PROTEIN LOCALIZATION & IDENTIFICATION DIRECTLY FROM CRYOEM MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Cryo-electron microscopy (cryoEM) produces three-dimensional maps of the electrostatic potential of protein complexes. Interpreting a map requires building an atomic model, which in turn requires knowing which protein occupies each region. Automated methods answer this by tracing the protein backbone and identifying side chains, and both steps fail below 5 Å resolution. In this work, we propose a method that assigns proteins to regions of the map without having to build an atomic model. We align a frozen foundation model of cryoEM maps with a frozen protein language model in a shared embedding space, training only two lightweight heads on cached features. Two components are essential: a soft-target contrastive objective that treats solvent as an explicit class and distributes target mass over every chain of a complex, and a rigid-motion-invariant attention module over the sampled voxels of a map. Given a map and the sequences of the proteins it contains, our model assigns regions of the map to their corresponding proteins, outperforming controls. This performance is retained at resolutions where backbone tracing is no longer reliable. When the protein composition is unknown, the same embedding can instead retrieve candidate sequences from a large catalogue, placing the correct protein among the top-ranked candidates in more than 84% of cases. Together, these results show that protein identity can be inferred directly from cryoEM density without first constructing an atomic model.

1 INTRODUCTION

Cryo-electron microscopy (cryoEM) produces three-dimensional maps of the electrostatic potential of biological macromolecules, and now does so routinely at resolutions sufficient for atomic interpretation (Cheng, 2018). A map is not itself a biological result. Its interpretation begins with an atomic model: a set of coordinates that places every amino acid of every chain into the map (Malhotra et al., 2019; Emsley et al., 2010). Subsequent analysis is carried out on the model rather than on the map, for example to determine which residues line a binding site, how a mutation alters a contact, or how a subunit reorients between states, and the model is deposited together with the map. Building a model requires knowing which protein occupies which region of the map.

In most cases this information is available, because the proteins have been purified deliberately and only their arrangement in the map is in question. When it is not, either the purification contains proteins that were not intended (Singh et al., 2020; Carr et al., 2025), or the workflow began with no list of proteins at all and their identities are recovered only after an atomic model has been built (Ho et al., 2020; Su et al., 2021; Schweighauser et al., 2022).

In both settings the input is a map together with a list of protein sequences, and the length of that list determines what has to be done with it. A list at the scale of a proteome has to be searched: thousands of candidates must be reduced to a few that are worth checking individually. Proteins already known to belong to the map have instead to be placed into their corresponding positions in it. We study both, in that order, since the placement starts from the result of the search. In our dataset a complex contains a median of seven distinct sequences and as many as 55, and the 225 held-out maps contain 1,966 distinct sequences in total.

Both questions are currently answered by building an atomic model. Automated model builders trace the protein backbone through the map, predict amino-acid probabilities for each residue, and match the resulting pseudo-sequence against a database (Jamali et al., 2024; Pfab et al., 2021).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

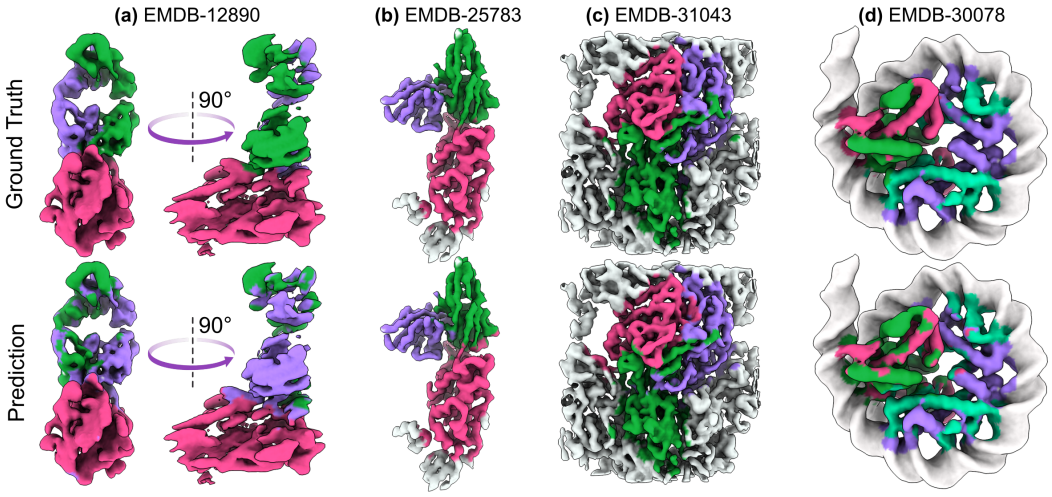

Figure 1: Voxel assignment on four representative test maps, each colored by chain identity for three candidate sequences. Top: the assignment given by the deposited atomic model. Bottom: the prediction, in which each voxel takes the candidate whose residue embeddings are most similar to its own. **(a)** A tetanus toxin fragment bound to a Fab, with strongly anisotropic resolution visible as stretched features in the side view (right). The confused regions are the Fab's heavy and light chains: two immunoglobulin folds of near-equal length arranged with pseudo-twofold symmetry, so that neither shape nor size separates them. **(b, c)** Maps containing chains outside the candidate set (grey), which the method effectively ignores. **(d)** A nucleosome. The histone chains are assigned correctly, and the DNA wrapped around them (grey) is left unassigned, absorbed by the background class. Except in (a), errors are confined largely to the boundaries between chains.

This approach depends on side-chain density being resolved, and degrades as resolution worsens: per-residue identification by a state-of-the-art tracer falls from 49% at 4–5 Å to near zero between 5 and 10 Å (Jamali et al., 2024). The alternative is to predict the structure of each candidate (Jumper et al., 2021; Lin et al., 2023; Abramson et al., 2024) and fit it into the map, or to guide the predictor with the map directly (Raghu et al., 2025); both require an unambiguous fit and scale with the length of the candidate list.

We instead formulate both questions as cross-modal retrieval, by analogy with open-vocabulary segmentation, in which an image is segmented given an open set of text labels. The map takes the role of the image, its voxels the role of pixels, and the candidate sequences the role of labels. Following dino.txt (Jose et al., 2025), which aligns a frozen self-supervised vision backbone (Oquab et al., 2024) with text to obtain pixel-level correspondence, we keep both unimodal models frozen (Zhai et al., 2022) and train only a small head on each side.

As in two-tower retrieval, we encode each modality with a separate network, or *tower*. Our two towers are built on the protein language model ESM-C (EvolutionaryScale Team, 2024; Hayes et al., 2025), and the flow-based model of cryoEM maps CryoFM2 (Zhou et al., 2025) .The language model is a bidirectional encoder, so its hidden states serve directly as sequence features. The map model, however, parameterizes a velocity field over noised maps. The intermediate activations we use as features therefore depend on a chosen noise level, and we find that this choice affects downstream accuracy as much as the choice of layer. To obtain features from clean maps, we distill CryoFM2 into an encoder. Following CleanDIFT (Stracke et al., 2025), a student learns to reproduce the model's features from across the noise schedule in a single deterministic forward pass, removing the dependence on both the general noise level and the particular draw of noise at that level. With both towers frozen, map features are a fixed function of the map and its orientation, so we compute them once and train on the cached features.

We depart from the vision-language recipe in three places as a result of differences in modality and in the amount of available data. First, we supervise the dense term directly. In dino.txt, pixel-level alignment emerges from an image-level objective over $\sim 10^8$ pairs, whereas a leakage-safe split

leaves us only a few hundred independent sequence clusters. Second, because residue boundaries in a cryoEM map are not sharp, we use a soft dense target rather than a hard segmentation mask. Instead of assigning each voxel to its nearest residue, we spread the target over nearby residues with a weight that decays with distance. The remaining weight is given to a learned class representing solvent. Third, because the map features are more local than the neighborhood that defines a residue's environment, we supply spatial context through an attention module biased by pairwise distance.

