for each test case to characterize more deeply. What happened to the other 1,999 seeds? Were the selected seeds overwhelmingly enriched during yeast screening? Were other high-ranking seeds enriched? Based on these observations, how many seeds would need to be screened to identify binders for most targets? Furthermore, this approach assumes that 2k randomly selected seeds would provide no enrichment for these same targets. Given the modest enrichment values reported in Table S2 and the unique properties of the ligands used for these test cases, it would be valuable to show experimentally that this assumption is true. For most binary complexes, I would not expect that 2k random proteins would identify a neosurface binder; however, for highly lipophilic ligands like progesterone, I would not be surprised to see similar enrichments from a random protein sampling.

R: As highlighted in Supplementary Table 1, only 100-120 seeds were selected and tested per target complex and not 2000 seeds (second column from the right, in brackets). However, each of these seeds were grafted on multiple recipient scaffold proteins, which led to ~2000 protein designs per target. We clarified this point in the main section of the manuscript: “Top-ranking seeds were selected (~100-120 for each target), refined, and grafted onto multiple recipient scaffolds, and approximately 2000 final designs per target complex were selected with computational filters”. As scaffold proteins may not fold properly or may bring various additional contacts with the target complex, we ensured to test multiple designs sharing the same seed. Thanks to the computational selection process, screening of these ~100 binding seeds (2000 designs if scaffolds are taken into account) sufficed to find successful binders while random binder screening typically requires library sizes in the order of 10^6 to 10^{11} [R1, R2, R3], not even taking into account the added complexity of ternary complex design. The log-enrichments of the binders shown in supplementary table S2 (30 to 200-fold enrichments in binding versus non-binding population) are in the expected range for binder screening with yeast display. Of course, these numbers are a reference and could vary according to the selected FACS gating and the binder affinity.

Moreover, the absence of binding with the native scaffold (no grafted seed) and the hotspot mutation on Fig. 3C, as well as the small molecule analogs on Fig. 3D, supports the necessity of the protein design steps in our pipeline.

To further support the absence of random selection, we computationally optimized the unsuccessful binders and re-screened the libraries, and highlighted 13 new binders from a variety of seeds (12 binders for PDF1:Actinonin and 1 for Bcl2:Venetoclax). Many designs from the first round of selection were predicted as unfolded by AlphaFold2 and were therefore optimized with LigandMPNN (See newly added Supplementary Fig. 19-22).

[R1] Almagro, J. C., Pedraza-Escalona, M., Arrieta, H. I. & Pérez-Tapia, S. M. Phage Display Libraries for Antibody Therapeutic Discovery and Development. Antibodies 8, 44 (2019).

[R2] Bashir, S. & Paeshuyse, J. Construction of antibody phage libraries and their application in veterinary immunovirology. Antibodies 9, 21 (2020).

[R3] Fierle, J. K. et al. A cell-based phenotypic library selection and screening approach for the de novo discovery of novel functional chimeric antigen receptors. Scientific Reports 12, (2022).

1.8 To validate functionally the identified CIDs, the authors use three different biochemical and cell-based assays, but only test one binder in each assay (Fig. 5). It would be much more informative if all three CIPs were tested in two functional assays, one extracellular and one intracellular (Fig. 5d,g). At present, the individual assays developed for each complex create uncertainty regarding the functional effects of the proteins in different systems.

R: We repeated each CID system with another type of assay: DBVen1619_2 in a intracellular split nanoLuciferase assay, DBAct553_1 in a cell-free expression system and DBPro1156_2 in an extracellular split nanoLuciferase assay. Unfortunately, DB3 scFv cannot be used intracellularly as it contains disulfide bonds which are required for proper folding, and therefore need to be located in an oxidizing environment. Supplementary figure 16 shows these additional assays. As a further demonstration of functionality in a different system, we incorporated our DBVen1619:Bcl2 system in a chimeric antigen receptor (CAR) T cell and successfully performed a drug-inducible tumor killing assay (see updated Fig. 5). This new assay holds promise for translational applications such as the establishment of drug-controlled and safer cell therapy. By showing 2 or 3 cell-based assays per CID, we demonstrated the functionality and generalizability of the designed proteins to various systems.

1.9 Regarding the first functional assay (Fig. 5a-c), the authors report an IC50 of 1.2 µM; however, the ternary complex KD reported for this complex is 18 nM. Since this is a cell-free system, why does the functional readout show a shift of nearly two orders of magnitude?

R: The EC_{50} shown in Fig. 5C represents the binding of the small molecule to the ternary complexes, while the K_D reported by SPR in Fig. 4B represents the binding of the designed binder to the progesterone-bound DB3. Moreover, the cell-free system contains 100 nM T7RNAP-fused DBPro1156_2, so we need >10-fold higher concentration of the small molecule to saturate all the binding sites available. Moreover, cell-free expression systems are complex environments where linker lengths and general steric orientation of the various components could also contribute to an apparently higher K_D value.

