# Artios – Supplementary Material

Ligo Biosciences Team

October 2026

## Contents

**1 Data Pipeline** **3**
1.1 Weighted PDB Dataset . . . 3
1.1.1 Filtering . . . 3
1.1.2 Sampling . . . 3
1.2 MGnify Distillation Dataset . . . 4
1.2.1 Fragmentation and Cleanup . . . 4
1.2.2 Structural Clustering and Sampling . . . 7
1.2.3 Pocket Detection . . . 7
1.3 Cropping . . . 8
1.4 Motif Construction . . . 9

**2 Models** **11**
2.1 ArtiosBackbone . . . 11
2.1.1 Architecture . . . 11
2.1.2 Featurisation . . . 11
2.1.3 Training Details . . . 12
2.1.4 Inference . . . 12
2.2 ArtiosPotts . . . 13
2.2.1 Architecture . . . 13
2.2.2 Featurisation . . . 13
2.2.3 Training Details . . . 13
2.2.4 Inference . . . 14

**3 Benchmarking** **15**
3.1 Atom Motif Enzyme Benchmark . . . 15
3.2 Sequence Design Benchmarking . . . 16

**4 Design Campaigns** **17**
4.1 Luciferase . . . 17
4.1.1 Design Pipeline . . . 17
4.1.2 Analysis of Designs . . . 18
4.1.3 In vitro Evaluation . . . 18
4.2 Esterase and Cutinase . . . 19
4.2.1 Esterase Design Pipeline . . . 19
4.2.2 Cutinase Design Pipeline . . . 20
4.2.3 In vitro Evaluation . . . 20
4.3 Phosphodiesterase . . . 21
4.3.1 Design Pipeline . . . 21
4.3.2 In vitro Evaluation . . . 22

**5 Experimental Methods** **22**
5.1 Gene Preparation and Protein Preparation . . . . . 22
5.2 Data Processing and Analysis . . . . . 23

**A Reference Tables** **23**
A.1 Additional Crystallization Aids . . . . . 23
A.2 *SeqVal* Sequence Design Benchmark Full Data . . . . . 24

# 1 Data Pipeline

This section describes the data sources and preprocessing pipeline used across Artios training. We first describe the weighted PDB dataset, which is used for ArtiosPotts and PDB-based ArtiosBackbone training, and then the MGnify-derived distillation dataset used in ArtiosBackbone training experiments. The overall PDB preparation procedure follows the data-processing strategy used for AlphaFold3 [1]: structures are read from mmCIF files, standardized using Chemical Component Dictionary definitions, expanded to the first biological assembly, and filtered before token-level cropping. We use the AF3 training cutoff of September 30, 2021 for PDB-derived training examples.

## 1.1 Weighted PDB Dataset

### 1.1.1 Filtering

For PDB examples, we parse atomic coordinates, polymer sequences, entity annotations, covalent bonds, and biological assembly metadata from the deposited mmCIF files. Alternate locations are resolved during parsing, selenomethionine residues are converted to methionine, arginine terminal nitrogen naming ambiguities are corrected, waters and hydrogens are removed, and atom sets are reconciled against CCD templates. Missing atoms and unresolved residues are restored as unresolved template atoms where possible, allowing the model to receive consistent chemical graphs while masking coordinates that were not experimentally observed.

The weighted PDB dataset follows the target and bioassembly filtering strategy described for the AlphaFold3 weighted PDB dataset [1], using the same release-date cutoff of 2021-09-30 and a 9 Å resolution threshold. The main Artios-specific additions and differences are:

- Polymer chains containing any unknown residue types are removed.
- The crystallization-aid and artifact-ligand cleanup is extended beyond the AF3 crystallization-aid list (Table 6), and training rows involving ligands classified as excluded ligands are removed.
- PDB entries in membrane, ribosome, and fibril exclusion lists are removed, and training rows with more than 4000 tokens are excluded.
- Unbound metal ions with fewer than three non-metal heavy-atom contacts within 3 Å are removed during bioassembly cleanup.
- CCD reconciliation follows the AF3 atom-name and leaving-atom cleanup, with additional handling for covalent leaving atoms that are not explicitly flagged in the CCD.
- Clashing-chain removal ignores covalently connected chain pairs and uses a 1.6 Å distance threshold.
- When reducing large assemblies, covalently connected polymer–ligand groups are treated as single molecules so that covalent ligands are not removed from their associated polymers.
- Asymmetric inter-chain polymer–ligand bonds that are not supported across symmetry-related copies are removed from the bond graph.
- Glycan residues are removed after cropping using the AF3 glycan CCD list.

### 1.1.2 Sampling

As in AlphaFold3, PDB training rows are constructed at the chain and chain-interface level. Chain rows represent individual molecular entities, while interface rows represent pairs of entities in contact.

Each row is assigned a cluster identifier, and weighted sampling is used to avoid over-representing highly redundant structures. The row weight is proportional to

$$\beta \cdot \frac{\alpha_{protein} n_{protein} + \alpha_{nucleic} n_{nucleic} + \alpha_{ligand} n_{ligand}}{N_{cluster} N_{entity}}, \tag{1}$$

where $\beta = 0.5$ for single-chain rows and $\beta = 1.0$ for interface rows, $n_{protein}$, $n_{nucleic}$, and $n_{ligand}$ are the numbers of entities of each type in the row, $N_{cluster}$ is the size of the row's cluster, and $N_{entity}$ accounts for repeated entity combinations within the same PDB entry. In the current Artios training configuration, the entity weights are $\alpha_{protein} = 1$, $\alpha_{nucleic} = 0$, and $\alpha_{ligand} = 1$.

We additionally cap the aggregate weight of one-protein-side interface rows to prevent heavily studied protein families from dominating the ligand and nucleic-acid contexts. This is needed because distinct ligands can create many separate interface clusters for the same protein family, even after the base cluster weighting above. For each protein cluster $k$, let $W_k$ be the total base sampling weight of its protein–ligand and protein–nucleic-acid interface rows. We set $\tau$ to the 90th percentile of the $\{W_k\}$ distribution and scale all such interface rows from clusters with $W_k > \tau$ by $\tau/W_k$. Monomer rows and protein–protein interface rows keep their base weights. In a 4 Å-resolution filtered weighted PDB index used for this diagnostic, this reduces the aggregate interface weight of representative kinase, HIV protease, and immunoglobulin/Fab clusters from 1620, 726, and 837 to $\tau = 16$, corresponding to roughly 101-, 45-, and 52-fold reductions, respectively. This cap was introduced to reduce mode-collapse behaviour observed under uncapped sampling.

## 1.2 MGnify Distillation Dataset

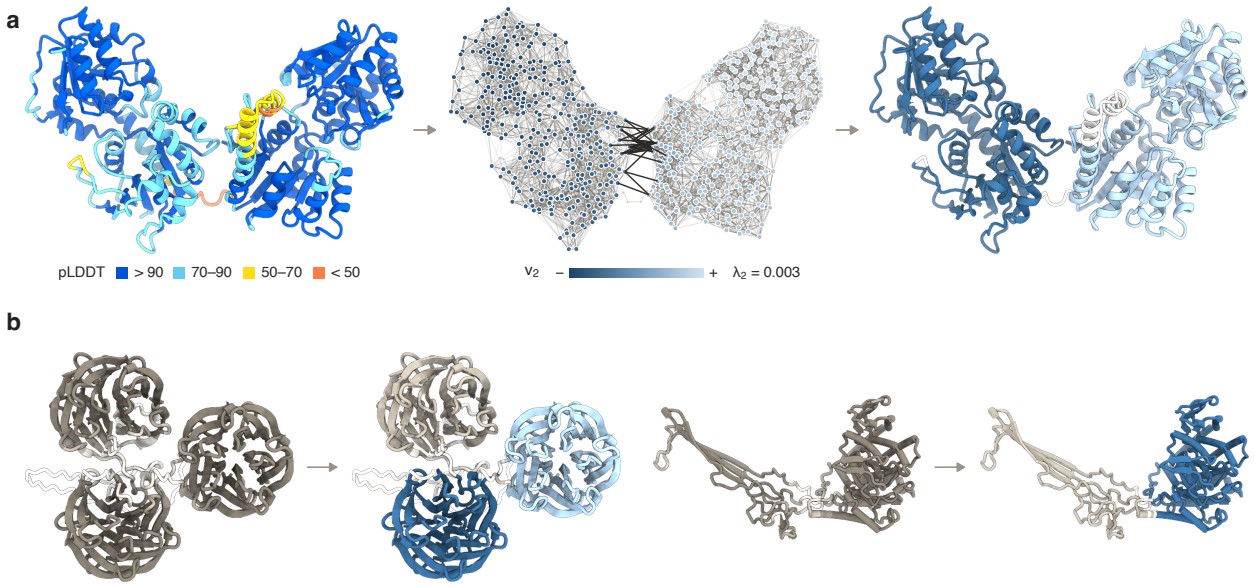
The MGnify distillation set is constructed from predicted structures of metagenomic protein sequences released through the OpenFold datasets portal and described in the OpenFold3-preview2 technical report [2, 3]. These examples provide broad coverage of natural protein sequence space, but unlike experimentally determined PDB structures, the predictions often contain low-confidence termini, flexible linkers, and multi-domain chains whose relative domain placement is not necessarily meaningful. We therefore process the dataset at the level of compact, high-confidence fragments rather than treating each predicted full-length chain as a single training unit.

### 1.2.1 Fragmentation and Cleanup

We first standardize the predicted structures by removing hydrogens, repairing residue naming where needed, converting selenomethionine to methionine, and annotating residues with CCD reference information. Low-confidence regions are then removed using the predicted confidence scores: residues with pLDDT below 65 are marked as unusable before fragment construction.

A central challenge in using predicted MGnify structures is that confidence filtering alone does not define the right training unit. Low pLDDT removes obvious disorder, but high-confidence linkers can still connect domains whose relative placement is not a useful generative target. We found that global metrics such as full-chain length and radius of gyration were too coarse for this decision, while domain-level annotations such as TED, The Encyclopedia of Domains [4], could split well-packed proteins more aggressively than desired for generative training. We therefore annotate compact fragments with a split-then-merge graph procedure (Algorithm 1): spectral bisection proposes weakly connected boundaries in a spatial graph over confident $C_\alpha$ atoms, and spatial contact clustering rejoins fragments that remain part of a single compact unit. This removes disordered tails and high-confidence linkers while preserving ordered domains that would be discarded by global filters on the full predicted chain (Supplementary Figure 1).

The retained fragments are then filtered by fragment-level quality metrics (Table 1) designed to remove examples with weak structural signal for generative training. Mean pLDDT and fragment length enforce confident, sufficiently large training units; radius of gyration and corrected packing efficiency remove expanded or poorly packed fragments; SSE entropy, repeat score, and corrected contact order discourage repetitive or filamentous geometries; and the hydrophobic burial and SASA




**Supplementary Figure 1:** Fragment generation on full-length predicted structures. (a) A two-domain prediction coloured by pLDDT (left). Confident residues form a $C_\alpha$ graph whose Fiedler vector $\mathbf{v}_2$ separates the domains; the cut edges (black) lie at the inter-domain neck (centre). The two domains are returned as fragments (right). (b) Two further predictions before and after fragmentation. Blue fragments are kept; pale residues are trimmed by the pLDDT or spectral steps, and grey fragments fail the quality filters (Table 1).

filters remove fragments with exposed hydrophobic cores. Here, *corrected* indicates that the corresponding metric is normalized by fragment length, making the threshold less sensitive to protein size. Fragments failing any of these checks are removed before pocket annotation and clustering; Supplementary Figure 2 shows an example rejected by each filter.