Our contributions are as follows. (i) We distil CryoFM2 across its noise schedule into a clean-input student, giving a timestep-free feature extractor for cryoEM maps; under a linear readout it beats the teacher at its own best decoupled timestep on seven of eight per-voxel tasks. (ii) We formulate the assignment of proteins to a cryoEM map as dense alignment of two frozen foundation models, reaching 0.65 top-1 against a 0.31 volume prior. (iii) We propose a soft-target, background-aware contrastive objective over all chains of a complex, in which the number of cross-structure negatives is set independently of the batch size. (iv) We introduce VoxMix, a rigid-motion-invariant attention module over the voxels of a map, which raises voxel assignment from 0.40 to 0.65 top-1. (v) We rank all $1{,}966$ test sequences for a map with no shortlist, putting one of its proteins in the top 20 for 84% of maps.

An extended discussion of related work on model building, vision-language alignment and the two towers is given in Appendix A.

2 Method

2.1 Problem formulation

A cryoEM map $\mathcal{M}$ is a reconstruction of a specimen's electrostatic potential, sampled as a scalar volume on a regular grid. The specimen typically contains several distinct molecules, most of them proteins, where each protein is defined by its amino-acid sequence. We address two problems, neither of which uses an atomic model at inference.

Sequence retrieval. Given $\mathcal{M}$ and a database $\mathcal{D}$ of protein sequences, order $\mathcal{D}$ so that the sequences present in the specimen rank highest. In our evaluation set, $\mathcal{D}$ contains $1{,}966$ sequences. At this size, the practical goal is to prune most of the database rather than single out the top-ranked entry with confidence. We therefore measure recall at 20, a list short enough to check by other means.

Voxel-sequence alignment. Given $\mathcal{M}$ and the set $S = \{s_1, \ldots, s_K\}$ of sequences present in the specimen, whether retrieved or known in advance, predict for each sampled voxel a distribution over $S \cup \{\text{BG}\}$, where BG denotes background, including solvent. Here K is small, with a median of seven, so we evaluate the $\arg\max$ directly and report per-voxel accuracy and macro IoU.

We score assignment within-map to eliminate shortcuts based on map-level properties such as size and composition. We score retrieval against both $\mathcal{D}$ and length-matched control pools for each true sequence (Appendix B.2).

2.2 Two frozen towers

Each modality is encoded by its own frozen tower, and trained heads map the outputs of both towers into a shared embedding space (Figure 2).

Map tower. We Fourier-resample each map to 1.5 Å/voxel and normalise it in line with CryoFM2's preprocessing. We then pad and divide each structure into non-overlapping 96 Å cubes which are then processed independently. The encoder is a CleanDIFT student of CryoFM2 (Zhou et al., 2025; Stracke et al., 2025) and therefore needs no noise level. We take features from the second decoder block, which produces 256 channels at 3 Å per voxel and, in a benchmark on real maps with a cluster-level split, gave the best secondary-structure prediction of the blocks we tested.

We train and evaluate on only a select subsample of voxels from each map. We define foreground voxels as those lying within 4 Å of a modelled protein heavy atom and background voxels as those positioned at least 12 Å from every such atom. The shell of voxels between them is excluded from selection. Following pruning, background voxels comprise approximately 18% of total samples. We

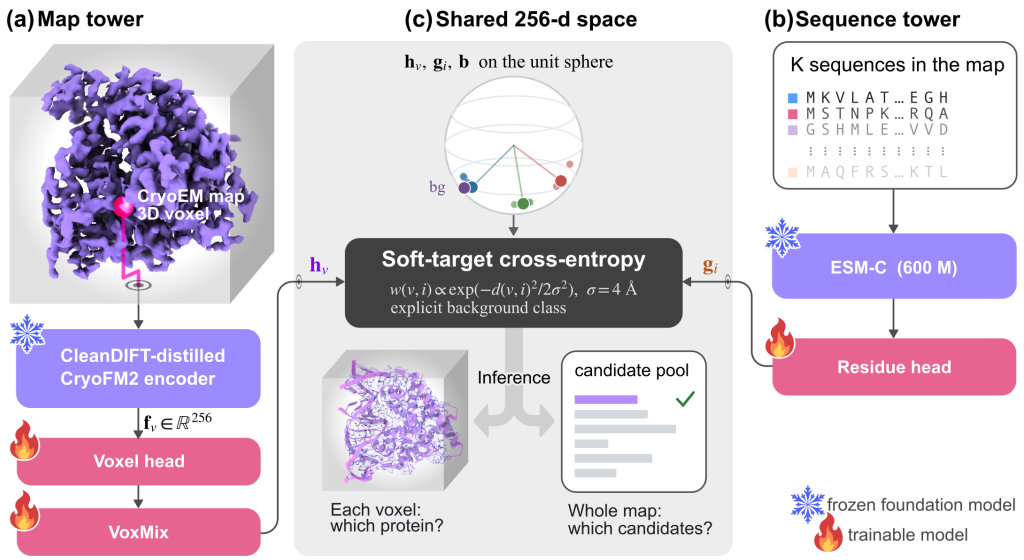

Figure 2: Overview of the method. **(a)** A CleanDIFT-distilled CryoFM2 U-Net encodes the map; features are read at sampled voxels and passed through a small head and a distance-biased mixing module. **(b)** Frozen ESM-C embeds the distinct sequences of the specimen, and a second head projects them. **(c)** Voxel, residue and background embeddings share a unit sphere and are trained with a soft-target cross-entropy. During inference, the same similarities assign each voxel to a protein and rank the candidates for a whole map.

enforce that samples are at least 4.5 Å apart, which is roughly one half-decay length of the features from the base CryoFM2 model (note that this is distinct from the CleanDIFT student). To prevent large chains or complexes from dominating the gradient, we balance the number of selected voxels across maps and weight each voxel in inverse proportion to the number of samples from its chain. We also ensure that every sampled voxel is at least 48 Å from each face of the map to avoid boundary effects. (Appendix D). The atomic model is used to select voxels and to relate them to residues during training and evaluation; Section 4 describes associated limitations.

Sequence tower. We embed each sequence with frozen ESM-C (EvolutionaryScale Team, 2024), taking activations from block 32 of 36 (1152 dimensions per residue, before the final normalisation), and cache them in `fp16`.

Heads. Each trainable head consists of a LayerNorm, a linear projection to $d = 256$ and a residual MLP with a zero-initialised output layer so that the MLP begins training as the identity. On the map side, distance-biased attention (Section 2.3) follows the head. The final components involve a background embedding $\mathbf{b} \in \mathbb{R}^d$ and a CLIP-style temperature, which are both learned (Appendix D). $\mathbf{h}_v$ denotes the embedding of voxel v and $\mathbf{g}_i$ that of residue i. The final voxel, residue, and background embeddings are ℓ_2-normalised, so every score is a scaled cosine similarity.

2.3 VoxMix: distance-biased voxel attention

The map features have a half-decay length of 4–5 Å, whereas the environment of a residue extends to the contact scale of roughly 10 Å. Rather than caching features from a coarser block, we add a module that aggregates the existing features over this larger neighbourhood (Figure 4, Appendix C).

VoxMix is a small self-attention stack over the sampled voxels of a single map. For voxels u, v at positions x_u, x_v, each attention head h adds a learned bias to the attention logit:

$$b_h(u,v) \;=\; \sum_{c=1}^{C} W_{hc}\, \exp\Big(-\gamma\big(\|x_v - x_u\| - \mu_c\big)^2\Big)\,, \tag{1}$$

This is a radial basis expansion of the pairwise distance, as in continuous-filter convolutions (Schütt et al., 2017), with $C = 16$ centres μ_c spanning $[0, 30]$ Å. The bias depends only on distance, and the attention is otherwise standard (Vaswani et al., 2017). We zero-initialise the output projection and final feed-forward layer, so at initialisation the map-side tower reduces to the pointwise head.