Referee #2:

2.1 This manuscript reports a computational method for designing protein surfaces for binding to neosurfaces. A neosurface is a protein-ligand interface where small molecules serve as 'molecular glue' that facilitates protein-protein binding. The computational methodology is closely related to that of a study published last year (ref. 9). The key difference is that in ref. 9, protein-protein interfaces are considered without ligand modification. The protein-ligand interface is featurized with properties including hydrophobicity and electrostatics. A fingerprint is built from these features using a neural network. The neosurface fingerprint is then input to the design/selection pipeline to generate a protein binding partner, where complementary interfaces are found by searching a library of protein surface fingerprints.

This technique is validated on a small number of known tertiary complexes and then used to design proteins that target three neosurfaces. The in-silico predictions are experimentally validated in a series of experiments including binding, mutagenesis and crystallography that characterize the affinity and structure of the binding complex and identify beneficial mutations. It is found that the binding is specific and targets the predicted neosurface interface. It is also shown that the complexes studied here can be employed in cellular systems and used as, for example, biosensors. The work is put in context with highly-visible protein structure prediction models in supplemental materials (see Figs S2 and S16).

The protein design predicted by the model is extensively validated through a series of wet-lab experiments and all presented data are robust and of high-quality. The presentation and conclusions are also clear and well-supported.

R: We thank the reviewer for accurately portraying our work and the encouraging assessment of its quality.

2.2 However, while the computational approach is clearly valid and useful, it can be considered an added feature to the method presented by the authors in Refs. 8,9 for treating the case of neosurfaces. It is therefore unclear if it represents a significant advancement in the field for publication in Nature.

R: We thank the reviewer for bringing up this point, however we would like to highlight that the purpose of our study is the generalization capabilities of a molecular representation and the experimental validation of the design approach in novel application domains. Our simple but nontrivial changes allowed us to apply a method that has already proven its usefulness for the design of protein-protein interactions directly to ternary complexes which opens the door to many applications with high practical relevance. Arguably the most interesting methodological finding of the paper is precisely the fact that the surface-centric design approach can be seamlessly transferred to ternary complexes without heavy re-engineering and specialized models. This level of generalizability is uncommon and highly non-trivial in deep learning approaches and was only possible because of the conceptual advances leading to a right level of abstraction of representation in molecular surfaces. Thanks to its generalization capability, the same principles could potentially be extended to other challenging design tasks such as molecular glues, which are highly-valuable in therapeutic applications.

2.3 The following suggestions may strengthen the manuscript in a future submission: Very recently, AlphaFold3 has been released which incorporates small molecules into their model [https://doi.org/10.1038/s41586-024-07487-w]. This represents an advancement over AlphaFold2, which is used in Fig. S16. How would AlphaFold3 perform or enhance the task of targeting protein neo-surfaces?

R: The published AlphaFold3 results are impressive and we foresee a great impact on the design of small molecule-induced protein interactions, similar to how AlphaFold2 revolutionized monomeric protein design. However, the current release of the software only includes a limited list of supported ligands and is not compatible with our test cases. Due to its limited availability, we could not extensively test the model or incorporate it in our design pipeline. We plan to study its capabilities in future research.

To support how our methodology fits into the context of the availability of the AlphaFold3 server, we added the following discussion: “Novel tools like AlphaFold3 [63] demonstrated good performances for ligand:protein complex prediction, however the limited scope of use posed non-negligible hurdles for further advancement in the field of drug/protein design. We foresee that approaches like surface fingerprinting could represent a suitable alternative for targeting neosurfaces”.

[63] Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).

2.4 The current computational method appears applicable to given protein-ligand interfaces, but leaves open the question of the design or selection of the ligand or ‘molecular glue’ itself. This would seem to introduce (another) large search space. How feasible is this as an extension of the present work?

R: This could indeed be an interesting future avenue. At the technical level, we can already profile different chemical entities bound to their receptors and score the propensity of the formed neosurfaces to bind to other target proteins existent in naturally occurring proteomes. Therefore, we could foresee an easy way to couple small-molecule generation pipelines where new ligands can be generated in the context of the receptor proteins and the emerging neosurface scored for the propensity of forming a productive interface. MaSIF metrics could be incorporated as filtering criteria or to guide the sampling in generative pipelines.

2.5 The computational approach of this work utilizes physical and chemical high-level ‘expertly chosen’ features such as Poisson-Boltzmann electrostatics and hydrophobicity scales when compared with recent approaches that may use only atoms/coordinates or sequences to represent the inputs. Does this choice allow the use of neural networks of lower complexity in the present work? Does the use of physical modeling such as electrostatics calculations limit the throughput of the technique?