**Algorithm 1** Fragment annotation

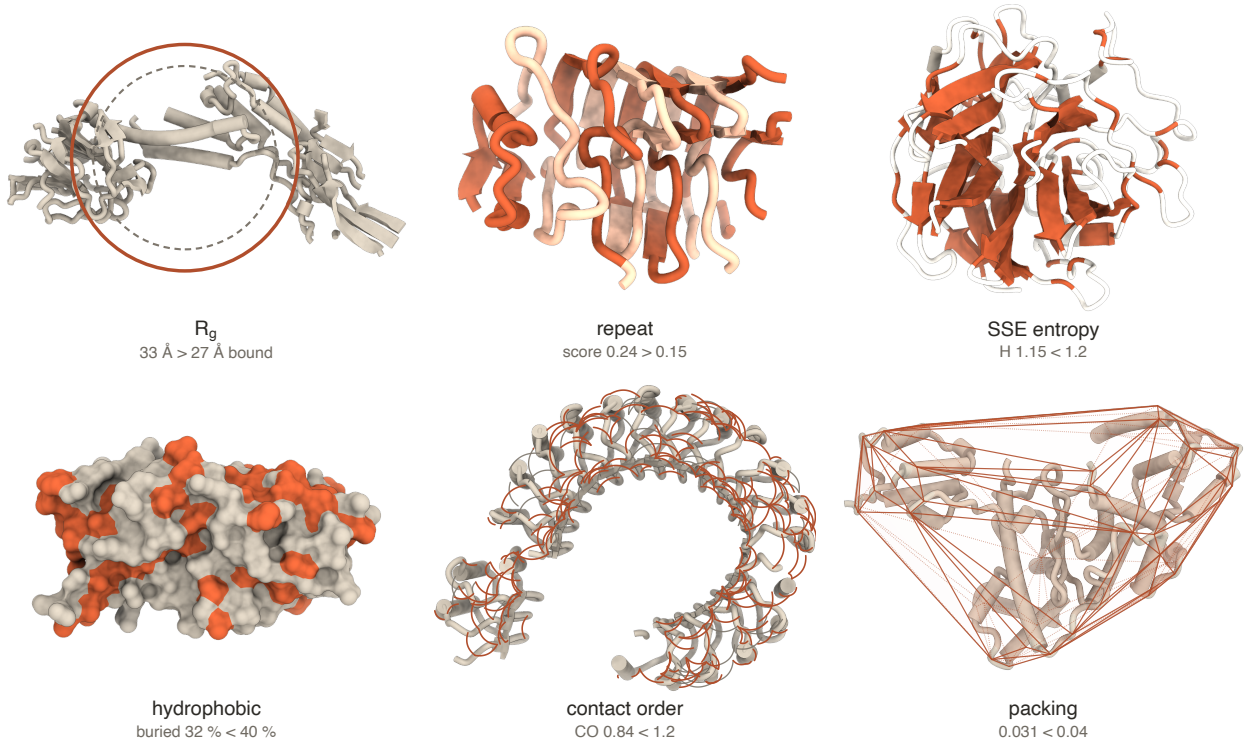
**def** FragmentPredictedStructure($S$):

*# $S$: predicted protein structure with coordinates and pLDDT values.*

mark residues with pLDDT $< 65$ as unusable
$V =$ remaining $C_\alpha$ atoms
build a 15-nearest-neighbour graph on $V$ with edge weights $w_{ij} = \exp(-d_{ij}^2/(2 \cdot 8^2))$
SpectralSplit($V$)
assign contiguous usable residues to preliminary fragments
**for each** pair of preliminary fragments $(F_i, F_j)$:
    **if** at least 4 residues in $F_i$ have any atom within 4 Å of an atom in $F_j$: merge $F_i, F_j$
fill internal spectral gaps of at most 8 residues when all gap residues have pLDDT $\geq 65$
remove fragments failing any quality filter in Table 1
**return** fragment identifiers for the retained atoms

**def** SpectralSplit($U$):

**if** $|U| < 50$: **return**
compute $L_{\text{sym}} = I - D^{-1/2} W D^{-1/2}$ for the subgraph induced by $U$
let $(\lambda_2, \mathbf{v}_2)$ be the second eigenpair of $L_{\text{sym}}$
**if** $\lambda_2 > 0.05$: **return**
mark the 4 residues with smallest $|\mathbf{v}_{2,i}|$ as unusable boundary residues
SpectralSplit($\{i \in U : \mathbf{v}_{2,i} < 0\}$)
SpectralSplit($\{i \in U : \mathbf{v}_{2,i} \geq 0\}$)




**Supplementary Figure 2:** Fragment quality filters. Each panel shows a predicted fragment rejected by one filter (Table 1), with the responsible feature in red and the metric against its threshold below. $R_g$: radius of gyration (solid) against the length-dependent bound (dashed). Repeat: repeat units of a $\beta$-solenoid, alternately shaded. SSE entropy: strands and coil with almost no helix. Hydrophobic: exposed hydrophobic residues. Contact order: a sample of $C_\alpha$ contacts, red where shorter-range than required. Packing: the $C_\alpha$ convex hull.

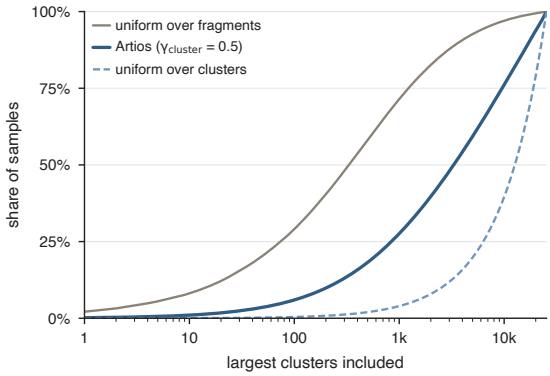
| Metric | Computation | Accepted range |
|---|---|---|
| Mean pLDDT | Average pLDDT over atoms in the fragment | $\geq 70$ |
| Fragment length | Number of unique residues in the fragment | $L \geq 150$ |
| Radius of gyration | Root-mean-square distance of fragment atoms from their centroid | $R_g \leq 2.4\,L^{0.4}$ |
| SSE entropy | Shannon entropy of the DSSP helix, sheet, and coil fractions | $H_{\text{SSE}} \geq 1.2$ |
| Repeat score | Maximum autocorrelation, at lag $\geq 15$, of the detrended mean $C_\alpha$ distance as a function of sequence separation | $\leq 0.15$ |
| Buried hydrophobic fraction | Fraction of hydrophobic residues with residue SASA $< 10$ Å$^2$ | $\geq 0.4$ |
| Hydrophobic SASA fraction | Hydrophobic-residue SASA divided by total residue SASA | $\leq 0.4$ |
| Corrected contact order | Mean sequence separation of non-local $C_\alpha$ contacts within 8 Å, divided by $\sqrt{L}$ | $\geq 1.2$ |
| Corrected packing efficiency | `L / convex_hull_volume(CA coords) * L**0.3` | $\geq 0.04$ |

**Table 1:** Fragment quality filters used before pocket detection for the MGnify distillation dataset. All bounds are inclusive. $L$ denotes fragment length.

### 1.2.2 Structural Clustering and Sampling

After fragmentation, structural redundancy is handled at the fragment level rather than the full-chain level. MGnify fragments are clustered by structural similarity using Foldseek as a scalable first pass. Because fast Foldseek clustering can separate sequence-divergent fragments that nevertheless share the same fold, we additionally audit cluster representatives with a TM-align-based structural comparison and merge representative clusters when both directional TM scores support the same fold-level assignment [5].

The resulting cluster identifiers are used during training to reduce over-representation of highly repeated natural folds. Rather than sampling fragments uniformly, which would heavily bias training toward the largest fold families, or sampling clusters uniformly, which would over-emphasize small clusters, we use a cluster-balanced sampler with an additional length-dependent reweighting (Supplementary Figure 3). For a fragment from a structural cluster of size $N$ and token length $L$, the per-row weight is proportional to $N^{\gamma_{\text{cluster}}-1}L^{\gamma_{\text{len}}}$. The current motif-scaffolding configuration uses a cluster-size exponent $\gamma_{\text{cluster}} = 0.5$, so the aggregate sampling mass of a cluster scales as $N^{0.5}$ before length reweighting, and a separate length exponent $\gamma_{\text{len}} = -1.0$. This length term shifts sampling toward shorter fragments, increasing the probability of drawing structures that fit within the 256-token crop used by the Artios motif-scaffolding model.




**Supplementary Figure 3:** Share of MGnify samples drawn from the largest clusters under three sampling schemes. Foldseek clusters before TM-align merging, multi-member only; length reweighting omitted.

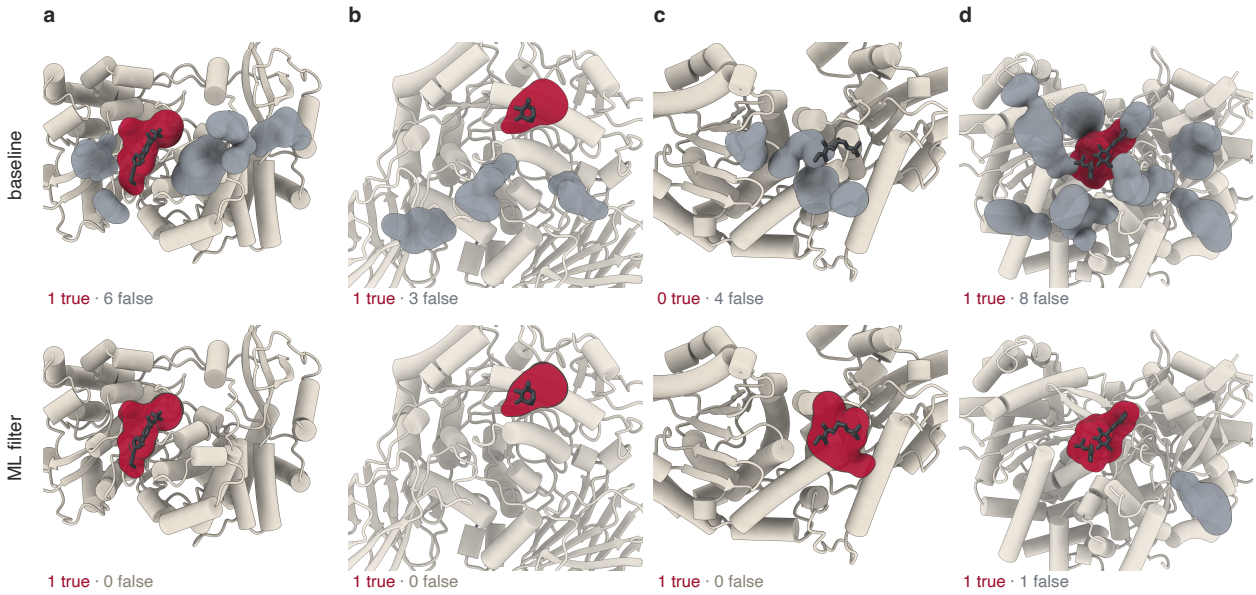
Training samples are drawn from this weighted fragment index before crop selection. The MGnify distillation training index contains 976,880 complete fragment examples assigned to 10,798 structural clusters. Fragment sizes have a median of 354 tokens, with an interquartile range of 289–445 tokens. In the broader MGnify fragment clustering analysis, 71.5% of multi-member fragments fall in the top 1,000 clusters [5].