Because the bias depends only on pairwise distances and attention is restricted to voxels of the same map, VOXMIX is invariant to rigid motions of the voxel coordinates and never mixes information across maps (Appendix G).

Part of the gain from mixing, however, does not come from learning. Because neighbouring voxels usually belong to the same chain, averaging over a neighbourhood improves accuracy by making predictions locally consistent. We control for this effect in Appendix B.2.

2.4 OBJECTIVE

The dense term is a soft-target cross-entropy anchored on voxels (Figure 3, Appendix C). Residues are indexed by sequence and position, so copies of the same sequence share residues. For voxel v in map j, let $d(v,i)$ be the distance to the nearest heavy atom of residue i in any copy, and let $\mathcal{N}(v)$ be the $K = 8$ residues of map j with the smallest $d(v,i)$ among those within 3σ. The target is

$$w(v,i) \;\propto\; \exp\left(-\frac{d(v,i)^2}{2\sigma^2}\right) \mathbb{1}[i \in \mathcal{N}(v)]\,, \qquad w(v,\mathrm{BG}) \;\propto\; \beta, \tag{2}$$

normalised to sum to one, with $\beta = 0.1$. A background voxel has no residue within 3σ and so places all its mass on BG; a foreground voxel places a median 2% there. We fix $\sigma = 4$ Å, which gives the target a half-decay length of $\sim$4.7 Å, close to the range measured for the corresponding CryoFM2 block (Section 2.3). Writing $\mathcal{C}(v)$ for the set of residues the voxel is scored against, the prediction is

$$p(i \mid v) \;=\; \frac{\exp\big(\mathbf{h}_v^\top \mathbf{g}_i/\tau\big)}{\sum_{i' \in \mathcal{C}(v)} \exp\big(\mathbf{h}_v^\top \mathbf{g}_{i'}/\tau\big) \;+\; \exp\big(\mathbf{h}_v^\top \mathbf{b}/\tau\big)}, \tag{3}$$

with $p(\mathrm{BG} \mid v)$ defined analogously. The denominator then runs over $\mathcal{C}(v)$ and a single background logit, and the loss is $\mathrm{CE}\big(w(v,\cdot), p(\cdot \mid v)\big)$. With a one-hot target and no background class, this reduces to one direction of the CLIP loss.

The objective has three further properties.

Cross-chain targets. We define $w(v,\cdot)$ over the residues of every chain in the map. A voxel on chain B is then evidence for B and evidence against every other chain in the complex. Because a per-chain target is defined over a single chain, it cannot express both. For chain A, it must label the voxel either as background, which is wrong, or as a plain negative, which discards its membership in B. Defining the target over all chains makes a complex of several different proteins a single training example rather than a set of binary segmentations.

Explicit background class. Background voxels, including solvent, place their target mass on the learned embedding $\mathbf{b}$, rather than being characterised only by low similarity to every residue.

Composition of negatives. The set $\mathcal{C}(v)$ determines how much of the denominator comes from the voxel's own map. Under a single softmax over every residue in a batch of M maps, the voxel's own map supplies only about $1/M$ of the candidates. A residue-anchored dual term at weight λ and cluster-disjoint batching complete the objective (Appendix D.2).

3 RESULTS

All results are reported on the 225 test maps. Data, splits, baselines, controls and statistics are given in Appendix B, whose intervals and metrics are used throughout, and inference in Appendix D.1.

3.1 WHAT THE MAP TOWER ENCODES

Before asking whether the towers can be aligned, we ask what the map side knows. We probe the density-side representations at the frozen sites of a per-voxel benchmark with eight labels — the aligned vision tower, the CleanDIFT student at the layer the tower consumes and at its best layer per task, and the undistilled CryoFM2 teacher at its best decoupled timestep — using linear heads whose strength is selected on validation, on a map-level split disjoint from the towers' training maps, with every arm scored on identical rows (Table 1).

Three things are visible. The student beats the teacher on seven of the eight tasks, by 0.070 on molecule class and 0.049 on secondary structure. Removing the timestep was the motive; the representation improved as well, which is why we report the student as a contribution and expect it to serve other readouts of cryoEM density.

Second, alignment to sequence did not erase what the input encoded, and it sharpened the geometric part. The vision tower differs from its own input by about 0.03 at most on any task, and where it moves, it moves towards the properties that separate chains within a map: it is the best density-side arm on secondary structure (0.751), molecule class (0.820) and B-factor (0.710), and gives up a little hydrophobicity (0.134 against 0.146).

Third, amino-acid identity is nearly unreadable from density at these resolutions. Every density-side arm sits between 0.14 and 0.20 on the 20-way task, against a shuffled control at 0.087, and the dense decoder does not change this. The method is built around that fact: per-residue identification is what backbone tracing needs and what these maps do not support, so we assign a chain by aggregating over its residues rather than by identifying them.

The dense-decoder columns qualify this, and the qualification is specific rather than general. EMNUSS (He & Huang, 2021) is a nested 3D U-Net that remains a state-of-the-art architecture for secondary-structure annotation of cryoEM maps; it has some seven million parameters, is trained from scratch for each task on the whole volume, and sees multi-scale context through its skip connections. Given that capacity, the student and the teacher become interchangeable, differing by less than 0.005 on seven of the eight tasks. The two feature sets therefore carry the same information, and the distillation gain under the linear probe is a gain in how accessibly that information is organised, not in how much of it there is. The same columns show what the features add that the decoder cannot recover from density alone: molecule class ($+0.24$), solvent accessibility ($+0.14$ and $+0.15$), secondary structure ($+0.06$) and molecule type ($+0.05$), all properties of a residue's environment — against nothing for amino-acid identity, which stays near 0.16 under every input. B-factor runs the other way, with raw density beating both feature sets (0.369 against 0.30), consistent with a generative prior that normalises away the local amplitude B-factor measures. Our alignment head is light and sits on frozen features, so the linear columns are the ones that apply to it.

3.2 VOXEL-SEQUENCE ALIGNMENT

Table 2 compares the three trained models with the volume prior, which assigns every voxel to the sequence contributing the most residues to the specimen, summed over its copies; it is the strongest model-free baseline we tested. Most of the margin over the prior comes from VOXMIX: a pointwise head exceeds the prior by 0.088 in top-1 accuracy, and eight VOXMIX blocks raise this margin to 0.333, almost four times as large. Doubling the depth to sixteen blocks adds a further 0.007 in top-1 accuracy and 0.010 in macro IoU. Both differences fall within the minimum detectable difference of 0.046, and each depth was trained with a single seed, so we treat the two depths as equivalent in further analyses.

Macro IoU gives a stricter view of the same result. The prior's top-1 accuracy of 0.309 corresponds to a macro IoU of only 0.089, because top-1 accuracy rewards correct prediction of the largest chain, whereas macro-averaging weights every chain equally. Every trained model's margin over the prior is larger under macro IoU than under top-1 accuracy (0.392 against 0.341 at depth 16). The models' accuracy is therefore not confined to the largest chain but extends to the smaller chains, which the prior never predicts.

