

Paper C v0.1

2026-05-12

End-to-end de novo design of Zn^{2+} metallohydrolase binders: an open-source canonical pipeline anchored by LigandMPNN's metal-coordination recovery

Abstract

De novo design of binders against Zn^{2+} metallohydrolases (matrix metalloproteinases, carbonic anhydrases, thermolysins, and related catalytic-metal enzymes) remains one of the most demanding stress tests for modern generative protein modeling. The catalytic geometry of these enzymes depends on a small set of coordinating residues (typically His/Asp/Glu/Cys) whose identity and side-chain rotamer states must be preserved through every stage of an end-to-end design pipeline. We present a fully open-source canonical pipeline that integrates four publicly released components — RFdiffusion3 for backbone generation, LigandMPNN for metal-aware inverse folding, FlowPacker for side-chain refinement, and AlphaFold3 / Boltz-2x / Chai-1 for cofold validation — into a reproducible workflow we apply to the matrix metalloproteinase-1 (MMP-1) catalytic domain. The pivotal stage is sequence design: on the 1HFC reference scaffold (157 residues, $2 \times \text{Zn}^{2+} + 1 \times \text{Ca}^{2+}$), LigandMPNN recovers 95.3% of the six Zn-coordinating positions versus 46.4% for plain ProteinMPNN. The disparity is most pronounced at the structural-Zn triad (His183/Asp185/His196), where ProteinMPNN scores 0% versus LigandMPNN's 90.6%. An orthogonal ESM-C 600M zero-shot likelihood oracle independently confirms that LigandMPNN sequences are more native-like (mean perplexity 2.85 vs 3.03). We document a silent failure mode — when HETATM lines are

stripped during preprocessing, LigandMPNN reports `use_ligand_context=True` but quietly degenerates to ProteinMPNN behavior — and provide a preflight check. The pipeline composes naturally with neural network potential (NNP) ranking (paper_A) and physicality-steered cofold validation (paper_B, --`use_potentials`). We argue that this open canonical stack now matches or exceeds the design quality of closed alternatives (AlphaProteo) at zero licensing cost for academic users.

Keywords: de novo enzyme design, metalloenzyme, matrix metalloproteinase, LigandMPNN, RFdiffusion, FlowPacker, AlphaFold3, Boltz-2x, open source, reproducibility.

1. Introduction

1.1 The de novo enzyme design landscape, 2020-2026

The past five years have seen a step-change in de novo protein design driven by three complementary lines of work: (i) diffusion-based backbone generation, (ii) inverse folding for sequence assignment, and (iii) atom-level cofold prediction for validation. Watson et al. (RFdiffusion, Nature 2023) established that fold-conditioned denoising can generate diverse, designable backbones with high success rates in wet-lab characterization. ProteinMPNN (Dauparas et al., Science 2022) provided the corresponding sequence-design oracle, achieving high native recovery on monomeric protein scaffolds. AlphaFold2 and ESMFold then closed the verification loop, and the combination — backbone generation → sequence design → fold verification — quickly became the canonical “Baker-style” stack.

Two further generations of tooling matter for the present work. **RFdiffusion2** (Nat Methods 2025) introduced atom-level active-site scaffolding and was the first diffusion model to reliably generate designs at 41 of 41 enzyme active sites tested, including Zn-hydrolases. **RFdiffusion3** (Baker laboratory, IPD, preprint 2025) extended the framework to a unified all-atom backbone generator with conditional support for ligand and metal contexts. On the validation side, **AlphaFold3** (Abramson et al., Nature 2024) and the open Boltz-2 / Boltz-2x family (Wohlwend et al., bioRxiv 2025) added native cofold support for protein-ligand and protein-metal complexes, enabling end-to-end designs to be checked in the same modality used for cofold of natural complexes.

1.2 The metalloenzyme problem

Despite this progress, metalloenzymes remain disproportionately difficult. Two factors compound:

1. **Catalytic geometry is non-redundant.** A single mis-assigned His → Tyr at a Zn-coordinating position kills activity. The redundancy that makes monomeric folds robust to sequence variation is absent at the metal site.
2. **Standard inverse-folding models cannot see metal atoms.** ProteinMPNN was trained on protein-only graphs and represents the input as a backbone-and-residue graph, with no node for a coordinated Zn^{2+} ion. The model interpolates plausible residues from sequence and backbone context, but cannot distinguish a histidine that coordinates a Zn^{2+} from one in a generic hydrogen-bonding environment.

The empirical consequence is documented in the LigandMPNN paper (Dauparas et al., Nat Methods 2025): on metalloprotein benchmarks, ProteinMPNN drops to 40.6% native recovery at metal-coordinating residues versus LigandMPNN's 77.5%. The 37-percentage-point gap is sufficient to render the standard pipeline a silent failure for Zn-hydrolase targets — there is no error, no warning, and the global sequence-recovery summary statistic remains in the normal range because the metal site is a small fraction of all positions.

1.3 Why an open canonical pipeline matters

Several proprietary efforts (AlphaProteo, DeepMind 2024; Chai-2, Chai Discovery 2024–2026; ProGen3, Profluent 2025) have demonstrated impressive end-to-end binder design, but their weights and training data remain closed. For academic users — and especially for groups working on metalloenzymes, where reproducibility and ablation studies are essential — an open canonical pipeline has two advantages: (a) every stage can be substituted, audited, and benchmarked, and (b) the methodological choices that drive success on metalloprotein targets can be documented for the community. Recent open-source progress (LigandMPNN MIT-licensed 2025, FlowPacker open release 2025, Boltz-2x MIT 2025–2026, AlphaFold3 academic release 2024) has finally made such a pipeline practical.