### 1.2.3 Pocket Detection

Because the MGnify distillation structures do not contain bound ligands, we annotate putative ligand-binding pockets to make the dataset useful for motif scaffolding and pocket-centred cropping. Pocket detection is run independently on each retained fragment so that pockets are identified in the same compact structural units used for training.

Our detector is derived from the alpha-sphere pocket construction used in Pocketeer [6], which implements a simplified fpocket-style Delaunay-tessellation and alpha-sphere procedure [7], with an additional P2Rank-inspired machine-learning filter to suppress geometrically plausible but non-ligand-binding cavities. We first compute alpha spheres from the protein geometry, retain buried spheres in high-confidence regions, and cluster nearby buried spheres into candidate pockets. Each sphere is featurized using local geometry, atom density, and chemical-environment features, and a random-forest classifier scores whether it resembles a ligand-proximal sphere. Candidate pockets are then scored by a second pocket-level model that combines pocket geometry, aggregated sphere features, and sphere-score statistics. Compared with scoring the same geometric candidates by an earlier logistic model, this two-stage filter removes many false pockets at similar recall (Supplementary Figure 4).

For distillation training, pockets are required to pass density and confidence filters, and examples are preferentially restricted to fragments containing at least one valid pocket. These pocket annotations are used both for motif sampling and for pocket-centred cropping, allowing the backbone model to see ligand-free natural protein environments that are nevertheless enriched for plausible binding-site geometry.




**Supplementary Figure 4:** The learned filter removes plausible but non-binding cavities. Pockets kept from the same candidates by a geometric baseline (top; alpha-sphere pockets scored by a logistic model) and by the two-stage ML filter (bottom), on four enzymes held out from detector training: (a) 5HMZ, (b) 8RW3, (c) 1KOJ, (d) 5JTT. Red, pocket within 3 Å of a crystal ligand (dark sticks; removed before detection); grey, false pocket. Across 78 held-out structures (898 candidate pockets), precision rises from 0.33 to 0.63 at similar recall (0.84 and 0.81).

**Algorithm 2** Pocket annotation

**def** AnnotatePockets($F$):

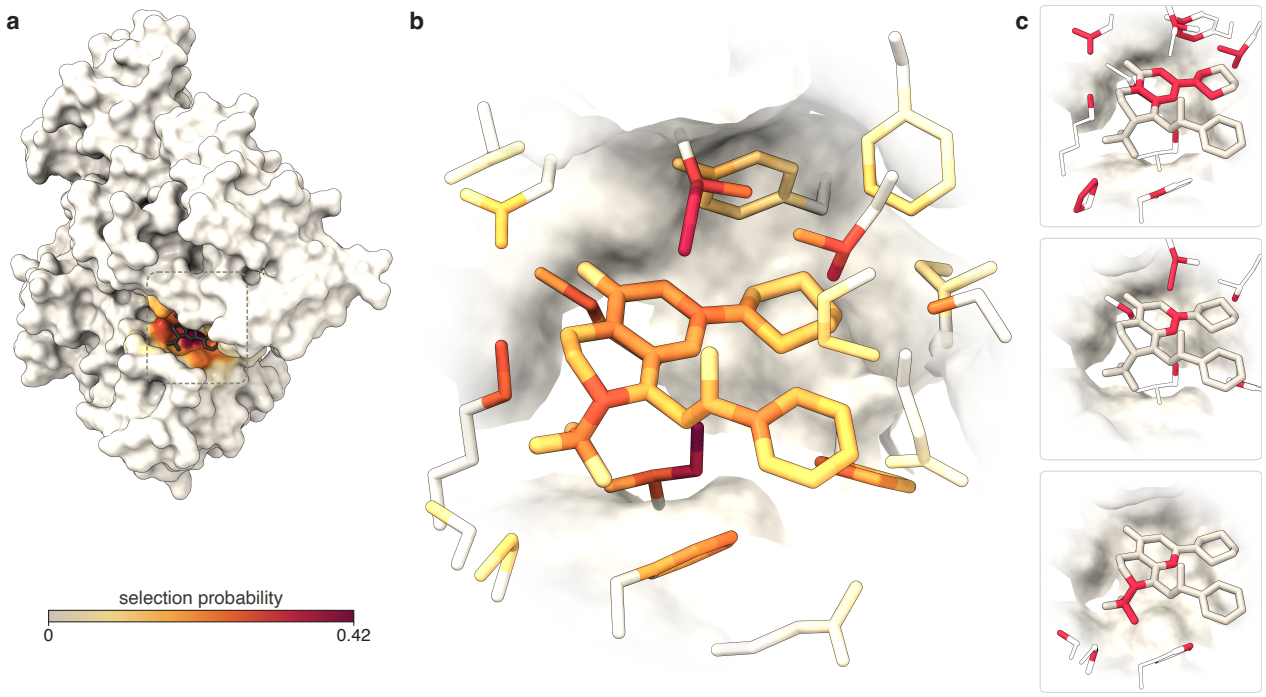
*# $F$: retained protein fragment with coordinates and pLDDT values.*

remove hydrogens, waters, and hetero atoms from $F$
$\mathcal{S}$ = empty Delaunay circumspheres with radii in $[3, 6]$ Å
mark $s \in \mathcal{S}$ as buried if its defining atoms have SASA $\leq 10$ Å$^2$
discard spheres whose defining atoms have pLDDT $< 70$
**for each** sphere $s$:
    $\phi_s$ = sphere geometry, protrusion, chemistry, density, and local context features
    $q_s$ = learned sphere random-forest probability from $\phi_s$
$\mathcal{C}$ = clusters of buried spheres with merge distance 1.75 Å and at least 35 spheres
$\mathcal{P} = \emptyset$
**for each** candidate cluster $C \in \mathcal{C}$:
    **if** local atom counts around Centroid($C$) are too low: **continue**
    $\psi_C$ = pocket geometry, aggregate sphere features, and summaries of $\{q_s : s \in C\}$
    **if** the learned pocket gradient-boosting score from $\psi_C$ is at least $\tau$: add $C$ to $\mathcal{P}$
assign each atom touched by a retained pocket sphere to a pocket identifier
**return** pocket identifiers for atoms in $F$

### 1.3 Cropping

Training examples are cropped at the token level after structure preparation. We use the contiguous, spatial, and spatial-interface cropping procedures described in the AlphaFold3 Supplement [1]. The weighted PDB dataset uses spatial crops and spatial-interface crops with equal probability; contiguous and pocket-centred crops are disabled.

For the MGnify distillation dataset, we extend the same reference-token idea used by spatial




**Supplementary Figure 5:** Per-atom motif selection probability on Vps34 with a bound inhibitor (PDB 4UWH), estimated from 20,000 sampler calls. (a) Molecular surface coloured by selection probability; the box marks the view in (b). (b) Ligand and nearby side chains (residues with an atom selected in at least 5% of calls). Side-chain tips closest to the ligand are chosen most often, and backbone atoms almost never ($\leq 0.01$; not shown). (c) Three individual draws, motif atoms in red: each is a sparse, partial subset, and (b) is their average.

and spatial-interface cropping to define pocket-centred cropping. Instead of choosing the crop centre from a chain or cross-chain interface, we first select an annotated pocket, map the atom-level pocket annotation to the corresponding tokens, and choose the spatial crop centre from this pocket-token set. The final crop is then formed by taking the nearest tokens in three-dimensional space. Pocket-centred cropping is used only for the distillation dataset, where each training example is first restricted to a single fragment that preferentially contains an annotated pocket.

The precise crop sizes and model-specific crop differences are described in the corresponding ArtiosBackbone and ArtiosPotts training sections.

## 1.4 Motif Construction

To train motif-conditioned generation, we sample an atom-level motif from each cropped structure. When ligands are present, motif sampling is ligand-centred with probability 0.9; otherwise, the sampler falls back to annotated pockets when available and then to polymer-centred motifs. Ligand-centred motifs preferentially choose chemically non-trivial ligands, and the ligand portion of the motif is usually a partial connected subgraph rather than the full ligand. Specifically, one to four ligand regions are selected, and each region is expanded by breadth-first traversal on the ligand bond graph to reveal a small connected set of atoms.

After choosing the ligand or pocket seed, the surrounding protein motif is sampled from atoms within a random radius of 5–8 Å. Residues in this neighbourhood are included stochastically, usually revealing only residue-specific tip atoms rather than complete side chains (Supplementary Figure 5). The local atom selection is performed on the residue bond graph (Algorithm 4), giving the model partial active-site constraints rather than fully specified residue identities and geometries. Motifs containing only a single atom or more than 15 residues are rejected. With small probability, ligand atoms are dropped after selecting the surrounding protein atoms, producing pocket-only motifs, and nearby supplementary ligands can be included when they lie within 5 Å of the seed ligand.

The resulting motif annotation is then transformed before model input: motif backbone atoms are duplicated as conditioning atoms, non-motif side chains are removed, and motif residues are unindexed so that the backbone model is not told where in sequence the motif must occur. We also compute a motif-depth conditioning feature and randomly drop this feature during training with probability 0.1.

**Algorithm 3** Motif sampling

**def** SampleMotif($S, \mathcal{L}, \mathcal{P}$):

*# $S$: cropped structure, $\mathcal{L}$: ligands, $\mathcal{P}$: annotated pockets.*

**if** $\mathcal{L} \neq \emptyset$ and Bernoulli(0.9) $= 1$:
  $L^\star =$ sample ligand from $\mathcal{L}$, preferring chemically non-trivial ligands
  $M_{\text{lig}} =$ one to four connected regions on the bond graph of $L^\star$
  $\mathbf{x} =$ SampleAtom($M_{\text{lig}}$)
  $M_{\text{env}} =$ SampleLocalAtoms($S, \mathbf{x}$)
  **if** $M_{\text{env}} = \emptyset$: **return** $\emptyset$
  set $M_{\text{lig}} = \emptyset$ with probability 0.1
  $M_{\text{near}} =$ small connected regions from ligands within 5 Å of $\mathbf{x}$
  $M = M_{\text{env}} \cup M_{\text{lig}} \cup M_{\text{near}}$
*# If no ligand-centred motif is sampled, use pocket annotations when available.*
**else if** $\mathcal{P} \neq \emptyset$:
  $P^\star \sim$ Uniform($\mathcal{P}$)
  $\mathbf{x} =$ Centroid($P^\star$)
  $M =$ SampleLocalAtoms($S, \mathbf{x}$, mask $= P^\star$)
**else**: **return** $\emptyset$
**if** $|M| = 1$ or $|\text{Residues}(M)| > 15$: **return** $\emptyset$
**return** $M$

**Algorithm 4** Local motif atom sampling

**def** SampleLocalAtoms($S, \mathbf{x}$, mask $=$ all):

*# The mask restricts candidates to an annotated pocket when one is supplied.*

$r \sim U(5\ \text{Å}, 8\ \text{Å})$
$C =$ polymer atoms in $S$ within $r$ of $\mathbf{x}$ and inside mask
$b_{\text{tip}} \sim$ Bernoulli(0.95)
**if** $b_{\text{tip}}$: restrict $C$ to residue-specific tip atom names
**if** $C = \emptyset$: **return** $\emptyset$
$M = \emptyset$
**for each** residue $R$ represented in $C$:
  **if** Bernoulli(0.8) $= 0$: **continue**
  $G_R =$ covalent bond graph of the atoms in $R$
  $a =$ sample atom from $C \cap R$
  $d \sim$ Geometric(0.5) $- 1$
  $M_R =$ atoms in $R$ reachable from $a$ within $d$ bonds in $G_R$
  **if** $b_{\text{tip}}$: restrict $M_R$ to residue-specific tip atom names
  $M = M \cup M_R$
**return** $M$

# 2 Models

Our computational pipeline (overview figure in the main text) uses a two-stage design paradigm: the first stage uses ArtiosBackbone to scaffold a backbone–ligand complex around a motif, followed by ArtiosPotts for sequence design. In this section, we describe the models' architectures and training details.