Where the errors are. Figure 1 shows the assignment on four test maps, and two patterns account for most of what goes wrong. First, errors concentrate at the boundaries between chains. The target

Table 1: What the representations encode, on an eight-label per-voxel benchmark with a map-level split disjoint from the towers' training maps. *Linear probes* are linear heads at frozen sites; *dense decoder* is a 3D U-Net over volumes, a far higher-capacity readout. Metrics are accuracy for *ss* and *aa* and R^2 for *hydro* and *SASA abs* throughout; for the molecule tasks, macro F_1 under the probes and balanced accuracy under the decoder, and for *SASA rel* and *B-factor*, Spearman ρ under the probes and R^2 under the decoder. Numbers are therefore compared within a readout and not across. A shuffled control, which permutes the site labels, scores 0.390 on secondary structure and at or below zero on the regressions. The molecule tasks include nucleotide and heteroatom sites, so the sequence tower does not apply to them.

	linear probes					dense decoder		
task	vision	sequence	student (tap)	student (best)	CryoFM2	student	CryoFM2	density
secondary structure	**0.751**	0.799	0.740	0.748	0.699	**0.800**	0.799	0.740
amino acid	0.143	0.662	0.148	**0.196**	0.180	**0.164**	0.157	0.153
molecule type	0.141	n/a	0.139	**0.181**	0.158	**0.202**	0.197	0.148
molecule class	**0.820**	n/a	0.801	0.799	0.729	0.953	**0.956**	0.714
hydrophobicity	0.134	0.674	**0.146**	0.145	0.129	**0.206**	0.203	0.186
SASA, absolute	0.363	0.446	**0.370**	**0.370**	0.343	**0.651**	0.646	0.510
SASA, relative	0.672	0.699	**0.681**	**0.681**	0.658	**0.747**	0.744	0.596
B-factor	**0.710**	0.492	0.679	0.679	0.670	0.296	0.301	**0.369**

Table 2: Voxel assignment on 225 test maps. Improvements are over the volume prior, which scores 0.3090 top-1 and 0.0891 macro IoU on the same voxels and is the strongest model-free baseline. The remaining floors — an amino-acid classifier, shuffled pairing, an untrained head and a randomly initialised map tower — are given in the text.

	top-1		macro IoU	
arm	accuracy	over prior	IoU	over prior
VoxMix, depth 16	**0.6495**	+0.3405	**0.4813**	+0.3922
VoxMix, depth 8	0.6422	+0.3332	0.4717	+0.3826
pointwise	0.3968	+0.0878	0.2440	+0.1549

spreads over residues within $\sigma = 4$ Å of a voxel and VoxMix attends over a 30 Å radius, so a voxel at an interface carries evidence for both of its neighbours and is decided by a small margin. Second, confusion over whole domains arises when two candidates share a fold. In Figure 1(a) the heavy and light chains of a Fab, two immunoglobulin folds of near-equal length in pseudo-twofold arrangement, are mixed throughout, and neither tower can separate them: two chains with the same fold produce the same density at 3.9 Å, and ESM-C embeddings track fold as well as sequence. This is the within-map form of the homolog problem that cluster-disjoint batching removes from training, and it has no training-side fix, since both chains sit in one map. The strong anisotropy of that map does not itself derail the assignment. What the method gets right is equally telling: chains outside the candidate set, and the DNA of the nucleosome in Figure 1(d), are absorbed by the background class rather than forced onto a candidate, which is what that class exists for and is consistent with the vision tower's molecule-class readout in Table 1.

3.3 Token mixing and its control

Mixing yields the largest single improvement. At depth 2, top-1 accuracy rises from 0.397 to 0.584 and macro IoU from 0.244 to 0.402, a difference of $+0.184$ $[+0.130, +0.242]$. Post-hoc smoothing of the pointwise scores recovers part of it: the best scale by top-1, $\sigma = 12$ Å, reaches 0.472, so spatial coherence alone accounts for 40% of the gain, and mixing still exceeds the best smoothing by $+0.106$ $[+0.057, +0.147]$. By macro IoU the best scale is $\sigma = 8$ Å, smoothing recovers only 31% of the gain, and mixing leads by $+0.110$. Smoothing improves regions that are already correct but cannot produce a small chain that was never predicted, which is what the macro-averaged metric exposes. The shuffled floor also rises under mixing, from 0.235 to 0.289, which measures the same coherence effect a second way, and the volume prior collapses under macro-averaging from 0.309 to 0.089, showing that a third of its top-1 score came from the largest chain (Table 5, Appendix C).

Table 3: Retrieval over the full test pool (1,966 sequences, 225 maps), with VOXMIX at depth 16; all rows come from one training run. Each voxel is matched to its most similar residue in a sequence, and these similarities are reduced over voxels. *Presence scoring* takes the 99th percentile, so a sequence scores highly if any part of the map matches it strongly. *Dominance scoring* takes the mean, which favours sequences that occupy most of the map and under-ranks small chains. Recall@n sets n to each map's true number of sequences; chance is simulated over the pool.

arm	mAP	recall@n	top-1 present
presence scoring	**0.2227**	**0.2029**	0.2044
dominance scoring	0.1726	0.1570	**0.2311**
shuffled pairing	0.0525	0.0428	0.0267

3.4 SEQUENCE RETRIEVAL

We score retrieval over the whole test pool of 1,966 sequences, with no per-map shortlist. The query is the foreground voxels of one map, the ground truth is the set of sequences present in the specimen (a median of seven), and the output is a ranking of the entire pool. We score each sequence by the 99th percentile, over voxels, of the similarity to its best-matching residue, rather than by the mean. The percentile suits a multi-label task: a 50-residue chain in a 2,000-residue complex occupies about 2% of the voxels, so its mean score is dominated by voxels belonging to other chains.

At least one sequence present in the specimen appears in the top 20 for 0.813 of maps, against a chance level of 0.128, and recall at 20 is 0.318, against 0.010. VOXMIX contributes part of this: the pointwise model reaches 0.729 on the first measure, and a two-layer model reaches 0.809. Reducing over a sequence's residues with a log-sum-exp instead of a maximum raises the two measures to 0.831 and 0.339, and an eight-layer model reaches 0.836 and 0.334.

Ranking quality over the whole pool is weaker (Table 3). Mean average precision is 0.2227, about 20 times chance and above every control: shuffled pairing reaches 0.0525, amino-acid composition 0.0214, and length alone 0.0213. The margin over chance is $+0.2116$ $[+0.1651, +0.2950]$, and the margin over shuffled pairing is $+0.1702$ $[+0.1288, +0.2511]$. However, recall at the true count is 0.203, and the top-ranked sequence is present in the specimen for only 0.204 of maps.

Together, these results show that the ranking reliably places a present sequence in a shortlist small enough to check, but does not identify proteins on its own: the top-ranked sequence is wrong for about four maps in five. On a smaller pool of eleven length-matched candidates, the two-layer model of Table 4 ranks the correct sequence first for 0.698 of maps (Table 7, Appendix C), so accuracy depends strongly on the size of the candidate pool.

3.5 RESOLUTION

We examine three separate questions about resolution: whether the relevant information survives band-limiting, whether a model trained at one resolution transfers to others, and how much accuracy a fixed model loses as resolution degrades. All models in this section use VOXMIX at depth 2, except where stated.

A model trained and tested on maps low-passed to 8 Å reaches 0.5851, compared with 0.5831 at native resolution, so band-limiting to 8 Å removes little of the information the task uses; accuracy falls only beyond that, to 0.5135 at 12 Å and 0.4874 at 20 Å. Transfer across resolutions is poor. Table 6 (Appendix C) reports every train–test pair: 15 of the 20 off-diagonal cells fall below the volume prior, and only transfer between 1.5 and 4 Å is useful (0.549 and 0.467); even between 4 and 8 Å a model is at or below the prior. A fixed model is also sensitive within the deposited range: across real maps at 1.5–3.9 Å, top-1 falls by about 0.09 per Å with complex size held fixed, and a native-trained model applied to the same maps low-passed to 8 Å falls to 0.317, level with the prior (0.309).

Training on mixed resolutions removes most of this sensitivity. We sample one of the five resolution caches for each batch; keeping a single resolution per batch prevents in-batch negatives from being distinguished by resolution alone. The resulting model scores 0.4259–0.5323 across the full 1.5–20 Å range, compared with 0.3098 for the prior, and comes within 0.048–0.069 of the model trained

specifically for each resolution. We recommend this mixed-resolution model for practical use, since a map arrives at whatever resolution it was reconstructed at.