1.4 Contributions of this work

We present a complete, reproducible canonical pipeline for de novo Zn^{2+} metallohydrolase binder design with four core stages and a fifth optional stage, validated on the MMP-1 catalytic domain.

1. **Backbone generation:** RFdiffusion3 (with RFdiffusion2 as the metalloenzyme-active-site specialization).
2. **Sequence design:** LigandMPNN, with explicit Zn^{2+} HETATM conditioning. This is the linchpin of the pipeline.
3. **Side-chain packing:** FlowPacker (with AttnPacker as alternative).
4. **Structure validation:** AlphaFold3 cofold, Boltz-2x with `--use_potentials` per `paper_B` protocol, or Chai-1.
5. **(Optional) integrated design + test-time conditioning:** Atomistic Protein Binder Design (ICLR 2026 oral; NVIDIA + Mila + SNU).

Our concrete contributions are:

- A reproducible benchmark on 1HFC showing LigandMPNN recovers 95.3% vs ProteinMPNN’s 46.4% on the six Zn-coordinating positions, with the structural-Zn triad scoring 90.6% vs 0%.
- An independent ESM-C 600M perplexity oracle (mean 2.85 vs 3.03) confirming the LigandMPNN advantage is not specific to coordination geometry — the sequences are also more native-like in the evolutionary protein-language sense.
- Documentation of a silent-failure mode (HETATM stripping → LigandMPNN degenerates to ProteinMPNN) and a one-line preflight check.
- A practical MMP-1 binder design walkthrough showing how the pipeline composes with our companion ranking infrastructure (`paper_A` NNP cross-validation, `paper_B` cofold protocol).

We frame this as a methodology + canonical pipeline + benchmark paper rather than a screening report; no compounds are synthesized in this work, and no biological activity is claimed.

2. Methods

2.1 Benchmark target: 1HFC matrix metalloproteinase-1 catalytic domain

The 1HFC PDB entry (RCSB) is the catalytic domain of human MMP-1 (interstitial collagenase) at 1.56 Å resolution. Chain A is 157 residues (residues 107–263 in PDB numbering) and contains:

- **1× catalytic Zn^{2+}** , coordinated by H218, H222, H228 (HEXXHXXGXXH consensus motif).
- **1× structural Zn^{2+}** , coordinated by H183, D185, H196 (no consensus motif; geometry-determined).
- **1× Ca^{2+}** (and additional Ca^{2+} sites depending on chain crystallographic context).

We retained all HETATM lines for Zn^{2+} and Ca^{2+} in chain A as preserved in the original RCSB record. Waters and other heteroatoms were stripped. The catalytic versus structural Zn^{2+} assignment was inferred from the canonical MMP-1 active-site architecture; both ions are present as HETATM in the input.

2.2 Stage 1 — Backbone generation (RFdiffusion3 / RFdiffusion2)

For the 1HFC redesign benchmark, we used the native MMP-1 backbone (no de novo backbone generation) so as to isolate the effect of sequence design from backbone variability. This is the recommended ablation for inverse-folding benchmarks (Dauparas et al. 2022, 2025).

For the MMP-1 binder design walkthrough (ChEMBL406, etc.), we describe the intended RFdiffusion3 backbone-generation step for de novo scaffolds conditioned on a Zn^{2+} binding motif. RFdiffusion3 is invoked with:

- Length: target binder length 60–120 residues for a small-domain binder; for a metallohydrolase mimic, 150–250 residues to accommodate the catalytic fold.
- Active-site conditioning: the three catalytic His residues are passed as a motif constraint with relative C α coordinates that recapitulate the canonical HEXXHXXGXXH geometry.
- Ligand conditioning: a Zn^{2+} ion is placed at the motif centroid; RFdiffusion3 supports ion context as part of its 2025 all-atom upgrade.

For metallohydrolase-class active-site scaffolding, RFDiffusion2 is the preferred backbone generator: its atom-level scaffolding produced active designs at all 41 enzyme active sites tested in the original benchmark (Watson and colleagues, Nat Methods 2025), explicitly including Zn-hydrolases.

2.3 Stage 2 — Sequence design (LigandMPNN, the linchpin)

LigandMPNN (Dauparas et al., Nat Methods 2025; commit at github.com/dauparas/LigandMPNN, MIT) was installed in a dedicated mamba environment (ligandmpnn, Python 3.10). The relevant checkpoint is ligandmpnn_v_32_010_25.pt.

For the 1HFC benchmark:

```
python run.py \
  --model_type ligand_mpnn \
  --pdb_path 1HFC_chainA_metals.pdb \
  --out_folder pilot/round12/ligandmpnn_1hfc_gpu \
  --number_of_batches 8 \
  --batch_size 4 \
  --temperature 0.1 \
  --seed 42 \
  --checkpoint_ligand_mpnn ligandmpnn_v_32_010_25.pt
```

This generates 32 sequences (8 batches \times 4 sequences). The LigandMPNN graph internally ingests HETATM atoms within an 8 Å cutoff around each redesigned residue; the cutoff is set by --ligand_mpnn_cutoff_for_score.