## 2.1 ArtiosBackbone

ArtiosBackbone is a diffusion-based generative model that generates protein backbone–ligand complexes. The model can be conditioned on a broad set of atom-level constraints, enabling what is commonly referred to in the literature as motif scaffolding [8–10]. This enables reliable recapitulation of active-site geometries in a diverse set of protein structures, a fundamental design task required for enzyme design.

### 2.1.1 Architecture

The model adopts a simple pair-biased diffusion transformer architecture, related to recent diffusion and flow-based protein structure generators [1, 11], with no AF3-style Pairformer module and no pair updates. The input structure is tokenized as follows: the protein backbone is tokenized per residue, ligand atoms are tokenized per heavy atom, and the side chains involved in the motif are tokenized per heavy atom. Single representations are generated for all tokens using basic features (such as noised coordinates, charge, atom name, etc.), and a pair representation is generated using token–token distances and relative positional encodings. As in AlphaFold3 [1], the pair representation enters the network as an additive bias on the attention logits. The same pair bias is applied to all transformer blocks.

### 2.1.2 Featurisation

ArtiosBackbone uses the shared all-atom context produced by the data pipeline, but only a subset of these tensors are used as conditioning features. At each diffusion step, the model receives noised all-atom coordinates, the corresponding EDM noise level, and, when self-conditioning is active, the previous denoised coordinate prediction. Reference-conformer features encode atom identity and chemistry, including element, formal charge, atom-name characters, conformer mask, and reference coordinates. These atom-level features are used primarily for ligand atoms and motif-related side-chain atoms, then aggregated to token-level single features and token-pair features based on reference-coordinate displacements and inverse squared distances.

Token identity features, including residue index, token index, chain identity, entity identity, and symmetry index, enter through relative positional encodings in the pair representation. Motif conditioning is supplied through motif-token masks, motif residue identity, motif coordinates, pairwise motif-distance bins, and a normalised motif-depth feature. Motif depth measures how deeply the motif is embedded in the complex, computed as the distance between the motif and non-motif centres of mass normalised by the radius of gyration of the resolved structure. This scalar gives the model a global cue for whether the motif should sit near the surface or more deeply buried in the scaffold; in the network, it is discretised into bins and embedded into the single representation. Motif residue identities and relative coordinates are embedded and added to the single representation, while intra-motif token distances are added to the pair representation. The final sequence positions of motif residues are withheld from the feature set, enabling unindexed motif scaffolding [8]: at inference time, the network must decide where to place the protein motif residues within the generated backbone while preserving the local motif geometry. This positional freedom is important for enzyme design, where active-site residues often need to be embedded into newly generated scaffolds rather than copied at predetermined sequence indices.

### 2.1.3 Training Details

We adopt the EDM parameterisation of Karras et al. [12], as used in AlphaFold3 [1]: for each training sample, we sample a noise level

$$\hat{t} \sim \sigma_{data} \cdot \exp(-1.2 + 1.5 \cdot \mathcal{N}(0,1)) \tag{2}$$

where $\sigma_{data}$ is a constant determined by the variance in the training coordinates (set to 16, following AF3). We centre and randomly augment the ground-truth sample $\mathbf{x}_0^{GT}$, then add noise to obtain the noisy sample $\mathbf{x}_{\hat{t}} \sim \mathbf{x}_0^{GT} + \sigma_{data} \cdot \hat{t} \cdot \mathcal{N}(0,1)$. During training, we use self-conditioning with probability 0.5: when selected, the model first makes a no-gradient denoised prediction from the same noisy input, and this prediction is then provided as an additional input to the supervised denoising pass. We feed $\mathbf{x}_{\hat{t}}$ into the network and supervise the denoised output $\hat{\mathbf{x}}_0$ with the AF3 weighted, rigid-aligned MSE loss:

$$L_{diffusion} = \frac{\hat{t}^2 + \sigma_{data}^2}{(\hat{t} \cdot \sigma_{data})^2} \cdot L_{MSE} \tag{3}$$

where $\hat{t}$ is the sampled noise level and

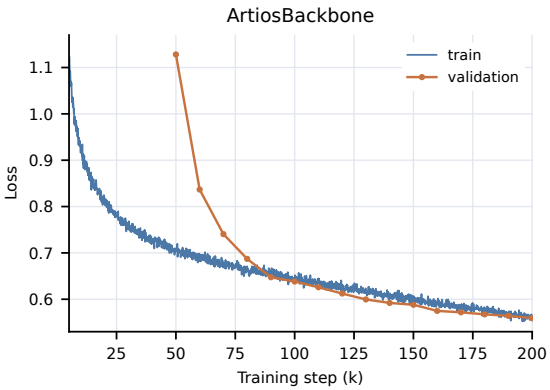
$$L_{MSE} = \frac{1}{3}\text{mean}\left(w_l \left\|\hat{\mathbf{x}}_0 - \mathbf{x}_0^{GT-\text{align}}\right\|^2\right) \tag{4}$$

where $\mathbf{x}_0^{GT-\text{align}}$ denotes the ground-truth structure that has been aligned to $\hat{\mathbf{x}}_0$ and $w_l$ denotes the weighting of certain entities.

$$\begin{aligned} w_l = 1 &+ \mathbb{I}_{protein} \cdot \alpha_{protein} + \mathbb{I}_{ligand} \cdot \alpha_{ligand} \\ &+ \mathbb{I}_{nucleic} \cdot \alpha_{nucleic} + \mathbb{I}_{motif} \cdot \alpha_{motif} \end{aligned} \tag{5}$$

with $\alpha_{protein} = 1, \alpha_{ligand} = 10, \alpha_{nucleic} = 0$, and $\alpha_{motif} = 5$.

In the motif-scaffolding training configuration, ArtiosBackbone is trained on a dataset-level mixture of weighted PDB examples and MGnify distillation examples. The two datasets are sampled with equal aggregate probability after normalizing per-example weights within each dataset. The PDB component uses the weighted chain/interface sampler described in Section 1.1; the MGnify component uses the cluster-balanced, length-reweighted sampler described in Section 1.2.




**Supplementary Figure 6:** ArtiosBackbone diffusion loss during training and EMA validation after the EMA cold-start period.

All ArtiosBackbone training examples use a 256-token crop. PDB examples use spatial and spatial-interface crops with equal probability. MGnify examples are first restricted to a single high-confidence fragment, preferentially one containing an annotated pocket, and are then cropped using pocket-centred crops with probability 0.8 and spatial crops with probability 0.2.

ArtiosBackbone is trained with a global batch size of 128 (16 per GPU). **ArtiosBackbone was trained in 26.5 hours on a single node of eight NVIDIA A100 (80 GB) GPUs (about 210 GPU-hours).**

### 2.1.4 Inference

We use the default AF3 noise schedule:

$$\hat{t} = \sigma_{data} \cdot (s_{max}^{1/p} + t \cdot (s_{min}^{1/p} - s_{max}^{1/p}))^p \tag{6}$$

where $s_{max} = 160$, $s_{min} = 4 \cdot 10^{-4}$, $p = 7$, and $t$ is uniformly distributed in the interval $[0, 1]$. Our default inference setting is $n_{\text{step}} = 200$. During sampling, we use self-conditioning across diffusion steps by feeding the previous denoised coordinate prediction into the next denoising step. We use no guidance methods. At inference time, motif-depth conditioning is optional: when a target normalised depth is provided, the same binned scalar condition biases generation toward placing the motif at that depth; otherwise, sampling is conditioned only on the motif geometry.

## 2.2 ArtiosPotts

ArtiosPotts follows and builds on the recent trend in sequence design by joining a Potts head onto the end of a ProteinMPNN-style GNN [13], transforming MPNN embeddings into an energy function that can be used to sample sequence space, following Caliby and L-Caliby [14, 15]. As others have demonstrated, using a Potts model for sequence design provides a demonstrable improvement in both traditional folding metrics (such as pLDDT and scRMSD) and ligand-aware scores (ligand pLDDT, Ligand-RMSD); see also the sequence design benchmark in the main text and Section 3.2. Framing the output of the model with an energy function that pivots on pairwise interactions between residues is a powerful inductive bias that favours sequence design, substantially improving upon the traditional autoregression-based training and sampling methods used by long-standing models in the space, including ProteinMPNN and LigandMPNN [13, 16]. We additionally find that the MCMC sampling method permits inference-time steering towards arbitrary objectives, which is key for tuning active-site environments for enzyme design.

### 2.2.1 Architecture

The architecture of ArtiosPotts has two stages: a ProteinMPNN/LigandMPNN-style graph neural network [13, 16] that encodes structure into node and edge embeddings, followed by a Potts head that transforms those features into the parameters of an energy model over sequence space. Unlike LigandMPNN, which creates three separate graphs (protein-only, intra-ligand, and protein–ligand), ArtiosPotts builds a unified graph over the whole input complex using k-nearest-neighbours between protein backbone $C_\alpha$ atoms, all heavy ligand atoms, and all side-chain heavy atoms of the protein motif residues. The Potts head maps the GNN's node and edge embeddings to a Markov random field over the 20 amino acids, outputting per-residue fields $h_i$ and pairwise couplings $J_{ij}$.

### 2.2.2 Featurisation

ArtiosPotts uses the same tokenised complex as ArtiosBackbone, but featurises it as a sparse structural graph. Graph nodes correspond to protein residues, ligand heavy atoms, and motif side-chain heavy atoms. The neighbourhood graph is built from k-nearest neighbours between token centre atoms. Node features include molecule-type indicators, non-protein residue type, and ligand or motif-related reference chemistry, where atom-level element, charge, and conformer-mask features are averaged to their corresponding tokens.

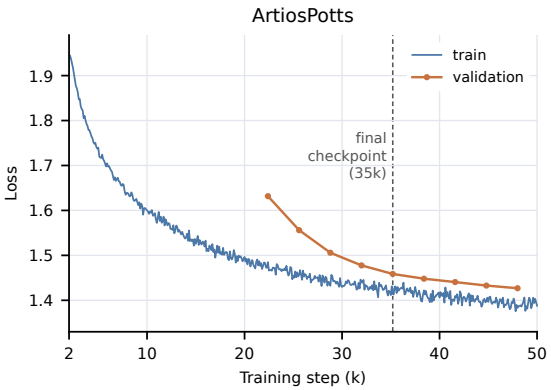
Edge features combine local structure and sequence-relative context. For each graph edge, ArtiosPotts computes all pairwise distances between the five backbone atom slots of the source and target tokens, expands these distances with radial basis functions, and concatenates them with a relative positional encoding derived from residue, token, chain, entity, and symmetry indices. During Potts training, the decoder is provided an all-mask sequence input, so the learned fields $h_i$ and couplings $J_{ij}$ are conditioned on structure and ligand or motif context rather than on the native sequence.