On real maps, whose deposited resolutions span 1.5–3.9 Å, the depth-8 model's top-1 falls from 0.692 below 3.0 Å to 0.606 at 3.5–3.9 Å, and macro IoU from 0.535 to 0.418. The margin over the volume prior falls further, but mainly because lower-resolution maps contain fewer distinct sequences and so have a stronger prior; with complex size held fixed, resolution has no significant effect on the margin. Because the corpus contains no maps worse than 3.9 Å, we cannot evaluate on genuinely low-resolution maps (Section 4).

4 LIMITATIONS

Three limitations qualify these results, and Appendix F discusses each together with three smaller ones. Training and evaluation voxels are selected by their distance to modelled heavy atoms, so the full pipeline still depends on an atomic model even though the network does not. Resolutions below the corpus's 1.5–3.9 Å are simulated by band-limiting, which omits the loss of signal-to-noise, the sharpening and the disorder of genuinely low-resolution data. And over the full pool the ranking yields a shortlist rather than an identification: the top-ranked sequence is present in the specimen for about a fifth of maps.

5 CONCLUSION

We have shown that two frozen foundation models, aligned through lightweight trainable heads, can assign proteins to regions of a cryoEM map. Given the sequences present in a specimen (a median of seven), the model assigns 65.0% of voxels correctly, compared with 30.9% for the longest-candidate prior. Given no sequence list, it ranks 1,966 candidates and places a correct sequence in the top 20 for 83.6% of maps at the depth best suited to retrieval. This is a shortlist for downstream verification, for example by structure fitting or mass spectrometry, rather than an identification: the top-ranked sequence is correct for only about a fifth of maps. When trained on band-limited maps, the model retains its accuracy at resolutions as coarse as 8 Å. Models trained at a single resolution do not transfer to other resolutions, but one model trained across resolutions recovers most of this loss from 1.5 to 20 Å. We find that two components contribute most: a soft-target objective that spans every chain and treats solvent as an explicit class, and a distance-biased attention module whose benefit remains after controlling for simple spatial smoothing.

REPRODUCIBILITY STATEMENT

Appendix B gives data construction, the cluster-level split, hyperparameters and the evaluation protocol; Sections 2.4 and 2.3 give the objective and the mixing module, with the remaining details in Appendix D and the self-tests in Appendix G. Code, the distilled map encoder and the trained heads will be released with the camera-ready version, together with the scripts that build the voxel cache and reproduce every table. The weights of both frozen towers, CryoFM2 and ESM-C, are already public.

AI USE STATEMENT

We used generative AI tools for implementation, for testing, and to aid and polish the writing. AI-assisted code was reviewed and tested by the authors, and the experiments it produced are covered by the assertions and self-tests listed in Appendix G. The text was polished with AI assistance, and every sentence was evaluated carefully by an author. We did not use generative AI to generate research ideas or to retrieve or summarise the literature on our behalf. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

REFERENCES

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(8016):493–500, 2024.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z Wu, Jin Sub Lee, Elise S Bruguera, Jehan A Keval, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, pp. 2026–06, 2026.

Kenneth D. Carr, Dane Evan D. Zambrano, Connor Weidle, Alex Goodson, Helen E. Eisenach, Harley Pyles, Alexis Courbet, Neil P. King, and Andrew J. Borst. Protein identification using cryo-EM and artificial intelligence guides improved sample purification. *Journal of Structural Biology: X*, 11:100120, 2025.

Yifan Cheng. Single-particle cryo-EM — how did it get here and where will it go. *Science*, 361 (6405):876–880, 2018.

Grzegorz Chojnowski, Adam J. Simpkin, Diego A. Leonardo, Wolfram Seifert-Davila, Dan E. Vivas-Ruiz, Ronan M. Keegan, and Daniel J. Rigden. findMySequence: a neural-network-based approach for identification of unknown proteins in X-ray crystallography and cryo-EM. *IUCrJ*, 9(1):86–97, 2022.

Paul Emsley, Bernhard Lohkamp, William G. Scott, and Kevin Cowtan. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography*, 66(4):486–501, 2010.

EvolutionaryScale Team. ESM cambrian: Revealing the mysteries of proteins with unsupervised learning. `https://evolutionaryscale.ai/blog/esm-cambrian`, 2024.

Nabin Giri, Liguo Wang, and Jianlin Cheng. Cryo2StructData: A large labeled cryo-EM density map dataset for ai-based modeling of protein structures. *Scientific Data*, 11(1):458, 2024.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Jiahua He and Sheng-You Huang. Emnuss: a deep learning framework for secondary structure annotation in cryo-em maps. *Briefings in bioinformatics*, 22(6):bbab156, 2021.

Chi-Min Ho, Xiaorun Li, Mason Lai, Thomas C. Terwilliger, Josh R. Beck, James Wohlschlegel, Daniel E. Goldberg, Anthony W. P. Fitzpatrick, and Z. Hong Zhou. Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu. *Nature Methods*, 17(1): 79–85, 2020.

Kiarash Jamali, Lukas Käll, Rui Zhang, Alan Brown, Dari Kimanius, and Sjors H. W. Scheres. Automated model building and protein identification in cryo-EM maps. *Nature*, 628(8007): 450–457, 2024.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. DINOv2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.16334.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Sony Malhotra, Sylvain Träger, Matteo Dal Peraro, and Maya Topf. Modelling structures in cryo-EM maps. *Current Opinion in Structural Biology*, 58:105–114, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Jonas Pfab, Nhut Minh Phan, and Dong Si. DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes. *Proceedings of the National Academy of Sciences*, 118(2):e2017525118, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Rishwanth Raghu, Axel Levy, Gordon Wetzstein, and Ellen D. Zhong. Multiscale guidance of protein structure prediction with heterogeneous cryo-EM data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Manuel Schweighauser, Diana Arseni, Mehtap Bacioglu, Melissa Huang, Sofia Lövestam, Yang Shi, Yang Yang, Wenjuan Zhang, Abhay Kotecha, Holly J. Garringer, Ruben Vidal, Grace I. Hallinan, Kathy L. Newell, Airi Tarutani, Shigeo Murayama, Masayuki Miyazaki, Yuko Saito, Mari Yoshida, Kazuko Hasegawa, Tammaryn Lashley, Tamas Revesz, Gabor G. Kovacs, John van Swieten, Masaki Takao, Masato Hasegawa, Bernardino Ghetti, Maria Grazia Spillantini, Benjamin Ryskeldi-Falcon, Alexey G. Murzin, Michel Goedert, and Sjors H. W. Scheres. Age-dependent formation of TMEM106B amyloid filaments in human brains. *Nature*, 605(7909):310–314, 2022.

Kashish Singh, Benjamin Graf, Andreas Linden, Viktor Sautner, Henning Urlaub, Kai Tittmann, Holger Stark, and Ashwin Chari. Discovery of a regulatory subunit of the yeast fatty acid synthase. *Cell*, 180(6):1130–1143.e20, 2020.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. CleanDIFT: Diffusion features without noise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2412.03439.

Chih-Chia Su, Meinan Lyu, Christopher E. Morgan, Jani Reddy Bolla, Carol V. Robinson, and Edward W. Yu. A 'Build and Retrieve' methodology to simultaneously solve cryo-EM structures of membrane proteins. *Nature Methods*, 18(1):69–75, 2021.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Thomas C. Terwilliger, Oleg V. Sobolev, Pavel V. Afonine, and Paul D. Adams. Automated map sharpening by maximization of detail and connectivity. *Acta Crystallographica Section D*, 74(6): 545–559, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Marcin Wojdyr. GEMMI: A library for structural biology. *Journal of Open Source Software*, 7(73): 4200, 2022.

wwPDB consortium. EMDB — the Electron Microscopy Data Bank. *Nucleic Acids Research*, 52 (D1):D456–D465, 2024.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yi Zhou, Yilai Li, Jing Yuan, and Quanquan Gu. CryoFM: A flow-based foundation model for cryo-EM densities. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.08631.