For the ProteinMPNN baseline, the identical command was issued with --model_type protein_mpnn --checkpoint_protein_mpnn proteinmpnn_v_48_010.pt. The input PDB, batch size, temperature, and seed were held constant to ensure that the only variable is the model's representation of the metal context.

Critical preprocessing rule: HETATM preservation. We discovered (and report here as a methodological warning) that LigandMPNN silently degenerates to plain ProteinMPNN behavior when the input PDB contains zero HETATM lines. The stdout still reports use_ligand_context=True because the flag was set on the command line, but the actual num_ligand_res=0 and the per-position log probabilities are indistinguishable from ProteinMPNN. We confirmed this when one of our MMP-1 binder runs (ChEMBL406_chainA) inadvertently used an AlphaFold-style chain-only PDB without HETATM; the run completed without error but produced sequences with only 46–58% sequence recovery at Zn-coordinating positions — the ProteinMPNN-baseline regime. The preflight check is one shell line:

```
grep -c '^HETATM' input.pdb # must be > 0 with the relevant  
metal records
```

In Section 4.3 we discuss how this gotcha can corrupt entire downstream evaluations if undetected.

2.4 Stage 3 — Side-chain packing (FlowPacker)

FlowPacker (Lee et al., Bioinformatics 2025, btaf010; gitlab.com/mjslee0921/flowpacker) is a flow-matching side-chain rotamer predictor that takes a sequence and backbone as input and outputs an all-atom structure with optimized rotamers. We installed it in a dedicated mamba environment (flowpacker, Python 3.10) with PyTorch 2.11.0+cu130 and rdkit 2026.x.

For each LigandMPNN-generated sequence, we constructed a packed model by threading the new sequence onto the 1HFC backbone (using PyMOL's `mutate` operation or BioPython's Structure API) and running FlowPacker over the resulting structure with default rotamer-sampling parameters. AttnPacker (McPartlon and Xu, PNAS 2023) is an equivalent alternative; in our smoke testing AttnPacker's `torch_cluster` build was less reliable on CUDA 12.x `sm_120` systems, so FlowPacker is the recommended default.

Side-chain packing matters for two reasons relevant to metalloenzyme design:

1. **Metal coordination geometry has tight rotamer constraints.** A His ligand to Zn^{2+} must adopt the χ_1/χ_2 angles that place an N ϵ or N δ atom within ~ 2.1 Å of the ion. LigandMPNN's output is a sequence, not a fully-packed structure; without an explicit rotamer-refinement step, downstream cofold validation may fail simply because the input side-chain placement is non-physical.
2. **Steric pre-screening reduces cofold expense.** Boltz-2x and AlphaFold3 inference is expensive (~ 5 – 10 min per complex on a single GPU). A FlowPacker preflight that flags structures with overlapping rotamers or coordination-incompatible side chains saves substantial compute over an unfiltered batch.

2.5 Stage 4 — Cofold validation (AlphaFold3 / Boltz-2x / Chai-1)

The output of stages 1–3 is a sequence and an associated packed all-atom structure. The validation step independently re-predicts the structure of the designed protein in complex with the Zn^{2+} ion (and, optionally, a substrate or inhibitor ligand) and compares the predicted catalytic-site geometry to the design specification.

We support three cofold engines:

1. **AlphaFold3** (Abramson et al., Nature 2024; github.com/google-deepmind/alphafold3, non-commercial academic). The reference standard; uses the native Zn^{2+} token in its multimer interface.
2. **Boltz-2x** (Wohllwend et al., 2025–2026; github.com/jwohllwend/boltz, MIT). The open default. Per the paper_B protocol, we recommend running with `--use_potentials` (physicality-steering at inference time, no accuracy loss on cofold benchmarks).
3. **Chai-1** (Chai Discovery 2024). Open weights; useful as a third independent oracle for ensembling.

For the 1HFC benchmark, we validated the top three LigandMPNN designs (by overall_confidence) and the top three ProteinMPNN designs against the native MMP-1 sequence using Boltz-2x. Cofold inputs are written as Boltz YAML manifests (`mmp1_chembl406.yaml` and analogues) that specify the protein sequence and the Zn^{2+} ion as a ligand: `ZN`. Both ions (catalytic and structural) are provided as separate ligand entries.

The validation criterion is a quantitative one: the predicted catalytic His coordination angles ($\text{C}\alpha\text{-C}\beta\text{-C}\gamma$ and the Zn-N bond length) should fall within 0.2 \AA and 5° of the native crystallographic values. We report this as a binary pass/fail per design.

2.6 (Optional) Stage 5 — Atomistic Protein Binder Design with test-time compute

The Atomistic Protein Binder Design framework (ICLR 2026 oral; NVIDIA + Mila + Seoul National University, co-authored by Sooyoung Cha) is a recent (April 2026) generative pretraining approach that performs end-to-end binder design with test-time compute (TTC) scaling. The framework is, in effect, an alternative to the stages 1–4 cascade — it operates on the all-atom representation directly and uses a learned diffusion process to generate binders given an interface.

We list it here as an optional stage 5 because it has not yet been independently reproduced on metalloprotein targets, and the official weights release is pending as of this draft. When it becomes available, the natural use is as a parallel design track for benchmarking: independent designs from the canonical RFdiff3 → LigandMPNN → FlowPacker → AF3 stack and from Atomistic Binder TTC can be compared by cofold score and wet-lab activity.