R: Yes, this observation is correct, in our approach we try to capture the biochemical intuition that protein and ligand interactions are determined by chemical and geometrical features and hence the choice for a “feature-enriched” surface representation. The rich input features allow us to use a small model with only ~66K trainable parameters (compared to 93M in AlphaFold for example), and are the reason our architecture could generalize to neosurfaces without retraining, which might be impossible with neural networks trained on raw atom coordinates and types. Most of these models do not support some of the chemical elements found in small molecules. Additionally, the low complexity of the model together with its good performance indicate a favorable bias-variance tradeoff whereas more complex models can be assumed to exhibit higher predictive variance when confronted with out-of-distribution inputs.

While the assumption about the computational bottleneck is equally correct (see for example Fig. 2 in this previous work from our lab [R4]), the slow surface triangulation and feature calculations are rarely a limitation in practice as they can be precomputed and stored. Furthermore, all precomputation steps run on CPU hardware and can be massively parallelized without requiring expensive GPUs.

We refer the reviewer to a previous work from our lab [R4] in which we showed a variant of MaSIF architecture (dMaSIF) that learns surface features directly from the atomic coordinates in an end-to-end fashion. In particular, we showed that the ‘expertly chosen’ pre-computed chemical features used in MaSIF could be learned from scratch, using a bigger neural network with more parameters. While there may be multiple advantages to such end-to-end differentiable architectures, importantly our experiments did not show significant advantage of learned features over pre-computed ones, while coming at the expense of more parameters and hence potentially poorer generalization and interpretability. This was one of the main reasons why in this work we opted for a simpler MaSIF architecture.

[R4] Sverrisson, Freyr, et al. "Fast end-to-end learning on protein surfaces." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

2.6 Note that Ref. 29 has been published [DOI: 10.1126/science.adl2528].

R: We thank the reviewer for this observation. We corrected the reference in the revised version.

Referee #3:

3.1 In the manuscript by Marchand et al., the authors build on their previously published model, MaSIF, to develop MaSIF-neosurf, expanding the target domain of protein binder design to protein-ligand complexes. The authors first demonstrate that the new method can enrich 14 ligand-dependent complexes over a library of decoy complexes (Fig. 1). They then use this method to design de novo protein binders targeting three neosurfaces (Fig. 2). Several initial and optimized designs have been characterized through binding, biochemical assays, and structural validations (Figs. 3 and 4). The optimized designs have also been tested in split complementation assays in cell-free and mammalian cell systems (Fig. 5).

This paper is well-done and represents a collaborative effort combining deep learning methods, yeast display, biochemical and cellular evaluations, and detailed structural studies to explore potential use cases for creating binders to protein-ligand complexes, an outstanding challenge in protein design. Although there is a question as to how efficient this method is compared to experimental methods, such as AbCIDs (Hill et al., Nat. Chem. Biol. (2018) 14:112-117), this study is a first step toward expanding deep learning methods to design selective binders for protein-ligand complexes. The implications of this study are particularly useful for developing molecular glues in synthetic biology, and, of course, protein design.

I could not detect any significant flaws in the work. Nevertheless, I have a few comments that I believe will improve the strength of the main takeaways for a broad audience. Regarding comments and suggestions, I have four.

R: We thank the reviewer for the positive assessment of our work and the constructive comments.

3.2 Comparison to experimental methods: Deep learning methods are undoubtedly improving, but how do they compare to existing experimental methods in terms of the success rate of both pipelines, the best potency straight out of initial design, preferences for certain targets, and whether deep learning methods have surpassed the efficiency of experimental methods in this task? What are the pros and cons of this method compared to experimental methods such as AbCIDs? This comparison and discussion would be informative for a general audience to better understand these methods and choose proper methods for their targets.

R: Experimental methods, such as AbCIDs and other antibody screening platforms are agnostic to where and how these proteins engage their respective target. With computational design methods, the binding site is pre-determined which is an important advantage of these approaches. Most experimental methods need to screen in the range of 10^6-10^9 variants to obtain a few binders, while the number of tested designs in our work is 3 to 6 orders of magnitude lower. Methods like AbCIDs are mostly focusing on antibody-based proteins, which are of course an important format in biotechnology, while deep learning tools give a broader range of protein folds and size frequently with high thermal stability. On the other hand, we could observe that most deep learning tools were successful on hydrophobic and rigid interfaces, while antibodies demonstrated broader versatility and are still a difficult format to design computationally. We added a few sentences in the discussion to clarify those points to a broader audience: “Experimental methods like antibody screening platforms [29] are agnostic to where and how these proteins engage their respective target. Deep learning tools, like the one presented here, can control these parameters and offer more modalities in terms of protein shapes, folds, sizes, and thermal stability. However, some challenges remain as state-of-the-art deep learning-based structure validation methods like RoseTTAFold[30] failed to predict our validated complexes (Supplementary Fig. 2) [...]. Most deep learning methods, including ours, exhibit superior performance on hydrophobic patches, while significant challenges persist in accurately modeling polar interfaces.[9,43]”

[9] Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. Nature 617, 176–184 (2023).