### 2.2.3 Training Details

Relative to the Potts-model training procedures used in Caliby and Chroma [14, 17], we make a few simplifications. For every training example, the MPNN decoder receives an all-mask sequence input, so the predicted fields $h_i$ and couplings $J_{ij}$ depend on the structure, ligand, motif, and graph

context, but not on a partially observed native sequence. We train these parameters directly with a composite likelihood objective on the native sequence: for each graph edge, the model scores the joint native residue identities of the edge endpoints while excluding the direct self-edge contribution, and the resulting per-position losses are averaged over valid protein tokens outside the fixed motif context. We also experimented with MSA-based supervision, but did not observe improvements on the sequence-design benchmarks considered here, so we omit it for simplicity.

We add Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to the coordinates of all atoms during training, with $\sigma = 0.3$ Å. We train with AdamW with a weight decay of $1 \times 10^{-2}$. The model is trained for 35k steps with a batch size of 32 on the data described in Section 1. Unlike ArtiosBackbone, ArtiosPotts uses only the weighted PDB component of the training data distribution, without the MGnify distillation set. Within the PDB component, we use the chain/interface weighting described above, but apply a Potts-specific boost to monomer-only PDBs: structures with chain rows but no interface rows are reweighted to account for 40% of the PDB sampling mass. This counteracts the overrepresentation of multichain and interface-rich structures, which contribute many more rows per PDB entry. We train with a learning rate of $5 \times 10^{-4}$, warming up linearly from 0 over the first 2,000 steps. Our training crop size is 384 and the data type is bf16. All results quoted in this paper are taken from the EMA checkpoint at 35k using an EMA decay factor of 0.9999. **ArtiosPotts was trained in about 3 hours on two NVIDIA A100 (40 GB) GPUs (about 6 GPU-hours).**




**Supplementary Figure 7:** ArtiosPotts protein negative log likelihood loss during training and EMA validation after the EMA cold-start period. The dashed line marks the 35k-step checkpoint used in this work.

### 2.2.4 Inference

At inference time, we sample sequences from the learned Markov random field via MCMC sweeps, starting from a crudely initialised sequence. Given $h_i$, $J_{ij}$, and a sequence state $S = (s_1, \ldots, s_N)$, the Potts energy is

$$E(S) = \sum_i h_i(s_i) + \frac{1}{2} \sum_i \sum_{j \in \mathcal{N}(i)} J_{ij}(s_i, s_j), \tag{7}$$

where $\mathcal{N}(i)$ are the graph neighbours of residue $i$, excluding $i$ itself. Each Gibbs update resamples residue $i$ from its conditional distribution at temperature $T$, with all other residues $S_{-i}$ held fixed:

$$p(s_i = a \mid S_{-i}) = \frac{\exp\left(-U_i(a)/T\right)}{\sum_b \exp\left(-U_i(b)/T\right)}, \qquad U_i(a) = h_i(a) + \sum_{j \in \mathcal{N}(i)} J_{ij}(a, s_j). \tag{8}$$

We use the coloured Gibbs sampling method used in Chroma [17] to draw from this distribution, which entails partitioning the sequence into subsets that contain no nearest neighbours and then sampling Equation 8 within each colour, updating the sequence state $S$ as we move between colours. The temperature $T$ decays linearly from 1.0 to 0.1 over the first 80% of the sweeps.

Like Caliby, we average our per-residue fields $h_i$ and pairwise couplings $J_{ij}$ over an ensemble of structures:

$$h_i = \frac{1}{n} \sum_{k=1}^{n} h_i^{(k)}, J_{ij} = \frac{1}{n} \sum_{k=1}^{n} J_{ij}^{(k)}$$

where $\{(h_i^{(k)}, J_{ij}^{(k)}) : 1 \leq k \leq n\}$ are the fields and couplings of a structural ensemble of size $n$. We generate synthetic ensembles by adding a small amount of Gaussian noise to the input structure's coordinates. In Caliby, ensemble conditioning substantially improved the AlphaFold2 self-consistency of designed sequences [14].

**Inference-time sequence steering.** Because ArtiosPotts samples from an explicit sequence energy model, we can add lightweight sequence-level terms during MCMC without retraining the network. In the default setting, we use a Chroma-style local composition penalty [17] that penalizes low-entropy amino-acid windows, reducing low-complexity sequence collapse. We also enforce hard alphabet constraints when useful, such as excluding cysteine and methionine at mutable positions. For campaign-specific design goals, we add residue- or residue-group energy biases directly to the Potts fields. The same framework also supports optional net-charge and active-site enrichment penalties, but these were not used in the default benchmark setting.

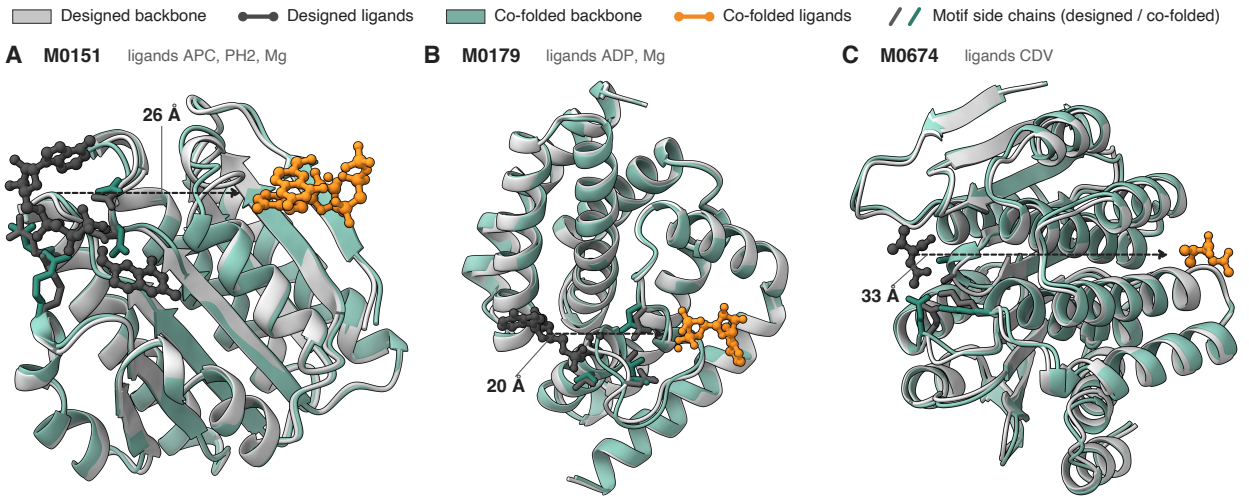
# 3 Benchmarking

In this section, we provide comprehensive *in silico* benchmarks of our computational enzyme design pipeline.

## 3.1 Atom Motif Enzyme Benchmark

We benchmark our system's ability to perform motif scaffolding by using the Atomic Motif Enzyme benchmark from [8]. The AME benchmark is a set of 41 input specifications used to test a model's capability at what is referred to as "motif scaffolding". This entails the design of a protein sequence that will co-fold with any relevant substrates and recapitulate the input specification's motif geometry. The AME set covers a broad range of input motifs, comprehensively assessing the system's understanding of the interplay between structure, sequence, and ligand positioning.

Under the standard success criteria [8, 18, 19], a design passes if its sequence refolds to the designed backbone with high confidence and reproduces the motif; **the position of the ligand is not checked**. A designable protein whose catalytic residues are in the right place but whose pocket cannot hold the substrate therefore counts as a success. Supplementary Figure 8 shows three such designs. As argued in the main text, we treat such designs as failures.




**Supplementary Figure 8:** Generated structures considered a success under the standard AME filters that fail under the revised filters. Backbones were generated with Complexa and sequences designed with ArtiosPotts. The Protenix co-fold of each designed sequence is superposed on the designed structure by C$\alpha$ atoms. Designed backbone shown in light grey, designed ligands in dark grey; co-folded backbone shown in green, co-folded ligands in orange. Motif side chains are shown as sticks in the designed (dark grey) and co-folded (green) structures, and magnesium ions as small spheres. Dashed arrows join the designed and co-folded positions of the organic substrate, labelled with its pocket-aligned centroid RMSD. **A** AME input M0151. **B** AME input M0179. **C** AME input M0674.

We append onto the standard set of AME filters a ligand positioning filter: the centroids of all ligands present must not move more than 3 Å after being co-folded with the designed sequence

(criterion 3 below). The final set of filters assess the systems ability to generate a diverse set designable backbones which recapitulate the motif and substrate geometry. We consider this a comprehensive *in silico* assessment of a model's aptitude for enzyme design. We present the adjusted AME below:

We design 125 protein-ligand complexes for lengths $\in \{200, 220, 240, 260, 280, 300, 320, 340\}$ leaving us with 1000 complexes. If a model with no codesign was performed, we design a single sequence with a sequence design model, then fold the sequence with AF2, followed by co-folding with the ligand(s) using Protenix V1). We then check for the following criteria:

1. **AF2 backbone self consistency:** The generated protein backbone and the AF2-folded backbone have RMSD $< 2.0\,\text{Å}$, and the backbone is confidently predicted (pLDDT $> 80$).
2. **AF2 motif geometry recapitulation:** The AF2-folded structure reproduces the motif with all-atom RMSD $< 2.5\,\text{Å}$.
3. **Protenix ligand position consistency:** Every ligand present has a pocket-aligned centroid RMSD $< 3.0\,\text{Å}$ between the Protenix co-fold and the generated structure. We align using the pocket residues of the generated structure, defined as all residues with a heavy atom within $8\,\text{Å}$ of any ligand atom.

We compare how our two-stage enzyme pipeline – using ArtiosBackbone to generate a backbone that scaffolds the input motif, and ArtiosPotts to design the full sequence – performs relative to other state-of-the-art enzyme design pipelines and find that our system presents significant improvement over the rest of the field.

### 3.2 Sequence Design Benchmarking

LigandMPNN is commonly considered the state-of-the-art design model for protein–ligand complexes. ArtiosPotts uses a similar GNN to LigandMPNN to encode the protein–ligand structure, but differs in that LigandMPNN uses autoregression-based decoding, whereas ArtiosPotts leverages a Potts head to perform MCMC sampling on the whole sequence. Caliby has demonstrated that Potts-based sampling provides a substantial improvement over autoregression-based decoding on folding consistency but is trained without ligand context.

In this section, we detail how ArtiosPotts improves upon LigandMPNN and Caliby on the task of designing both natural and de novo protein–ligand complexes. Success on this task requires the following criteria to be met:

1. *Folded backbone consistency:* The designed sequence should fold into a geometry consistent with the input structure. To measure this, we use the following filters:
   (a) Protenix backbone RMSD $< 2.0$
   (b) Protenix pLDDT $> 80.0$
2. *Ligand consistency:* The ligand position should be consistent with the original position when it is co-folded with the designed sequence. To measure this, we use the following filters:
   (a) Protenix ligand centroid RMSD $< 3.0$ for all ligands present.

We do not include recovery of the native sequence in these criteria, as it has been shown that optimising for native sequence recovery is negatively correlated with folding consistency [14].