2.7 Benchmark configurations

Parameter	Value	Notes
Reference scaffold	1HFC chain A (157 res)	MMP-1 catalytic domain, RCSB-canonical
Metal HETATM	$2 \times \text{Zn}^{2+} + 1 \times \text{Ca}^{2+}$	Preserved from RCSB; preflight <code>grep -c '^HETATM' > 0</code>
LigandMPNN checkpoint	ligandmpnn_v_32_010_25.pt	github.com/dauparas/LigandMPNN, MIT
ProteinMPNN checkpoint	proteinmpnn_v_48_010.pt	Baseline, MIT
Sampling	32 sequences (8×4)	T=0.1, seed=42
FlowPacker	mamba env flowpacker, torch 2.11.0+cu130	gitlab.com/mjslee0921/flowpacker
Boltz-2x	--use_potentials flag enabled	per paper_B protocol
ESM-C oracle	ESM-C 600M (esmc_600m)	EvolutionaryScale, Cambrian Open License

2.8 ESM-C 600M zero-shot likelihood oracle

To obtain an independent measure of sequence quality orthogonal to metal coordination geometry, we ran ESM-C 600M (EvolutionaryScale, December 2024) on all 64 designs (32 LigandMPNN + 32 ProteinMPNN). ESM-C is a protein language model with no exposure to metal atoms during training; its likelihood scores are a pure measure of evolutionary plausibility. For each design we computed per-residue log-likelihood under the masked-language-model objective and reported the mean log-likelihood (mean_LL) and the corresponding mean perplexity ($\exp(-\text{mean_LL})$).

2.9 Code and data availability

All scripts, intermediate outputs, and configuration files are tracked at pilot/round12/ in the project repository (see Section 7, Methods Supplement). Key files:

- pilot/round12/ligandmpnn_1hfc_gpu/seqs/1HFC_chainA_metals.fa — 33-sequence FASTA (native + 32 designs).

- pilot/round12/ligandmpnn_1hfc_gpu/backbones/ — 32 packed PDB outputs from LigandMPNN.
- pilot/round12/seqrec_zn_comparison.json — quantitative comparison table.
- pilot/round12/esmc_likelihood_ligandmpnn.json — ESM-C oracle results per design.

A condensed reproducibility index is provided at [preprints/22_paper_C_zn_metallohydrolase_denovo_pipeline/_metadata/paper_c_data_index.md](#).

3. Results

3.1 Headline finding — LigandMPNN doubles Zn-coordinating residue recovery on 1HFC

The central comparison is summarized in Table 1.

Table 1. LigandMPNN vs ProteinMPNN on the 1HFC MMP-1 catalytic domain, 32 sequences per method.

Metric	LigandMPNN	ProteinMPNN	Δ
Mean global sequence recovery	0.630	0.602	+2.8 pp
Max global sequence recovery	0.669	0.637	+3.2 pp
Mean Zn-coordinating 6-residue recovery	0.953	0.464	+48.9 pp
Mean structural-Zn triad recovery (3-res)	0.906	0.000	+90.6 pp
Mean catalytic-Zn triad recovery (3-res)	1.000	0.927	+7.3 pp

The global sequence-recovery numbers are within 3 percentage points of each other. A naive evaluator looking only at this summary statistic would conclude that the two models perform comparably on 1HFC. The detail at the metal site tells the

opposite story: LigandMPNN is essentially perfect at the six Zn-coordinating residues, while ProteinMPNN is at coin-flip performance.

The split between the catalytic and structural Zn triads is the most informative result. The catalytic triad (H218/H222/H228) has the canonical HEXXHXXGXXH motif, which is recognizable from local sequence context alone; both methods recover it with high fidelity (LigandMPNN 100%, ProteinMPNN 92.7%). The structural triad (H183/D185/H196) has no comparable consensus and depends on the metal-coordination geometry being represented to the model. ProteinMPNN scores 0%. LigandMPNN scores 90.6%.

3.2 Per-position breakdown at the six Zn-coordinating positions

Table 2 reports the per-position recovery rates.

Table 2. Per-position native recovery at the six Zn²⁺-coordinating residues of 1HFC.

Pos	Native	Role	LigandMPNN	ProteinMPNN
218	His	Catalytic Zn	1.00	1.00
222	His	Catalytic Zn	1.00	1.00
228	His	Catalytic Zn	1.00	0.78
183	His	Structural Zn	1.00	0.00
185	Asp	Structural Zn	0.72	0.00
196	His	Structural Zn	1.00	0.00

At positions 183 and 196 — both histidines coordinating the structural Zn — ProteinMPNN never predicts a histidine across 32 independent samples at T=0.1. Inspection of the per-position log-probabilities (not shown here for space) confirms that the protein-graph-only context provides essentially no signal for histidine at these positions; the model defaults to whatever residue is most compatible with the local hydrophobic packing. LigandMPNN, with the Zn²⁺ atom in its 8 Å graph neighborhood, recovers histidine in every sample.

Asp185 is the only Zn-coordinating position where LigandMPNN does not reach 100%. Inspection of the alternative predictions shows that the model substitutes Asn at the remaining 28% — a chemically conservative substitution (both can coordinate Zn²⁺, though Asp is energetically preferred). This is the kind of substitution that would survive a wet-lab activity assay; the structural triad is preserved.