[29] Hill, Z. B., Martinko, A. J., Nguyen, D. P. & Wells, J. A. Human antibody-based chemically induced dimerizers for cell therapeutic applications. Nat. Chem. Biol. 14, 112–117 (2018).

[30] Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science eadl2528 (2024) doi:10.1126/science.adl2528.

[43] Cao, L. et al. Design of protein-binding proteins from the target structure alone. Nature 605, 551–560 (2022).

3.3 Generalizability and success rate discussion: How potent should ligand-protein complexes be? What are the lower and upper potency boundaries within which this method can be applied?

R: The ligands we used have affinities ranging from ~10 pM (Venetoclax) to ~500 nM (Actinonin) for their respective target protein. We can safely assume that there is no lower bound for the ligand dissociation constant as higher affinities are desired to ensure a proper ternary complex formation. An upper bound is harder to estimate as the designed protein might provide a cooperative effect upon binding and virtually “trap” the ligand. We could assume that as long as the target protein can be saturated with bound ligands (>10-fold excess above K_D), any ligand-complex could be applied to this method. The main limiting factor of the upper boundary is then the solubility of the compound. In a way our approach is rather agnostic to the affinity of the ligand-receptor interaction, provided that the structure of the holo complex is accurate, nevertheless we can easily assume that high affinity ligand-receptors are more accessible for experimental characterization for the reasons mentioned above.

3.4 Public availability of site saturation mutagenesis data: The authors should consider making the site-saturation mutagenesis data publicly available in .csv or other machine-readable formats. This information could be useful as a benchmark in the field for similar design strategies and also for improvements to the current method.

R: We agree with the idea suggested by the reviewer as we are actively committed to transparency and open source research. We will provide all data used to plot each main and supplementary figure, including the SSM, in a dedicated repository. A statement and a link in the “Data availability” section was be added: “Data used to generate the figures and supplementary materials, as well as the relevant plasmid maps, were deposited on Zenodo (http://doi.org/10.5281/zenodo.13737922)”

3.5 Consistency in chemical drawings: The chemical drawing style is not consistent throughout the manuscript. The chemical drawing style in Fig. 3a is different from that in Fig. 3d for Bcl2. The same issue is present in Fig. S6a,b.

R: We thank the reviewer for this observation. We have now corrected the consistency of the chemical drawings in the mentioned figures.

Authors' response to referees' comments (version 2)"

Referee #1:

In the resubmitted manuscript, the authors have done a nice job of addressing all our comments. We were especially happy to see the added discussion around a protein-ligand complex that was unsuccessful with their approach (JQ1-BRD4; SI Fig. 23 and Discussion), which we believe provides a more balanced presentation of their work. Additional improvements include the enlarged set of decoys used in their benchmarking experiments, which strengthens the MaSIF validation; the use of LigandMPNN to improve the success of the seed grafting process; and the added functional validation experiments, in particular the CID-CAR experiment, to further validate the success and utility of MaSIF-neosurf. Given these modifications, we support publication of the manuscript in its revised form.

Remarks on code availability: The code deposited at https://github.com/LPDI-EPFL/masif-neosurf was reviewed and found to be complete and well documented, including instructions for running the main analyses provided in the manuscript.

R: We thank the reviewer for the positive assessment of our work.

Referee #2:

I thank the authors for their detailed response. They have addressed my concerns and I suggest no further changes.

I agree that the level of generalizability that this method displays is indeed uncommon. I look forward to future work which may more fully address comparisons between this method and emerging models in the literature, where generalizability serves as a key 'metric' for consideration.

Remarks on code availability: Please see my comments on the original submission

R: We thank the reviewer for the good appreciation of our study.

Referee #3:

The authors have addressed all my comments. I have no further comments and would recommend publishing the manuscript as it is.

R: We thank the reviewer for supporting our manuscript.

Referee #4:

The experiments with the CAR-T cells are well performed. There is some baseline killing in the absence of drug (leakiness) that should be commented on (Fig. 5k). In addition, while this provides proof of concept for the CAR dimerization, the data in Fig. 5i as well as Supp. 18 shows that the split CAR approach is likely less potent than a traditional CAR, which is not surprising and should be commented on. It should be noted that drug dimerization of CAR is not a novel approach and has already been in the clinic, but that does not appear to be the point of this manuscript so this should be acceptable as POC.

R: We thank the reviewer for the constructive comments. We commented our results following the reviewer suggestion: "Despite the observed desired effect, residual tumor killing in absence of the drug and a slightly lower potency compared to that of the 2G-CAR was observed."