To assess the task of designing sequences for natural protein–ligand structures, we curated a diverse set of protein–ligand complexes called *SeqVal*. Since ArtiosPotts is trained using the AF3 train–val splits, we derive *SeqVal* from AF3's Recent PDB evaluation set, a temporal and low-homology holdout: all entries were released between 2022-05 and 2023-01 (after the AF3 2021-09-30 training cutoff), filtered to be sequence-dissimilar from the AF3 training set, and capped at 1536

| | Filter | Remaining |
|---|---|---|
| | Unique structures in the low-homology, ≤1536-token set | 1,818 |
| (a) | ≥1 non-covalent protein–ligand interface | 503 |
| (b) | No nucleic acid in the assembly | 427 |
| (c) | Exactly one protein chain | 162 |
| (d) | Fewer than 5 ligand molecules | 157 |
| (e) | Longest fully-resolved backbone segment retains its pocket | 143 |
| (f) | Segment length ≥ 100 residues | 139 |
| (g) | Most-buried ligand ≥ 0.4 SASA burial | 127 |
| (h) | Successful curation to ArtiosPotts format | **114** |

**Table 2:** Construction of *SeqVal* from the AF3 Recent PDB set. Each row reports the number of structures remaining after the filter. Pocket retention (e) requires every organic ligand to keep ≥4 contacting residues within 5 Å or ≥0.4 SASA burial against the cropped segment.

tokens. Starting from the 1,818 unique structures in this set, we applied the sequence of filters in Table 2, yielding a final set of 114 single-receptor protein–ligand design targets.

The sequence design benchmark figure in the main text reports how ArtiosPotts and LigandMPNN perform on *SeqVal*. For each complex, we design 10 sequences with each model and co-fold each sequence with the complex's ligands using Protenix; metrics are computed on the top-ranked Protenix sample. We consider a complex to have been successfully designed by a model if at least one of its sequences passes filters (1) and (2). Full data for this benchmark are presented in Table 7.

The panels of that figure are constructed as follows. In the filter cascades (panels A and B), the ligand centroid filter is applied first, followed by the backbone RMSD and pLDDT filters; a complex passes a step if at least one of its 10 sequences passes every filter up to that step, and absent ligand slots pass. The AF2 cascade replaces the Protenix backbone RMSD and pLDDT with those of the AlphaFold2 prediction of the designed sequence, but keeps the Protenix ligand centroid filter, since AF2 does not model ligands. In panel D, the best value among a complex's 10 designs is taken separately for each metric (maximum pLDDT, minimum backbone RMSD), so the points for one complex in different panels may come from different designs. Three complexes have no folding metrics and are omitted from panels D and F, leaving 111. Panel F shows the mean and standard deviation of the per-sequence amino-acid fractions over the designs for these 111 complexes, with the native sequences of the same complexes for reference. For the 7R0U example in panel G, we show for each model the design with the lowest maximum pocket-aligned ligand centroid RMSD among its 10, superposed on the native backbone by C$\alpha$ atoms with iterative pruning of poorly aligned pairs.

# 4 Design Campaigns

## 4.1 Luciferase

In the de novo luciferase study, the target reaction was the oxidative chemiluminescence of the synthetic coelenterazine analogue diphenylterazine (DTZ). The proposed mechanism involves formation of an anionic imidazopyrazinone species, followed by single-electron transfer to triplet oxygen and oxygenation to a dioxetane intermediate [20]. Subsequent dioxetane thermolysis and $CO_2$ extrusion result in photon emission.

### 4.1.1 Design Pipeline

The computational design began with the construction of a minimal catalytic motif. The theozyme was designed to stabilize the anionic state of DTZ and lower the barrier for oxygen activation, principally through placement of a cationic arginine and histidine in proximity to the emerging delocalized anionic charge.

We generated 1,000,000 de novo protein backbones that preserved this four-atom motif. We designed equal numbers of proteins with lengths 160, 175, and 190 amino acids. Candidates were then passed through two mutually exclusive filter sets (Table 3). The main filter selected designs satisfying the target catalytic geometry, whereas the $\pi$-stacking rescue filter retained designs that narrowly failed the main filter but satisfied a relaxed catalytic geometry together with an explicit predicted aromatic stacking interaction with the ligand. Pairwise sequence identities ranged from 22% to 31% with a mean of 26%, placing them within the twilight zone of sequence similarity.

These filters produced 85 designs for experimental testing. We predefined a hit criterion as the mean negative-control signal plus three standard deviations. To require reproducibility across replicates, the lower bound of the sample signal, calculated as the sample mean minus one sample standard deviation, was also required to exceed this threshold. A design was classified as a hit as described:

$$\bar{x}_{\text{sample}} - s_{\text{sample}} > \bar{x}_{\text{neg}} + 3s_{\text{neg}}.$$

where $\bar{x}_{\text{sample}}$ and $s_{\text{sample}}$ are the mean and standard deviation of the replicate sample wells, and $\bar{x}_{\text{neg}}$ and $s_{\text{neg}}$ are the mean and standard deviation of the matched negative-control wells.

### 4.1.2 Analysis of Designs

Subsequent analysis of hits found that all pairwise TM-scores ranged from 0.30 to 0.52, indicating that each design adopted a distinct fold rather than a shared scaffold. None of the seven sequences had a statistically significant homolog in the NCBI nr database by BLASTp ($E \geq 0.3$), confirming that all were novel at the sequence level. A Foldseek search of the full PDB identified no close natural structural homologs for six of the seven designs, whose best matches were to functionally unrelated proteins with TM-scores of 0.48–0.68. The remaining design adopted a natural calycin/iLBP-like $\beta$-barrel fold, with comparable structural similarity to NanoLuc/nanoKAZ, FABP7, myelin P2, and UnaG (TM-align TM-score $\sim$0.75). Thus, its structural match appears to reflect the generic calycin/iLBP scaffold rather than luciferase-specific similarity.

### 4.1.3 In vitro Evaluation

Initial screens for luciferase activity were conducted with cell-free expression lysate and the synthetic luciferin substrate diphenylterazine (DTZ). DTZ, ordered from MCE, was diluted in PBS buffer. The assay followed the format described by Yeh et al. [20]. Endpoint luminescence measurements were acquired using a GloMax instrument (Promega). Lysate-based screens were performed in white 96-well half-area plates (Corning); 12.5 µL of cell-free lysate was combined with DTZ at a final concentration of 25 µM in a total reaction volume of 50 µL. Candidate wells and the six negative controls were run in triplicate. Each replicate was measured three times at one-minute intervals, and the mean of these readings was used.

For Michaelis–Menten kinetic measurements, 50 µL of serially diluted DTZ substrate was added to wells of a white 96-well microplate (Corning) containing 50 µL of purified luciferase, giving a final enzyme concentration of 1 µM and final substrate concentrations ranging from 0.39 to 50 µM. Luminescence was measured every 1 min for 10 min using a 0.1 s integration time, with 10 s shaking between each measurement. Initial velocities were estimated from the average luminescence intensities of the first three data points and fitted to the Michaelis–Menten equation as in Yeh et al.

**Table 3: Computational filters used across the DTZ luciferase design campaigns.** We initially assayed our main-filter candidates and our $\pi$-stacking interaction candidates. Following a successful design round, we tested longer sequences to investigate the model's ability to design larger enzymes using large luciferase filters.

| **Filter category** | **Main catalytic geometry filter** | **$\pi$-stacking rescue filter** | **Large luciferase filters** |
|---|---|---|---|
| Backbone motif RMSD | $< 1.0$ Å | $< 1.0$ Å | $< 1.0$ Å |
| Motif stitch RMSD | $< 0.3$ Å | $< 0.3$ Å | $< 0.3$ Å |
| AF2 backbone RMSD | $< 2.0$ Å | $< 2.0$ Å | $< 2.3$ Å |
| AF2 pLDDT | $> 80$ | $> 80$ | $> 80$ |
| Backbone RMSD (Protenix) | $< 2.0$ Å | $< 2.0$ Å | $< 2.3$ Å |
| pLDDT (Protenix) | $> 80$ | $> 80$ | $> 80$ |
| Arginine–DTZ distance | Arg NH1/NH2–DTZ N1 $< 4.0$ Å | Arg NH1/NH2–DTZ N1 $< 4.55$ Å | Arg NH1/NH2–DTZ N1 $< 4.3$ Å |
| Histidine–DTZ distance | His NE2–DTZ O1 $< 4.0$ Å | His NE2–DTZ O1 $< 4.55$ Å | His NE2–DTZ O1 $< 4.3$ Å |
| $\pi$-stacking interaction (Protenix) | Not required | $\pi$-stacking interaction count $> 2$ | Not required |
| Number of designs sent for validation | 73 | 12 | 48 |
| Hits | 6 | 1 | 1 |

## 4.2 Esterase and Cutinase

In cysteine hydrolases, catalysis can generally be understood in terms of a nucleophilic cysteine, a nearby histidine that functions as a general acid/base, and one or more residues that orient and tune the histidine. Activation of the cysteine thiol generates a reactive thiolate, which attacks the substrate carbonyl to form a tetrahedral intermediate. The developing oxyanion is stabilized by an oxyanion hole formed by backbone amides and/or side-chain hydrogen-bond donors. Collapse of this intermediate produces an acyl-enzyme thioester, which is subsequently hydrolysed to release product and regenerate the catalytic cysteine.

### 4.2.1 Esterase Design Pipeline

The computational design began with the construction of a minimal cysteine-hydrolase motif for ester hydrolysis. The target reaction was cleavage of 4-methylumbelliferyl butyrate (4MU-Bu), in which hydrolysis of the ester bond releases fluorescent 4-methylumbelliferone. We modelled the catalytic chemistry around the tetrahedral intermediate formed after nucleophilic attack of a cysteine thiolate on the substrate carbonyl. The motif therefore encoded the key features of a cysteine hydrolase active site: a catalytic cysteine positioned for nucleophilic attack, a histidine positioned to activate the cysteine thiol, an asparagine positioned to preorganize the histidine, and an oxyanion-hole interaction involving the cysteine backbone amide and a glutamine side chain.

RFDiffusion2 [8] and RFDiffusion3 [18] cysteine-hydrolase design efforts used a larger motif derived from natural cysteine hydrolases, including non-catalytic flanking backbone residues around the catalytic cysteine to bias the nucleophile toward the N-terminus of an $\alpha$-helix and thereby exploit the helix-dipole contribution to oxyanion stabilization [8, 18]. In preliminary design experiments, however, we found that our model performed better when we specified less of the motif and allowed the model to flexibly design its own secondary structure in these regions. We therefore did not include these flanking residues in our theozyme and conditioned the backbone model only on the compact catalytic arrangement. Interestingly, the model frequently regenerated an analogous helix-

N-terminal cysteine geometry even without explicitly fixing the flanking residues.

We generated 1,000,000 de novo protein backbones, consisting of 250,000 designs at each of four amino acid lengths: 170, 190, 200, and 210. Candidate backbones were sequence-designed using our sequence model, folded with AlphaFold2, and filtered for global fold confidence and motif preservation. Designs that passed these filters were then folded with Protenix and subjected to active-site geometry filters (Table 4).

The initial computational filters produced 88,043 designs, and the final Protenix-based geometry filters produced 297 structures across 149 unique clusters. From this filtered set, 85 designs were selected for cell-free expression and experimental screening. Activity was measured using the 4MU-Bu fluorescence assay, with excitation at 360 nm and emission at 450 nm. We predefined a primary hit as a design whose steady-state fluorescence signal exceeded the mean signal of the negative controls by more than three standard deviations, as described in RFDiffusion3.