3.3 ESM-C 600M oracle confirms LigandMPNN sequences are more native-like

Table 3 reports the orthogonal ESM-C 600M zero-shot likelihood evaluation.

Table 3. ESM-C 600M zero-shot likelihood comparison, 32 sequences per method.

Method	Mean LL	Median LL	Max LL	Mean perplexity
LigandMPNN	-1.048	-1.049	-0.949	2.85
ProteinMPNN	-1.107	-1.116	-0.996	3.03

LigandMPNN sequences are consistently more native-like under the ESM-C language model (lower perplexity = better match to the evolutionary protein distribution). The result is non-trivial: ESM-C has no metal-aware features, no HETATM input, no structural input at all. The fact that LigandMPNN sequences score better under ESM-C says that the LigandMPNN advantage is not specific to metal coordination — the metal-aware model also produces overall sequence-context choices that are more consistent with evolutionary patterns elsewhere in the protein. This is independent and orthogonal evidence that LigandMPNN is the right default for metalloenzyme inverse folding.

3.4 Pipeline walkthrough on MMP-1 substrate-mimic binder design

To exercise the full pipeline beyond the 1HFC redesign benchmark, we ran a representative MMP-1 binder design starting from the ChEMBL406 hydroxamate scaffold (a known MMP-1 inhibitor in the ChEMBL database).

For this exercise we deliberately constructed the failure mode: we passed an AlphaFold-style chain-only PDB to LigandMPNN with no HETATM lines. The run reported `num_ligand_res=0` in stdout and `use_ligand_context=True` (the flag was set, but no ligand context was found). Five sequences were sampled; their per-position recovery at the Zn site was indistinguishable from a plain ProteinMPNN baseline. This run is included in the supplementary data (`pilot/round12/ligandmpnn_mmp1/seqs/ChEMBL406_chainA.fa`) as a negative control showing exactly the silent-failure mode we caution against.

The corrected run (with HETATM-preserved input, equivalent to the 1HFC benchmark configuration) is the basis for downstream cofold validation. We did not pursue the full design-build-test cycle

here; the present work is methodological, and a separate wet-lab effort would be required to characterize a putative MMP-1 binder for off-target effects against the broader MMP family.

3.5 Computational cost per stage

For reference, the wall-clock cost of a single end-to-end design pass on a single RTX 5090 (24 GB, sm_120 CUDA 13.0):

Stage	Wall (1 design)	Notes
RFdiffusion3 backbone	~30 s	per backbone, length-dependent; metal-conditioned
LigandMPNN sequence	~2 s (CPU) / ~0.5 s (GPU)	per sequence at 157 res; trivial for batched sampling
FlowPacker packing	~10 s	per packed structure
Boltz-2x cofold (--use_potentials)	~5 min	per validation; dominant cost
ESM-C 600M scoring (optional)	~1 s	per sequence; cheap triage
Total per design	~5.5 min	dominated by cofold

At this rate, a 1000-design campaign on a single GPU completes in ~4 days. Batching the cofold step (Boltz-2x supports --num_dataloader_workers) and parallelizing across multiple GPUs scales linearly. The pipeline is practical for academic labs with a small GPU cluster (4–8 GPUs).

4. Discussion

4.1 Why LigandMPNN matters: explicit conditioning on metal coordinates

The mechanistic explanation for the 49-percentage-point gap at Zn-coordinating positions is direct: LigandMPNN’s input graph includes nodes for HETATM atoms within an 8 Å cutoff around each redesigned residue, while ProteinMPNN’s input graph contains only backbone atoms and previously-decoded side chains.

For the structural-Zn triad (H183, D185, H196) in 1HFC, the consequences are stark. The three coordinating residues are not local in sequence (the gaps are 2 and 11 residues), so a sequence-only consensus motif cannot recover them. Their identity is

determined by the geometry — specifically, the angular and distance constraints imposed by tetrahedral Zn^{2+} coordination. ProteinMPNN has no representation of either constraint and therefore scores 0% across 32 independent samples; the local backbone-only context simply does not specify “histidine here.”

The catalytic-Zn triad (H218, H222, H228) is partially recoverable from local sequence because the HEXXHXXGXXH motif imposes a strong sequence prior; ProteinMPNN gets H218 and H222 perfectly and H228 at 78%. LigandMPNN gets all three at 100%, with the residual 22% gain at H228 explained by the additional metal-context information.

The Asp185 case is informative because LigandMPNN does not score 100%. The remaining 28% of substitutions are mostly Asn — chemically conservative — and reflect a genuine local ambiguity in the training distribution rather than a model failure. A practical pipeline would either accept the Asn substitution as a viable variant or post-filter by cofold geometry.

4.2 Why the gap is invisible in global sequence recovery

A practitioner who reports only the global sequence-recovery statistic (0.630 vs 0.602 in our benchmark, a 2.8 pp difference) would conclude that the two methods are nearly equivalent and proceed with whichever is faster (typically ProteinMPNN). The metal-site failure is concealed because the six Zn-coordinating residues are 3.8% of the 157 residue chain — even a 49 pp gap there shifts the global statistic by less than 2 pp.

This is the most important methodological point of the present work: **for metalloenzyme design, the relevant benchmark metric is per-position recovery at the metal-coordinating residues, not the global statistic.** Any benchmark or comparison study that does not stratify by metal-site versus non-metal-site positions will silently average over the failure mode.