A Related work

ModelAngelo (Jamali et al., 2024) traces the backbone with a graph neural network, predicts residue identity, and matches the result by HMM search, outperforming human experts when resolution is favorable. DeepTracer (Pfab et al., 2021) and Phenix (Terwilliger et al., 2018) also build models directly from density, and findMySequence (Chojnowski et al., 2022) identifies an unknown protein from a built model by searching sequence databases. The second route predicts the structure of each candidate (Jumper et al., 2021; Lin et al., 2023; Abramson et al., 2024) and fits it into the map, most recently by guiding the sampling trajectory of a pretrained predictor with constraints derived from the density (Raghu et al., 2025). Its cost scales with the number of candidates and its discrimination requires an unambiguous fit. Both routes depend on resolved side chains. We target the regime in which side chains are unresolved, and we neither trace the backbone nor fit predicted structures.

Vision-language alignment. CLIP (Radford et al., 2021) and LiT (Zhai et al., 2022) established contrastive alignment of two modalities, the latter with a locked image tower. The closest analog to our setting is dino.txt (Jose et al., 2025), which concatenates the mean over patch tokens with the class token and obtains patch-level alignment from an image-level objective. We retain the two-tower, locked-backbone structure, but do not rely on alignment emerging from a map-level objective, which our sample size does not support. We retrieve with late interaction (Khattab & Zaharia, 2020) rather than with a pooled descriptor, since preliminary experiments indicated that pooling discards information in a map descriptor that is predictable from sequence.

Activations of denoising networks are useful features (Tang et al., 2023), but depend on a chosen noise level; CleanDIFT (Stracke et al., 2025) distills a clean-input student that consolidates them across the noise schedule. CryoFM (Zhou et al., 2025) is a flow-matching (Lipman et al., 2023) prior over cryoEM maps, implemented as a 3D U-Net (Ronneberger et al., 2015). Because its features have the same noise dependence, we apply the CleanDIFT construction as well. ESM-C (EvolutionaryScale Team, 2024) provides per-residue representations that support structure prediction (Candido et al., 2026); we use them frozen, with a trained projection.

B Experimental setup

B.1 Data and splits

We take maps with deposited atomic models from EMDB (wwPDB consortium, 2024) via Cryo2Struct (Giri et al., 2024). For each map we enumerate every polymer chain of at least 20 residues together with its sequence and heavy-atom coordinates (Wojdyr, 2022), deduplicate to distinct sequences, and record chain multiplicity so that homo-oligomeric copies act as multiple positives. The resulting inventory contains $17{,}498$ chains over $1{,}147$ maps, with a median of 9 chains per map (maximum 56), a median of 7 distinct sequences, and 93% of maps containing at least two; $7{,}186$ distinct sequences in total.

We cluster sequences at 30% identity with MMseqs2 (Steinegger & Söding, 2017), obtaining 672 clusters, and split at map level from those clusters, with test $>$ validation $>$ train precedence for the 110 maps that span more than one chain-level split. The map encoder was distilled on the training maps of this split under leakage assertions, so it has seen no validation or test map. We do not re-derive the partition, since any new partition would place validation or test residues in the training set of the encoder. Evaluation uses 225 test maps and about 1.8×10^5 test voxels; those maps hold $1{,}966$ distinct sequences, which is the candidate catalogue the map-level question is scored against.

B.2 Baselines and controls

We evaluate four voxel-level baselines in the same process as the trained model. The *volume prior* assigns every voxel to the longest candidate sequence, which is the size shortcut in its purest form. The *amino-acid classifier* is a per-voxel 20-way classifier on the same features, aggregated to chains by composition matching, and is the representation-space analogue of identify-then-match. *Shuffled pairing* permutes sequences across maps, removing the correspondence and nothing else. An *untrained head* and a *randomly initialised map tower* complete the set; the latter is a head trained from scratch on features from a random-weight encoder, rather than the trained head applied to

random input, which would measure distribution shift instead. It decides whether the result follows from learned features or from architecture and geometry alone, and it is necessary: in preliminary experiments an untrained 3D U-Net reached a pooled top-1 accuracy of 0.58–0.68 against a chance level of 0.10.

The mixing model requires one further control. We apply Gaussian smoothing to the per-sequence scores of the *pointwise* model over the same neighbourhoods, sweeping σ, which measures the improvement available from spatial coherence alone. We select the best σ per metric and require the mixing model to exceed it, rather than merely to exceed the unsmoothed model. The shuffled floor provides a second check: it should increase under mixing by the amount that local consistency helps when the pairing is incorrect.

At the map level the candidate list is the catalogue of all $1{,}966$ distinct sequences in the test maps, and we rank it in full. The floors are a shuffled pairing, a length-only arm scored against a map-side size proxy rather than against the answer, and amino-acid composition. We also report the narrow case used in earlier work, a pool of the eleven catalogue entries closest in length to the answer with cluster-disjoint distractors, in which length is not a sufficient statistic; the length-only control invalidated one of our earlier runs, in which a loosely matched pool made that arm perfect, so we report the realised spread of pool lengths alongside its accuracy.

B.3 Simulating lower resolution

The regime we target is 4–10 Å, so we low-pass filter the cached volumes with a soft raised-cosine Fourier filter at several cut-offs before encoding. Filtering is followed by re-normalisation, because the filter lowers the high percentile by which the normalisation divides: an unrenormalised volume is systematically dimmer, and the experiment would then measure loss of contrast as loss of resolution. Each resolution has its own retrained amino-acid baseline, since the high-resolution classifier would again measure distribution shift, and we verify that the filter changes the features, as an absence of degradation is also the outcome of an experiment that changed nothing (Appendix D).

B.4 Metrics and statistics

We report voxel-level accuracy over foreground voxels, those drawn by the sampler near a heavy atom, so that the background class is trained but not scored. We report per-voxel top-1 accuracy and, because top-1 accuracy rewards any mechanism that produces spatial coherence, per-chain IoU and recall macro-averaged over chains within a map and then over maps. The two measures disagree in an informative way: a volume prior that appears respectable in terms of top-1 accuracy collapses under macro-averaging, which weights a 50-residue chain equally with a $2{,}000$-residue one and removes the credit obtained from the largest chain.

All arms are evaluated in a single process on identical maps, voxels and residues, which we assert to be equal. Differences are reported with 95% paired bootstrap intervals whose resampling unit is the *sequence-cluster component*, not the map: the 225 test maps share 30%-identity clusters and collapse into 89 independent components, the largest two of which hold 53 and 32 maps. Treating maps as independent understates every interval by a factor of 2.2–3.2. The realised minimum detectable difference is accordingly ≈ 0.046 in top-1 accuracy, and we do not interpret differences below it. Hyperparameters are selected on the validation set, and the test set is read once.

Both heads use $d = 256$, a 512-unit hidden layer and one zero-initialised residual block; VoxMix uses two blocks, four heads, 16 RBF centres and $r_{\max} = 30$ Å, raising the trainable count from 0.89 M to 1.95 M. Remaining hyperparameters, cache configuration and preprocessing are in Appendix D.