#### 4.2.2 Cutinase Design Pipeline

We next performed a continuation esterase design campaign using the same minimal cysteine-hydrolase catalytic logic. The input motif retained the tetrahedral-intermediate-inspired ester geometry, but we changed to the eight-carbon-chain substrate 4MU-octanoate.

We generated 1,000,000 de novo protein backbones, with 250,000 designs at each of four lengths: 170, 190, 200, and 210 amino acids. In the redesign stage for this pipeline, selected backbones were expanded by a factor of 10, redesigned with our sequence model, folded with AlphaFold2, and designs that passed were evaluated with Protenix. This strategy reduced the number of Protenix evaluations required while increasing the number of final candidates available for selection.

The final filtering used the same fold-quality criteria as the first esterase campaign, including backbone motif RMSD, motif stitch RMSD, AlphaFold2 pLDDT and backbone RMSD, and Protenix pLDDT and backbone RMSD. Active-site filters were relaxed to ensure enough passing designs. Our filters produced 191 final sequences from 161 unique backbones. Of these, 58 were derived from the pre-redesign stream, with the remainder generated after redesign, indicating that redesign increased the final candidate pool without collapsing onto a small number of backbone scaffolds. A hit was defined as in the previous esterase campaign.

#### 4.2.3 In vitro Evaluation

Esterase activity in cell-free lysates was assayed using fluorogenic 4-methylumbelliferyl ester substrates (ordered from MCE). Lysate-based assays were performed in black 96-well microplates in a final volume of 100 µL. Each well contained 10 µL of cell-free lysate and 90 µL of substrate-containing assay buffer. For 4-methylumbelliferyl butyrate (4MU-butyrate) assays, substrate buffer was prepared from a 10 mM 4MU-butyrate stock in DMSO and diluted into PBS containing 5% DMSO. The final substrate concentration in the assay was 100 µM 4MU-butyrate. Where indicated, the longer-chain substrate, 4-methylumbelliferyl octanoate (4MU-octanoate), was assayed under the same format.

Fluorescence measurements using excitation at 365 nm and emission at 415–445 nm were taken at one-minute intervals for 60 min using a 0.1 s integration time and 10 s shaking between reads. Buffer-only reactions containing substrate but no protein were used to quantify assay background and non-enzymatic substrate hydrolysis. Sample traces were background-corrected by subtracting the buffer-only signal, and activity was evaluated relative to the purified no-DNA CFPS control.

**Table 4: Computational filters used across the esterase design campaigns.**

| Filter category | Esterase filter set 1 | Esterase filter set 2 |
|---|---|---|
| Backbone motif RMSD | $< 1.0$ Å | $< 1.0$ Å |
| Motif stitch RMSD | $< 0.6$ Å | $< 0.5$ Å |
| AF2 backbone RMSD | $< 3.0$ Å | $< 3.0$ Å |
| AF2 pLDDT | $> 70$ | $> 70$ |
| Backbone RMSD (Protenix) | $< 3.0$ Å | $< 3.0$ Å |
| pLDDT (Protenix) | $> 70$ | $> 70$ |
| Cys–His distance | HIS ND1–CYS SG $< 5.55$ Å | HIS ND1–CYS SG $< 5.5$ Å |
| Asn–His preorganization | ASN4 ND2–HIS NE2 $< 4.5$ Å or ASN4 OD1–HIS3 NE2 $< 4.5$ Å | ASN4 OD1–HIS NE2 $< 5.0$ Å |
| Gln–Cys backbone interaction | Not applied | GLN OE1–CYS N $< 6.0$ Å |
| Oxyanion-hole interaction: carbonyl O to Cys backbone N | LIG O1–CYS N $< 5.0$ Å | LIG O1–CYS N $< 5.5$ Å |
| Oxyanion-hole interaction: carbonyl O to Gln | GLN1 NE2–LIG O1 $< 5.0$ Å or GLN1 OE1–UNL1 O1 $< 5.0$ Å | GLN1 OE1–LIG O1 $< 5.0$ Å |

## 4.3 Phosphodiesterase

Zinc phosphodiesterases catalyse hydrolysis of phosphoester bonds through activation of a metal-bound water nucleophile and stabilization of a developing negative charge in the transition state. For our screening assay we use the hydrolysis of bis-*p*-nitrophenyl phosphate (bis-pNPP) as previously described[21]. This model reaction releases the chromogenic reporter,*p*-nitrophenolate , which can be monitored by absorbance at 400 nm.

### 4.3.1 Design Pipeline

The computational design began with a minimal binuclear-zinc phosphodiesterase motif based on the expected geometry for bis-pNPP hydrolysis. The motif specified two zinc ions positioned near the substrate phosphoryl oxygen, with distances chosen to support substrate polarization, metal-assisted water activation, and stabilization of the hydrolytic transition state. We evaluated two motif-conditioning strategies. In the first, the input motif contained only the two zinc ions and substrate atoms needed to define the reactive geometry. In the second, the motif additionally specified protein coordination around each zinc ion, using coordinating residue triads to preorganize the dinuclear metal site.

Backbones were generated using the backbone model to generate designs at three target lengths, 220, 260, and 300 amino acids. Since the motif was relatively simple, most candidates from the model passed initial screening filters. This meant we only scaled up to 150,000 backbone candidates. The minimal-motif run generated approximately 100,000 backbones, and the protein-coordination run generated approximately 50,000 backbones, giving the 150,000 total backbone evaluations. Sequences were designed using the Potts sequence model and all designs were folded with AlphaFold2. Designs passing the initial fold-quality and motif-geometry filters were then folded with Protenix and filtered again for fold recapitulation, ligand placement, dinuclear zinc geometry, pocket burial, ligand confidence, and zinc coordination. At this point in the model development we decided to remove the filters measuring backbone motif RMSD and motif stitch RMSD. In both bases, a negligible number of sequences were failing these checks and the candidates failing still appeared relatively plausible. For that reason we decided to remove these filters from now on.

We required AlphaFold2 and Protenix folded backbone RMSD values below 2.0 Å and pLDDT values above 80. Designs were further required to preserve the three motif positions within 3.0 Å, maintain a zinc–zinc distance below 4.0 Å, and place both zinc ions near the substrate oxygen, with Zn–O distance cutoffs of 5.0 Å and 4.0 Å for the two zinc sites. We also required the designed pocket to be buried, with ligand-chain SASA below 55, and retained only structures with ligand-chain pLDDT values above 75. Finally, designs were filtered using a zinc-coordination analysis

script to identify structures in which both metals were coordinated by three protein residues.

This final coordination filter produced 58 sequences from the 50,000-design protein-coordination run and 52 sequences from the 100,000-design minimal-motif run.

### 4.3.2 In vitro Evaluation

Proteins were purified in TBS and assayed for bis-pNPP hydrolysis in the presence of zinc sulfate. Zinc sulfate ($ZnSO_4 \cdot 7H_2O$) was prepared as a 10 mM stock, then diluted 1:100 in MilliQ water to generate a 100 µM working stock. bis-pNPP was prepared as a 1 mM stock in MilliQ water.

For each 100 µL reaction, 94 µL of 1.06 µM enzyme was dispensed into the assay plate and supplemented with 2 µL of 100 µM zinc sulfate, giving final concentrations of 1 µM enzyme and 2 µM zinc. Plates containing enzyme and zinc were incubated for 15 min to allow metal equilibration. Reactions were initiated by adding 4 µL of 1 mM bis-pNPP, giving a final substrate concentration of 40 µM. Matched buffer, and buffer-plus-zinc wells lacking enzyme were used as blanks and a control protein expected to be inactive under the assay conditions was used as a negative control.

Formation of *p*-nitrophenolate was monitored by absorbance at 400 nm for 10 hours with 6 min reads. A design was classified as a hit if its activity exceeded the matched negative-control background by more than three standard deviations. Where replicate measurements were available, we additionally required the lower bound of the sample signal, calculated as the sample mean minus one sample standard deviation, to exceed the negative-control threshold:

$$\bar{x}_{\text{sample}} - s_{\text{sample}} > \bar{x}_{\text{neg}} + 3s_{\text{neg}}.$$

The computational filters are listed in Table 5.

**Table 5: Computational filters used for the phosphodiesterase design campaign.**

| Filter category | Cutoff |
|---|---|
| AF2 folded backbone RMSD | $< 2.0$ Å |
| AF2 pLDDT | $> 80$ |
| Protenix folded backbone RMSD | $< 2.0$ Å |
| Protenix pLDDT | $> 80$ |
| Protenix pocket centroid RMSD, ligand positions 1–3 | $< 3.0$ Å |
| Zn–Zn distance | $< 4.0$ Å |
| Zn1–substrate oxygen distance | $< 5.0$ Å |
| Zn2–substrate oxygen distance | $< 4.0$ Å |
| Pocket burial, ligand-chain SASA | $< 55$ |
| Ligand-chain pLDDT, chains 1–3 | $> 75$ |
| Final metal-site coordination | both zinc ions coordinated by three protein residues |
| Final sequences retained | 58 from protein-coordination run; 52 from minimal-motif run |

# 5 Experimental Methods

## 5.1 Gene Preparation and Protein Preparation

Coding sequences were codon optimized for expression in *Escherichia coli* using Benchling's codon optimization tool with the default organism-specific codon usage table. Genes were ordered from Abstract Biosciences or Twist Biosciences and amplified by PCR using Q5 High-Fidelity DNA Polymerase (NEB) according to the manufacturer's instructions. All constructs contained a T7 promoter, T7 terminator, and a C-terminal 6xHis tag. PCR products were purified using the Monarch Mag PCR and DNA Cleanup Kit (NEB) according to the manufacturer's instructions. Linear DNA templates were expressed using the CFXpress cell-free expression system (GenScript) according to the

manufacturer's instructions with additional supplementation of GamS (NEB). Proteins were purified using Ni-NTA affinity resin in a 96-well format as previously described in the Semi-Automated Protein Production (SAPP) workflow [22]. Following elution, buffer exchange into PBS was completed on the proteins using AcroPrep Advance 96-well filter plates (Cytiva) with an appropriate molecular weight cut-off. Protein expression was assessed by SDS-PAGE using Invitrogen Novex Tris-Glycine Mini Protein Gels (Thermo Fisher Scientific) and visualized using GelCode Blue Safe Protein Stain (Thermo Fisher Scientific).

### 5.2 Data Processing and Analysis

## A Reference Tables

### A.1 Additional Crystallization Aids

Table 6 lists the crystallization aids removed by Artios in addition to the common crystallization aids listed in the AlphaFold3 Supplement [1].