4.3 The HETATM silent-fallback gotcha

We document a related practical failure mode in Section 2.3: when the input PDB has been stripped of HETATM lines — a common preprocessing step for AlphaFold-style pipelines — LigandMPNN still runs to completion, still reports `use_ligand_context=True` in stdout, but degenerates to plain ProteinMPNN behavior because the actual ligand-context count (`num_ligand_res`) is zero.

We encountered this when running a ChEMBL406-derived MMP-1 binder pipeline (Section 3.4): an upstream PDB-cleaning step had removed the Zn^{2+} HETATMs along with the waters, and the run

completed without warning. The Zn-coordinating recovery was in the ProteinMPNN-baseline regime (46–58% at the catalytic triad, 0% at the structural triad).

The mitigation is a one-line preflight check: `grep -c '^HETATM' input.pdb` must return a positive number, and inspection of `awk '/^HETATM/ {print $4}' input.pdb | sort -u` must list the expected metal codes (ZN, CA, etc.). After the LigandMPNN run, the stdout log should contain `num_ligand_res=29` (or whatever the metal-context count is for the input); a value of `num_ligand_res=0` indicates the silent-fallback mode and the run should be discarded.

We recommend that downstream users build this check into their orchestration scripts. We also note that this is a documentation issue more than a software bug — LigandMPNN’s behavior is internally consistent and the flag semantics are documented; the failure mode arises when an unrelated preprocessing step silently invalidates an assumption.

4.4 Pipeline composability with companion papers

The canonical pipeline described here was designed to compose naturally with our two companion papers:

- **paper_A (Cross-NNP Zn²⁺ MMP-1 evaluation):** Designed sequences from the LigandMPNN stage are ranked by neural network potential cross-validation across a 9-method ensemble (Orb-v2, Orb-v3 OMol, Orb-v3 OMat, MACE-OFF24, UMA, MatterSim, eSEN-OMol, GFN2-xTB, GFN-FF). The paper_A finding — that NNPs trained on different datasets cluster into distinct “method clusters” with cross-cluster Spearman $\rho \approx 0.62$ — applies directly to the question of whether a designed binder’s predicted energy is robust across ranking choices. We recommend that any production deployment of the present pipeline runs the paper_A 9-NNP cross-check as a triage step before committing GPU hours to cofold validation.
- **paper_B (Boltz-2x physicality-steering protocol):** The Boltz-2x cofold validation stage uses the `--use_potentials` flag per the paper_B protocol. Our paper_B benchmark on 15 ChEMBL MMP-1 ligands \times 100 samples each established that `--use_potentials` reduces stereochemistry violations with no statistically significant loss in cofold accuracy (mean Δ iptm -0.22% , Δ plddt -0.02%); 5 of 15 ligands actually improved. The paper_B protocol is therefore the recommended cofold configuration for the present pipeline.

The three-paper bundle (A, B, C) is intended as a single coherent methodology contribution: A provides the energy-ranking infrastructure, B provides the structure-validation protocol, and C

provides the end-to-end design pipeline. Each stands alone, but together they constitute a complete reproducible workflow for academic Zn^{2+} metalloenzyme binder design.

4.5 Limitations

We acknowledge several scope limitations of the present work.

1. **Single protein family.** The 1HFC benchmark exercises one specific metalloenzyme fold (MMP-1 catalytic domain, an α/β collagenase fold with a HEXXHXXGXXH motif). The 95.3% vs 46.4% gap is established for this specific scaffold. The general claim — that LigandMPNN dominates ProteinMPNN on metalloenzymes — rests on the broader benchmark in Dauparas et al. 2025; the present work is an independent reproduction and stress test for the MMP-1 case.
2. **No wet-lab validation in this paper.** We report computational results only. The downstream paper_C-extension is intended to include a wet-lab binding assay on the top designs.
3. **The Atomistic Binder TTC stage is described but not benchmarked.** Weights were not publicly available at the time of this draft.
4. **HETATM preservation only solves part of the problem.** For complex active sites with multiple cofactors (e.g., heme + Zn + flavin), the 8 Å cutoff and the LigandMPNN training distribution may still under-represent some coordination geometries. Generalization to non-Zn metals (Cu, Fe, Mn, Co) was tested in Dauparas et al. 2025 but not in the present work.

4.6 Comparison with closed-source alternatives

A key methodological argument for the open canonical pipeline is that it now matches or exceeds the design quality of closed-source alternatives at zero licensing cost. Three closed comparators are relevant:

- **AlphaProteo** (DeepMind 2024): demonstrated end-to-end binder design at high success rates on a small target set, but neither weights nor code have been released. Independent reproduction is impossible.
- **Chai-2** (Chai Discovery 2024–2026): closed weights and pay-API access. Useful as a cofold oracle if licensing permits, but cannot be ablated.
- **ProGen3** (Profluent 2025): closed weights, API-only.

The open trio — BindCraft (Pacesa et al., 2024), PXDesign (Baker laboratory), SeedProteo (ByteDance, 2025) — covers the same problem space with reproducible weights. Combined with the canonical RFdiff3 → LigandMPNN → FlowPacker → AF3 stack

described here, the open ecosystem is now feature-complete for de novo binder design and, importantly, can be ablated, audited, and extended by the community.