C Supplementary figures and tables

Table 4: VOXMIX depth, 225 test maps, one protocol and one seed per arm. Validation top-1 is from the training log. Differences between depth 8 and depth 16 are inside the realised minimum detectable difference of 0.046.

depth	val top-1	test top-1	macro IoU	hit@20
2	—	0.584	0.402	0.809
8	0.6682	**0.6422**	**0.4717**	**0.836**
16	0.6857	0.6495	0.4813	0.813

Table 5: Mixing compared with its smoothing control on 225 test maps, one rotation, identical voxels. Smoothing is applied to the per-sequence scores of the pointwise model; σ is given in Å. Macro metrics are averaged over chains within a map and then over maps.

arm	top-1	macro IoU	macro recall
mixing, 2 blocks	**0.584**	**0.402**	**0.552**
pointwise + smoothing, $\sigma = 12$	0.472	0.283	0.436
pointwise + smoothing, $\sigma = 8$	0.467	0.292	0.452
pointwise + smoothing, $\sigma = 20$	0.434	0.227	0.371
pointwise + smoothing, $\sigma = 4$	0.425	0.266	0.427
pointwise	0.397	0.244	0.403
volume prior	0.309	0.089	0.181
mixing, shuffled pairing	0.289	0.163	0.267

Table 6: Transfer across resolution: voxel-level top-1 for models trained at one low-pass cut-off (rows) and tested at another (columns), one process and identical voxels. The volume prior is 0.3098 everywhere. *Mixed* is one model trained by sampling a resolution per batch.

train ↓ / test →	1.5 Å	4 Å	8 Å	12 Å	20 Å
1.5 Å	**0.583**	0.548	0.337	0.270	0.234
4 Å	0.466	**0.591**	0.337	0.260	0.234
8 Å	0.271	0.278	**0.583**	0.300	0.247
12 Å	0.244	0.248	0.284	**0.513**	0.313
20 Å	0.238	0.245	0.247	0.282	**0.488**
mixed	0.530	0.532	0.514	0.465	0.426

Table 7: Map-level ranking on 225 test maps. Top-1 accuracy on a pool of 11 length-matched, cluster-disjoint candidates drawn from the 1,966-sequence catalogue.

arm	top-1	trained − arm
late interaction (mixing model)	**0.698**	—
amino-acid composition	0.173	+0.524 [+0.458, +0.596]
length only	0.098	+0.600 [+0.524, +0.671]
chance	0.091	+0.607 [+0.549, +0.669]

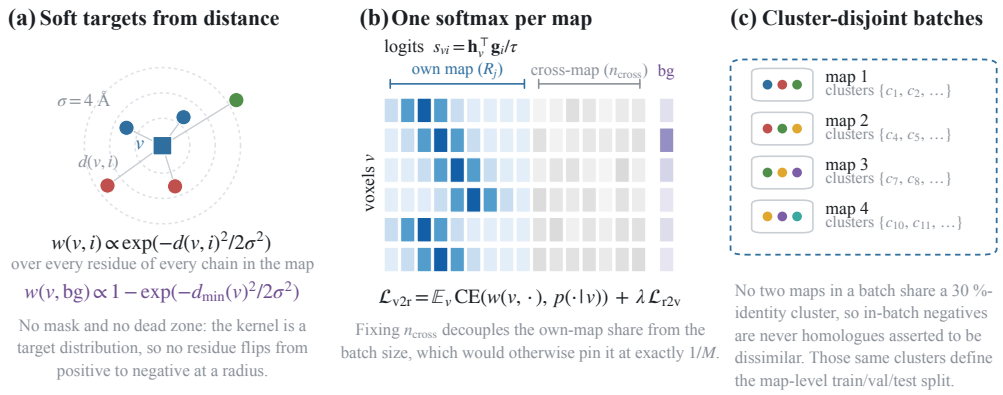

Figure 3: **The objective** (Section 2.4). **(a)** The target over residues is a distance kernel; solvent carries the complementary mass on an explicit background class. Nothing is thresholded, so no residue flips from positive to negative at a radius. **(b)** One softmax per map, over its own residues, a fixed number of residues from the other maps in the batch, and background, so the own-map share of the denominator is a parameter rather than $1/M$. **(c)** No batch holds two maps from one 30%-identity cluster; the same clusters define the map-level split.

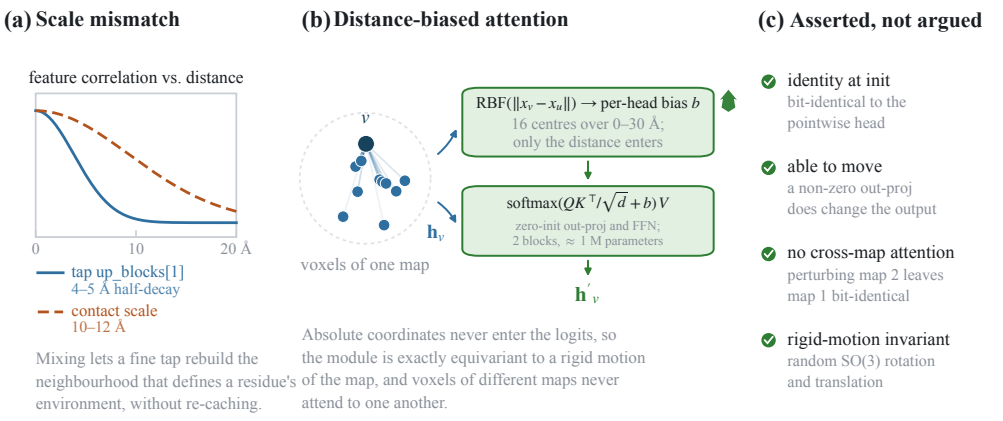

Figure 4: **VoxMix** (Section 2.3). **(a)** The feature tap is finer than the neighbourhood that defines a residue's environment; mixing supplies the missing scale. **(b)** Attention logits receive a learned per-head function of pair distance, expanded in radial basis functions; the module runs over one map at a time with zero-initialised residual branches. **(c)** Four properties checked by unit test.

D IMPLEMENTATION DETAILS

D.1 INFERENCE

To assign a voxel, we score it against every residue of every sequence in S; the score of a sequence is the maximum over its residues, and we assign the voxel to the highest-scoring element of $S \cup \{\text{BG}\}$. For evaluation we score only foreground voxels.

To rank candidate sequences for a whole map, we use late interaction (Khattab & Zaharia, 2020) rather than a pooled map descriptor. Let $m_v(s) = \max_{i \in s} \mathbf{h}_v^\top \mathbf{g}_i$ be voxel v's best match in sequence s, and $\mathcal{V}$ the sampled voxels of map $\mathcal{M}$. For retrieval among a small pool we average over voxels,

$$\text{score}(\mathcal{M}, s) \;=\; \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} m_v(s), \tag{4}$$

and to detect which sequences are present, where a sequence may occupy only a small part of the map, we take the 99th percentile of $m_v(s)$ over $\mathcal{V}$ instead of the mean.

With both towers frozen and the student deterministic, map features are a fixed function of the pair (map, rotation) and are computed once.

D.2 THE DUAL TERM AND CLUSTER-DISJOINT BATCHES

The similarity matrix has voxels along one axis and residues along the other, and the loss above normalises each voxel's row over residues. Normalising each residue's column over voxels instead gives a residue-anchored dual, which we add at weight λ. We keep it secondary for four reasons. First, only the voxel-anchored direction matches what we evaluate. Second, it normalises over a set defined by the data, the residues present, rather than over the voxels we sampled. Third, it is single-valued under homo-oligomeric symmetry: a voxel on one copy of a twelve-fold ring has one correct sequence, whereas a residue has twelve equally correct locations. Fourth, it admits a background class, since a voxel can belong to no residue, whereas every resolved residue must have some corresponding density.

We enforce that no batch contains two maps from the same 30%-identity sequence cluster. The residues of the other maps in a batch serve as negatives, so two homologous maps in one batch would push the embeddings of similar residues apart. Such false negatives would be systematic in this setting, since α-helices look alike across proteins.

D.3 VOXMIX PROPERTIES

We verify four properties by unit test (Appendix G): the output is identical to that of the pointwise head at initialisation; it changes once the output projection is non-zero; no attention crosses maps, so batch composition cannot enter a prediction; and the output is invariant to a random $SO(3)$ rotation and translation of the input coordinates. The last property is required for deployment, since an unknown map is presented in an arbitrary frame, and it is why we store coordinates in the rotated frame: a canonical frame would render any spatial module trivially invariant. A run with mixing therefore draws each batch from a single rotation, and we train a pointwise control under the same sampling regime.