**Table 6:** Additional crystallization aids removed during Artios structure preparation.

| CCD code | Description |
|---|---|
| PGO | S-1,2-propanediol |
| BR | Bromide ion; phasing reagent |
| LMT | DDM; dodecyl-$\beta$-D-maltoside detergent |
| TLA | L(+)-tartaric acid; precipitant |
| BOG | OG; octyl $\beta$-D-glucopyranoside detergent |
| CAC | Cacodylate ion; buffer |
| DTT | Dithiothreitol; reducing agent |
| B3P | BIS-TRIS propane; buffer |
| 2PE | Nonaethylene glycol; PEG variant |
| DMU | Decyl-$\beta$-D-maltopyranoside detergent |
| XE | Xenon; noble-gas phasing reagent |
| MPO | MOPS buffer |
| P33 | Heptaethylene glycol / PEG 330 |
| PE8 | Heptaoxatricosanediol; PEG variant |
| BU3 | (R,R)-2,3-butanediol; cryoprotectant |
| CPS | CHAPS detergent |
| PE5 | PEG 400 |
| PDO | 1,3-propanediol; cryoprotectant |
| DMF | Dimethylformamide; solvent |
| KR | Krypton; noble-gas phasing reagent |
| XPE | Decaethylene glycol; PEG variant |
| MBN | Toluene; solvent |
| TCE | TCEP; reducing agent |
| SOR | Sorbitol; cryoprotectant |
| BU2 | 1,3-butanediol; cryoprotectant |
| SDS | Sodium dodecyl sulfate; detergent |
| SBT | 2-butanol; solvent |
| BET | Trimethyl glycine / betaine; osmolyte or precipitant |
| MLT | D-malate; precipitant |
| JEF | Jeffamine; PEG/PPG variant and cryoprotectant |
| EMC | Ethyl mercury ion; phasing reagent |

## A.2 *SeqVal* Sequence Design Benchmark Full Data

Table 7 presents the full data for the sequence design model benchmarking on *SeqVal* described in Section 3.2.

**Table 7:** Per-structure filtering outcomes on *SeqVal*, ordered by protein length. For each model, $n$ is the number of its designs passing the full filter set and *Succ.* is ✓ when $n \geq 1$. **Protenix filters**: ligand pocket-centroid RMSD $< 3$ Å, Protenix backbone RMSD $< 3$ Å, Protenix pLDDT $> 70$. **AF2 filters**: ligand pocket-centroid RMSD $< 3$ Å, AF2 backbone RMSD $< 3$ Å, AF2 pLDDT $> 70$.

| | | | Protenix filters | | | | AF2 filters | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Potts | | MPNN | | Potts | | MPNN | |
| PDB | Len | #Lig | $n$ | Succ. | $n$ | Succ. | $n$ | Succ. | $n$ | Succ. |
| 7FCN | 100 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7QAV | 107 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8DKS | 115 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7W18 | 119 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8DJZ | 124 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8DIK | 127 | 1 | 9 | ✓ | 6 | ✓ | 6 | ✓ | 0 | ✗ |
| 7T7H | 144 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7RKB | 146 | 1 | 1 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7ZHD | 149 | 3 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7ROT | 154 | 3 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7VKR | 161 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7VKQ | 166 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7W5I | 167 | 2 | 0 | ✗ | 0 | ✗ | 1 | ✓ | 0 | ✗ |
| 7W5J | 167 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7W5H | 168 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8E83 | 176 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7TW7 | 180 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7TW8 | 180 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7TW9 | 180 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7R8P | 186 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7ZH6 | 195 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7OVU | 196 | 1 | 9 | ✓ | 0 | ✗ | 2 | ✓ | 0 | ✗ |
| 8AJP | 199 | 1 | 9 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7QA0 | 206 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7QA3 | 206 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7ZHA | 206 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7R7O | 209 | 1 | 8 | ✓ | 0 | ✗ | 3 | ✓ | 0 | ✗ |
| 7UQ1 | 210 | 2 | 5 | ✓ | 2 | ✓ | 5 | ✓ | 0 | ✗ |
| 7XQ2 | 210 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8E4Y | 212 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8E5O | 212 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7OAS | 216 | 1 | 8 | ✓ | 2 | ✓ | 0 | ✗ | 0 | ✗ |
| 7VNX | 216 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7NQE | 218 | 3 | 3 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7TVP | 225 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7Z2J | 225 | 1 | 7 | ✓ | 2 | ✓ | 0 | ✗ | 0 | ✗ |
| 7ZVR | 235 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8ETR | 235 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7Z05 | 239 | 1 | 3 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |

*continued on next page*

Table 7 – continued

| | | | Protenix filters | | | | AF2 filters | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Potts | | MPNN | | Potts | | MPNN | |
| PDB | Len | #Lig | $n$ | Succ. | $n$ | Succ. | $n$ | Succ. | $n$ | Succ. |
| 7VSK | 241 | 1 | 1 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7Q3A | 244 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7V8U | 258 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7WRG | 258 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7X15 | 261 | 1 | 2 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8DP7 | 265 | 1 | 2 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7W8F | 278 | 2 | 2 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7FOB | 282 | 1 | 2 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7EUU | 288 | 3 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7EUT | 289 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7PUP | 289 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7UUZ | 294 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7UV0 | 294 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7UUY | 295 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7WZV | 300 | 4 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7WZX | 300 | 4 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7X0B | 300 | 3 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7RIZ | 306 | 2 | 1 | ✓ | 3 | ✓ | 0 | ✗ | 0 | ✗ |
| 7WXP | 311 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7ULC | 315 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7U1O | 320 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7U91 | 321 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7U5A | 323 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7R55 | 324 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7VOB | 327 | 3 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7VOC | 327 | 4 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7WWD | 327 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7WWG | 327 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7WVT | 333 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7ROS | 336 | 1 | 1 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7ROU | 336 | 2 | 1 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8DL3 | 339 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8A9I | 341 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7W6Z | 342 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7Y88 | 354 | 2 | 4 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7XN9 | 363 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7F72 | 370 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7VTL | 373 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7Q6F | 382 | 1 | 8 | ✓ | 1 | ✓ | 0 | ✗ | 0 | ✗ |
| 7Q6I | 382 | 1 | 10 | ✓ | 2 | ✓ | 0 | ✗ | 0 | ✗ |
| 8D89 | 382 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7Q6D | 383 | 3 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7R8Q | 387 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8GZC | 387 | 3 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7Q6G | 388 | 1 | 6 | ✓ | 0 | ✗ | 2 | ✓ | 0 | ✗ |
| 7X0J | 406 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |

*continued on next page*

Table 7 – continued

| | | | Protenix filters | | | | AF2 filters | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Potts | | MPNN | | Potts | | MPNN | |
| PDB | Len | #Lig | $n$ | Succ. | $n$ | Succ. | $n$ | Succ. | $n$ | Succ. |
| 7X0K | 406 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7X0M | 406 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7X0L | 409 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7X0N | 409 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7WBM | 415 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7Z74 | 415 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7Z75 | 416 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7N6O | 433 | 1 | 2 | ✓ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7VU1 | 435 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7VU3 | 435 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7BMF | 444 | 3 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8E19 | 462 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7R6J | 466 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7REV | 466 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7T66 | 466 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7T68 | 466 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7YKN | 476 | 3 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7EXH | 486 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7EXJ | 486 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7EXQ | 486 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7EXR | 486 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7W1J | 494 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7W1I | 497 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7W1L | 497 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7SNJ | 506 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7SNO | 506 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7VUD | 535 | 4 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 7VEW | 609 | 1 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| 8DMF | 718 | 2 | 0 | ✗ | 0 | ✗ | 0 | ✗ | 0 | ✗ |
| **Total** | | | 104 | 23 | 18 | 7 | 19 | 6 | 0 | 0 |

# References


[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

[2] OpenFold Consortium. OpenFold Portal: Datasets. `https://portal.openfold.omsf.io/datasets`, 2026. Accessed 2026-06-16.

[3] The OpenFold3 Team. OpenFold3-preview2 technical report. Technical report, OpenFold Consortium, 2026. URL `https://portal.openfold.omsf.io/reports/of3p2_technical_report.pdf`.

[4] Andy M. Lau, Nicola Bordin, Shaun M. Kandathil, Ian Sillitoe, Vaishali P. Waman, et al. Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Sci-*

*ence*, 2024. URL https://www.biorxiv.org/content/10.1101/2024.03.18.585509v2. Published version of bioRxiv preprint.

[5] Arda Goreci. The unreasonable redundancy of nature's protein folds. *Ligo Biosciences Blog*, May 2026. URL https://research.ligo.bio/posts/unreasonable-redundancy-of-natural-protein-folds/.

[6] cch1999. Pocketeer. https://github.com/cch1999/pocketeer, 2024. Accessed 2026-06-17.

[7] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10(1):168, 2009. doi: 10.1186/1471-2105-10-168.

[8] Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, et al. Atom-level enzyme active site scaffolding using RFdiffusion2. *Nature Methods*, 23:96–105, 2026. doi: 10.1038/s41592-025-02975-x.

[9] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

[10] Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with Genie 2. *arXiv preprint arXiv:2405.15489*, 2024. doi: 10.48550/arXiv.2405.15489.

[11] Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025. doi: 10.48550/arXiv.2503.00710. ICLR 2025.

[12] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.48550/arXiv.2206.00364.

[13] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, Philip J. Y. Leung, Timothy F. Huddy, Samuel Pellock, Doug Tischer, Frederick Chan, Brian Koepnick, Hannah Nguyen, Alex Kang, B. Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

[14] Richard W. Shuai, Tianyu Lu, Subhang Bhatti, Petr Kouba, and Po-Ssu Huang. Ensemble-conditioned protein sequence design with Caliby. *bioRxiv*, 2025. doi: 10.1101/2025.09.30.679633. Preprint.

[15] Jinho Kim, Richard W. Shuai, and Po-Ssu Huang. Two-stage ligand-protein complex sequence design with L-Caliby. In *ICLR 2026 Workshop on Foundation Models for Science*, 2026. URL https://openreview.net/forum?id=NcQZyVzyMf.

[16] Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and David Baker. Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods*, 22:717–723, 2025. doi: 10.1038/s41592-025-02626-1.

[17] John B. Ingraham, Max Baranov, Zachary Costello, et al. Illuminating protein space with a programmable generative model. *Nature*, 623:1070–1078, 2023. doi: 10.1038/s41586-023-06728-8.

[18] Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I. Brent, Yanjing Li, Nathaniel Corley, Paul Kim, Jonathan Funk, Simon Mathis, Saman Salike, Aiko Muraishi, Helen Eisenach, Tuscan Rock Thompson, Jie Chen, Yuliya Politanska, Enisha Sehgal, Brian Coventry, Odin Zhang, Bo Qiang, Kieran Didi, Max Kazman, Frank DiMaio, and David Baker. De novo design of all-atom biomolecular interactions with RFdiffusion3. *bioRxiv*, 2025. doi: 10.1101/2025.09.18.676967.

[19] Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling atomistic protein binder design with generative pretraining and test-time compute. *arXiv preprint arXiv:2603.27950*, 2026. doi: 10.48550/arXiv.2603.27950.

[20] Andy Hsien-Wei Yeh, Christoffer Norn, Yakov Kipnis, Doug Tischer, Samuel J Pellock, Declan Evans, Pengchen Ma, Gyu Rie Lee, Jason Z Zhang, Ivan Anishchenko, et al. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, 2023. doi: 10.1038/s41586-023-05696-3.

[21] Gina El Nesr, Simon L. Dürr, Irimpan I. Mathews, Qi Wen, Kewei Zhao, Ritimukta Sarangi, Ursula Röthlisberger, Fanny Sunden, and Po-Ssu Huang. Zero-shot design of a de novo metalloenzyme. *bioRxiv*, 2026. doi: 10.64898/2026.04.23.720277.

[22] Jason Qian, Lukas F. Milles, Basile I. M. Wicky, Amir Motmaen, Xinting Li, Ryan D. Kibler, Lance Stewart, and David Baker. Accelerating protein design by scaling experimental characterization. *bioRxiv*, 2025. doi: 10.1101/2025.08.05.668824.