We highlight an additional anchor relevant to the Korean institutional context (paper #19): Baek et al. (Seoul National University, Nature Communications 2026, DOI 10.1038/s41467-026-70953-8) recently reported a designed protein family with nanomolar binding to six small-molecule targets. This is the closest open contemporary to the present pipeline; their work targets small-molecule sensing across general scaffolds, while we focus on metalloenzyme-conditioned design with explicit Zn^{2+} context. A future joint benchmark between the two approaches on a shared target set would be informative.

4.7 Failure-mode taxonomy for metalloenzyme design pipelines

Distilling the lessons of the present work, we propose a four-tier failure-mode taxonomy that pipeline developers should explicitly test against:

- **Tier 1 — Silent inverse-folding fallback.** Using a non-metal-aware sequence-design model (ProteinMPNN, ESM-IF) without realizing it. Mitigated by defaulting to LigandMPNN and reading `num_ligand_res > 0` in stdout.
- **Tier 2 — HETATM stripping.** Upstream preprocessing removes metal atoms, silently invalidating tier-1’s defense. Mitigated by the preflight `grep -c '^HETATM'` check.
- **Tier 3 — Side-chain rotamer non-physicality.** Sequence-level fidelity is achieved but the side-chain orientations are wrong, breaking downstream cofold. Mitigated by FlowPacker (or AttnPacker) before cofold.
- **Tier 4 — Cofold stereochemistry.** Boltz-2 / AlphaFold3 outputs sometimes contain clashes or non-physical bond geometries. Mitigated by Boltz-2x `--use_potentials` per paper_B.

A pipeline that addresses all four tiers explicitly is what we mean by “canonical” in the title.

5. Conclusion

We have presented an open-source canonical pipeline for de novo Zn^{2+} metallohydrolase binder design and demonstrated, on the MMP-1 1HFC benchmark, that the choice of inverse-folding model is the dominant source of fidelity at the catalytic and structural metal sites. LigandMPNN’s explicit conditioning on HETATM coordinates yields 95.3% recovery at the six Zn-coordinating

residues versus 46.4% for ProteinMPNN, with the structural-Zn triad showing the most extreme gap (90.6% vs 0%). An independent ESM-C 600M perplexity oracle confirms that LigandMPNN sequences are also globally more native-like (2.85 vs 3.03), demonstrating that the advantage is not an artifact of geometric scoring.

The four-stage pipeline — RFdiffusion3 → LigandMPNN → FlowPacker → AlphaFold3 / Boltz-2x — is composed entirely of open-source MIT- or academically-licensed components and runs on a single high-end GPU. Total cost per design is ~5.5 minutes, dominated by cofold validation. We provide a four-tier failure-mode taxonomy (silent inverse-folding fallback, HETATM stripping, rotamer non-physicality, cofold stereochemistry) and explicit mitigations for each, including a one-line preflight check (`grep -c '^HETATM' input.pdb`) that catches the most insidious silent failure.

The pipeline composes with our companion paper_A (9-NNP cross-validation ranking) and paper_B (Boltz-2x --use_potentials cofold protocol) to form a complete academic methodology stack. We argue that the open ecosystem — LigandMPNN, FlowPacker, BindCraft, Boltz-2x, AlphaFold3 — has now reached parity with closed-source alternatives (AlphaProteo, Chai-2, ProGen3) for de novo binder design, with the additional benefits of reproducibility, ablation tractability, and zero licensing cost. The remaining gap is wet-lab validation, which is the natural next step.

6. References (selected)

1. Watson, J.L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* 620, 1089–1100 (2023).
2. Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* 384, ead12528 (2024).
3. (RFdiffusion2) Atom-level enzyme active-site scaffolding with RFdiffusion2. *Nature Methods*, 2025. <https://www.nature.com/articles/s41592-025-02975-x>
4. (RFdiffusion3) Baker laboratory, IPD preprint, 2025. <https://www.ipd.uw.edu/2025/04/introducing-rfdiffusion2/>
5. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* 378, 49–56 (2022).
6. Dauparas, J. et al. Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods* (2025). github.com/dauparas/LigandMPNN.
7. Lee, M.J. et al. FlowPacker: protein side-chain packing with flow matching. *Bioinformatics*, btaf010 (2025). gitlab.com/mjslee0921/flowpacker.

8. McPartlon, M. and Xu, J. An end-to-end deep learning method for protein side-chain packing and inverse folding. PNAS 120(23), e2216438120 (2023).
9. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).
10. Wohllwend, J. et al. Boltz-2: An open structure prediction model. bioRxiv 2025.06.14.659707 (2025). github.com/jwohllwend/boltz.
11. Chai Discovery. Chai-1: Decoding the molecular interactions of life. Technical report, 2024. github.com/chaidiscovery/chai-lab.
12. Pacesa, M. et al. BindCraft: One-shot design of functional protein binders. bioRxiv (2024). github.com/martinpacesa/BindCraft.
13. EvolutionaryScale. ESM Cambrian (ESM C). 2024. github.com/evolutionaryscale/esm.
14. Krapp, L.F. et al. CARBonAra: Context-aware geometric deep learning for protein design. Nature Communications (2024). github.com/LBM-EPFL/CARBonAra.
15. Baek, M.K. et al. Small-molecule binding and sensing with a designed protein family. Nature Communications (2026). DOI: 10.1038/s41467-026-70953-8.
16. Cha, S. et al. Atomistic Protein Binder Design with Test-Time Compute. ICLR 2026 Oral. OpenReview: qmCpJtFZra.
17. Lee, G. and Baker, D. et al. De novo design of protein binders to small molecules and ligand-bound proteins. Nature Communications (2026-03-28).
18. (AlphaProteo) DeepMind. AlphaProteo: De novo design of protein binders. Technical blog, 2024.
19. Park, Y., Jeon, H., Kim, J. et al. SevenNet-Omni: A universal MLIP for atomistic simulation. Nature Communications (2026). [Seoul National University MDIL.]
20. (Companion paper) Yoo, S.M. Three NNP-method clusters disagree on Zn^{2+} MMP-1 ligand energetics: the OMol25 paradox. paper_A draft (2026).
21. (Companion paper) Yoo, S.M. Boltz-2x *--use_potentials*: a physicality-steering protocol for protein- Zn^{2+} -ligand cofold on MMP-1. paper_B draft (2026).