D.4 HEAD ARCHITECTURE

Heads. Each tower has a LayerNorm, a linear projection to $d = 256$, and a residual MLP with a zero-initialised output layer, followed by ℓ_2 normalisation; at step 0 each head is its linear projection. A learned background embedding $\mathbf{b} \in \mathbb{R}^d$ and a temperature τ, parameterised as $\log(1/\tau)$ and clamped to $[0.01, 1]$ (Radford et al., 2021), complete the model. $\mathbf{h}_v$ denotes the embedding of voxel v and $\mathbf{g}_i$ that of residue i.

E ABLATIONS AND ADDITIONAL CONTROLS

We report ablations on the validation maps. *Mixing depth.* One block reaches 0.534 top-1 accuracy and two blocks reach 0.600, against 0.342 for the pointwise model; we did not sweep depth further. *Fixed blur.* Replacing the learned attention with a fixed Gaussian average of the raw features over the voxels of each map, which gives a coarser receptive field without learning, yields a margin over the volume prior of $+0.174$, against $+0.370$ for mixing; applying the blur *underneath* mixing lowers the margin to $+0.332$. The attention therefore requires the unblurred fine features. *Sequence-side pair channel.* An ESMFold-style pair representation on the sequence tower, given relative position, ESM-C outer products, or both, reaches 0.346, 0.347 and 0.345 top-1 accuracy respectively, against 0.349 for the pointwise head. The channel does not help, and the ESM-fed variant is indistinguishable from its position-only control, so what little it contributes is positional.

E.1 MODEL SELECTION

Two rules follow from earlier errors in this project. We never select on training similarity or on feature-matching cosine, which twice mispredicted downstream accuracy, in opposite directions. We monitor representation collapse with the raw off-diagonal cosine and the relative variation rather than with effective rank, which reported 197.7 on a provably degenerate set because it is computed on centred data.

E.2 HYPERPARAMETERS AND CACHE

Training uses AdamW (Loshchilov & Hutter, 2019) with a one-cycle schedule, peak learning rate 3×10^{-4}, weight decay 0.01, 8 maps and 384 voxels per map per step, τ initialised at 0.07, and 12k steps. The voxel cache holds $3{,}000$ foreground voxels per map per rotation over 4 fixed rotations, at 15% background, 3 Å minimum separation, and at most 8 residue targets per voxel.

Voxel coordinates are recovered on CPU rather than rebuilt on GPU. This is sound because the sampler is a pure function of the volume's shape and the atom coordinates and never reads the volume's values, which is also why every low-pass cache shares one coordinate set. The recovery is verified by pushing the coordinates back through the target construction and requiring the cached residue indices and kernel weights to be reproduced exactly.

E.3 THE LOW-PASS LADDER

The raised-cosine filter commutes with the octahedral rotations used in caching, so filtering once before the rotation loop is correct. To confirm that the filter changes the features we report the centred per-voxel cosine between caches and the change in mean feature norm; the sampled voxel geometry is byte-identical across resolutions and so cannot confound the comparison.

E.4 PREPROCESSING

Volumes are Fourier-resampled to an isotropic 1.5 Å/voxel grid, divided by a high percentile of the value distribution, and shifted to the convention the map model expects, with no negative clipping. Boxes are padded to a multiple of the patch size so that per-patch features stitch into a whole-map feature volume aligned to the tap stride. Padding uses the per-map median rather than zero: zero in preprocessed units is a value well above background, so a zero pad would be a slab of signal.

E.5 VOXEL SAMPLING

Foreground voxels are drawn with probability proportional to a kernel of the distance to the nearest heavy atom, an isometry-invariant statistic that cannot bias orientation. Candidates are thinned by greedy Poisson-disk rejection on a spatial hash whose cells are one minimum separation wide, so the separation guarantee holds across cell boundaries. Per-chain quotas choose which candidates are trained on; they do not change the pool, the input or the features.

E.6 ROTATIONS

The cache stores 4 fixed rotations per map as the augmentation. Voxel coordinates are stored in the rotated frame, matching the features. Whenever coordinates are in play, evaluation restricts every arm to one rotation so that all arms score the same voxels.

Pretraining overlap. Of the 225 test maps, 28 appear in the pretraining set of the map model. On the 197 unseen maps, the pointwise model scores 0.379 against a volume prior of 0.300, a margin of $+0.079$ $[+0.032, +0.135]$; on the 28 seen maps the margin is $+0.095$ $[+0.013, +0.186]$. The seen maps are easier for both the model and the prior, whose accuracy is itself higher on them, so we compare margins rather than raw scores. The difference between the two margins is not detectable ($+0.016$ $[-0.084, +0.107]$), although with 28 seen maps the interval is wide.

Mixing depth. Depth is the one axis on which we could keep buying accuracy, and it saturates. Table 4 (Appendix C) scores depths 2, 8 and 16 under one protocol. Depth 8 is a large gain over depth 2 ($+0.058$ top-1), and depth 16 is not distinguishable from depth 8: $+0.007$ on top-1, -0.023 on hit@20, every difference far inside the 0.046 minimum detectable difference, from a single seed per arm. We report depth 16 for voxel assignment, where it is marginally ahead, and depth 8 for retrieval, where it is.

F LIMITATIONS

Voxel selection. Although the network takes only a map and a set of sequences as input, we select both training and evaluation voxels by their distance to modelled heavy atoms. Selecting foreground voxels by thresholding map values at a percentile failed on real maps: protein occupies only 2–17% of the box in the maps we examined, and some maps contain artifacts brighter than protein. The full pipeline therefore still depends on an atomic model, and a model-free way to select foreground voxels remains future work.

Simulated resolution. All maps in our corpus have real resolutions of 1.5–3.9 Å, because maps that cannot be modelled have no deposited model to supervise from. We obtain lower resolutions by band-limiting these maps. This does not reproduce the loss of signal-to-noise, the sharpening, or the disorder that accompany real low-resolution data, so our results in the 4–10 Å range come from degraded high-resolution maps rather than from genuinely low-resolution ones.

Ranking quality. Over the full pool, mean average precision is 0.22, and the top-ranked sequence is present in the specimen for about a fifth of maps. The method is therefore suited to producing a shortlist, which would then need to be confirmed by mass spectrometry, structure fitting, or, where resolution permits, model building.

Aggregation over chains. Voxel assignment aggregates evidence over hundreds of residues per chain. High assignment accuracy is therefore compatible with weak per-residue discrimination, and our results do not measure residue-level identification.

Test-set size. The test maps form 89 independent cluster components, so differences smaller than about 0.046 cannot be reliably detected.

Scope. At resolutions of about 3 Å or better, backbone tracing and structure prediction with fitting work well and are well validated. Our method is intended for lower resolutions, where these approaches become unreliable, and we have evaluated it there only on simulated data.

G ASSERTIONS AND SELF-TESTS

Each of the following is checked in code, and each guards a failure that the reported numbers would not reveal.

- The distilled map encoder returns bitwise-identical features at two nominal timesteps, so the timestep is not a hyperparameter.
- Both heads equal their linear projections at initialisation, and VOXMIX output equals the pointwise head at initialisation.
- VOXMIX output changes once its output projection is non-zero.
- Perturbing one map's voxels leaves another map's outputs bit-identical.
- VOXMIX output is invariant to a random $SO(3)$ rotation and translation of the input coordinates.
- Recovered voxel coordinates reproduce the cached residue indices exactly and the cached kernel weights to zero absolute error, on every map.
- The low-pass filter reduces high-frequency power to numerical zero and commutes with the octahedral rotation group.
- Train, validation and test share no map and no sequence cluster.
- Evaluation arms score identical voxel sets.