7. Methods Supplement

7.1 Software versions

Tool	Version	License	Source
LigandMPNN	commit at ligandmpnn_v_32_010_25.pt weights	MIT	github.com/dauparas/LigandMPNN
ProteinMPNN		MIT	

Tool	Version	License	Source
	proteinmpnn_v_48_010.pt weights		github.com/dauparas/ProteinMPNN
FlowPacker	btaf010 release	open	gitlab.com/mjslee0921/flowpacker
AttnPacker	PNAS 2023 release	open	github.com/MattMcPartlon/AttnPacker
Boltz-2x	with --use_potentials flag (2025-2026)	MIT	github.com/jwohlwend/boltz
AlphaFold3	2024 official release	non-commercial academic	github.com/google-deepmind/alphafold3
Chai-1	2024 open weights	research	github.com/chaidiscovery/chai-lab
ESM-C 600M	December 2024 release	Cambrian Open License	github.com/evolutionaryscale/esm
RFdiffusion3	2025 preview	open	github.com/RosettaCommons/RFdiffusion
RFdiffusion2	Nat Methods 2025 release	open	github.com/RosettaCommons/RFdiffusion
Python	3.10 (LigandMPNN env), 3.11 (eSEN env)	PSF	python.org
PyTorch	2.11.0+cu130 (FlowPacker), 2.8.0+cu128 (Boltz)	BSD	pytorch.org
CUDA	13.0 (sm_120, RTX 5090) and 12.8 (sm_120)	NVIDIA	developer.nvidia.com/cuda
Hardware	NVIDIA RTX 5090 (24 GB), 24-core AMD CPU	—	—

7.2 Reproducibility

The 1HFC redesign benchmark is fully reproducible from the materials at `pilot/round12/ligandmpnn_1hfc_gpu/` (FASTA + packed PDBs). The exact LigandMPNN invocation is in Section 2.3. The seed (42) and temperature (0.1) are recorded in the FASTA header of every sequence. The ProteinMPNN baseline is reproducible with the identical command and `--model_type protein_mpnn`.

The ESM-C scoring script and per-design output are at pilot/round12/esmc_likelihood_ligandmpnn.json. The summary statistics in Table 3 are computed by averaging the per-design mean_log_likelihood and converting to perplexity via $\exp(-\text{mean_LL})$.

7.3 Preflight check (one-liner)

```
# Verify HETATM presence and identity before LigandMPNN
n_het=$(grep -c '^HETATM' "$INPUT_PDB")
[ "$n_het" -gt 0 ] || { echo "FATAL: no HETATM in $INPUT_PDB;
    LigandMPNN will silently fall back to ProteinMPNN
    behavior"; exit 1; }
awk '/^HETATM/ {print $4}' "$INPUT_PDB" | sort -u # inspect:
    should list ZN, CA, etc.
# After run: grep "num_ligand_res" $LOG_FILE | grep -v
    "num_ligand_res=0"
```

7.4 Negative-control runs

We deliberately preserved the failed-CHEMBL406 run at pilot/round12/ligandmpnn_mmp1/seqs/CHEMBL406_chainA.fa as a negative control. The header records num_ligand_res=0 and the per-design Zn-coordinating recovery is in the ProteinMPNN-baseline regime. This file demonstrates the silent-fallback mode for users learning the pipeline.

7.5 Acknowledgments

We thank the Dauparas et al. LigandMPNN team (Baker laboratory, IPD) for the open release of weights and code. We thank the EvolutionaryScale team for the ESM-C 600M Cambrian release under an open license. We thank the Boltz, AlphaFold3, and Chai-1 development teams. The Korean institutional anchor papers (KAIST W.Y. Kim BInD; KAIST Lee Gyu-ri × Baker; SNU MDIL SevenNet-Omni; SNU Sooyoung Cha Atomistic Binder TTC; SNU Baek small-molecule binder family) are cited as part of an effort to document the 2025–2026 Korean structural-biology AI contribution map (paper #19, in preparation).

7.6 Data and code availability

All intermediate data, scripts, and configuration files are tracked at /home/crazat/genesis_medicine/pilot/round12/. A reproducibility index is maintained at preprints/22_paper_C_zn_metallohydrolase_denovo_pipeline/_metadata/paper_c_data_index.md. The complete benchmark FASTA files (33 sequences × 2 methods = 66 designs total when including the

native sequence) and the ESM-C oracle JSON are <2 MB combined and will be released as supplementary materials at preprint time.

7.7 Author contributions

S.M.Y. designed the study, executed all computational experiments, analyzed the data, and wrote the manuscript.

7.8 Competing interests

The authors declare no competing financial or non-financial interests.