PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Void-X: A generative void-filling model for predicting atomic packing in proteins

Jing Yang[a,b], Junying Yuan[a,c,1], and James J. Chou[a,b,1]

Affiliations are included on p. 10.

Contributed by Junying Yuan; received February 26, 2026; accepted May 5, 2026; reviewed by Nir Ben-Tal and Jianyang Zeng

Generative AI algorithms such as the transformer and diffusion models have greatly empowered de novo design of proteins capable of specifically interacting with designated structural sites on another protein. Most of these design methods employ a top–down approach, in which an overall protein shape is generated by an AI model to pack against a given structural site, followed by sequence design to optimize the interaction. Despite being trained on limited protein complex structures available in the database, the top–down approach has yielded encouraging results. Here, we propose a bottom–up approach that generates atom clusters for optimal packing against a specified structured region for informing the design of protein–protein interactions. To this end, we trained a masked discrete diffusion model, named Void-X, that uses the diffusion transformer to learn atomic-level interactions and fill atomic voids in protein interaction interfaces. Void-X was trained using 8.7 million spherical clusters of atoms from experimental structures in the Protein Data Bank. In each cluster, ~70% of the atoms are used as context (or prompt), and ~30% are masked for information recovery (or answer). By training the model with 172 million parameters, Void-X achieves an overall accuracy of 78.3% and 68.2% for intra- and interchain spherical clusters, respectively. Furthermore, we find that information entropy is a reliable indicator of the prediction accuracy for Void-X. This level of performance allows de novo generation of molecular interactions at the atomic level, offering an alternative approach of protein design complementary to the existing ones.

AI | protein design | discrete diffusion model

Significance

Computational protein design is advancing rapidly from designing protein folds to designing specific protein–protein and protein–drug interactions for expediting drug discovery. The majority of the recent design studies employ a top–down approach whereby the overall protein shape is first generated to pack against a given structural site, followed by sequence design to optimize the interaction. Here, we report a different approach wherein an atomic filling model was developed for learning atomic-level interactions. This model, named Void-X, comprises 172 million parameters and achieves a success rate of 53 to 78% depending on the different types of test applications. Void-X represents a prompt-answer model for predicting atomic packing at protein interaction interfaces, with potentially broad application in designing epitope-specific protein binders.

Accurate prediction of protein–protein interaction at structural level and precise manipulation of these interactions via protein design hold great potential in accelerating therapeutic development and solving unmet medical needs. With the rapid advances in protein delivery technologies such as adeno-associated virus (AAV) and mRNA lipid nanoparticle (LNP), it is increasingly feasible to deliver designed proteins, extracellularly or intracellularly, with tissue specificity.

Prediction of protein–protein interaction at atomic level has historically been addressed with molecular docking by either rigid-body methods (1–6) or molecular dynamics (MD) simulation (7–10), based on physicochemical properties such as van der Waals (VDW) forces, electrostatic interactions, hydrogen bonds, and hydrophobic effect. Protein–protein interaction, however, also involves the poorly understood factors such as the intrinsic dynamics of the molecular system (conformational entropy) and its relation with the entropic changes of the solvent and the entire system upon molecular interaction. These dynamics are difficult to model precisely but are known to be crucial (11, 12). For example, using ensembles of protein structures to account for conformational dynamics was demonstrated to be more effective than using a single conformation in the redesign of enzyme and predicting binding affinity (13, 14).

Although the complex dynamics of macromolecular interactions are not fully understood, a number of protein language model (PLM)-based methods have been developed to predict protein–protein interactions using the database relating protein sequence and structure (15–18). A principal advantage of PLMs is their capacity to be trained on extremely large collection of protein sequences (~3 billion), which allows them to capture intricate structural and functional patterns inherent in primary sequences (19–24). The general-purpose PLMs are typically trained on universal databases such as UniProtKB (25) and MGnify (26), but can also be adapted for specialized applications such as antibody and enzyme optimization (27–31). While PLM addresses the protein–protein interaction mostly at sequence level, AlphaFold2 combines multiple sequence alignment (MSA)-based PLM and 3D structure prediction network to predict protein–protein interaction at structural level (32). It has been demonstrated to predict the structures of single-chain proteins with near experimental

Author contributions: J. Yang, J. Yuan, and J.J.C. designed research; J. Yang and J.J.C. performed research; and J. Yang, J. Yuan, and J.J.C. wrote the paper.

Reviewers: N.B.-T., Tel Aviv University; and J.Z., Westlake University.

Competing interest statement: A patent for discrete diffusion model of protein design has been filed.

Copyright © 2026 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: junying_yuan@sioc.ac.cn or james_chou@mail.sioc.ac.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2607035123/-/DCSupplemental.

Published June 9, 2026.

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

accuracy and has been extended to predict multichain protein complexes, significantly improving the accuracy of interface interactions (33). The upgraded AlphaFold3 (34) can achieve an accuracy of 76.6% [DockQ (35) >0.23] for protein complexes.

The above methods for predicting protein–protein interaction or complex structure require the input of known sequences of the binding partners and are thus not sufficient for designing protein binders with unprecedented sequences. Several recent studies have reported innovative approaches to enable de novo design of protein sequence and fold that can specifically interact with a designated structural site of another protein (36–43), and most of these methods utilize generative AI algorithms such as the transformer (44) and diffusion (45, 46) models. For example, RFdiffusion (36), a diffusion-based model, is used for de novo protein backbone generation, after which sequences can be assigned using sequence design methods such as ProteinMPNN (47). Although AlphaFold by itself is insufficient for the purpose of de novo design as it requires the input of known protein sequences, its high predictive power has been used in the BindCraft method to generate error gradient for sequence optimization (38). Furthermore, protein design applications have been tailored for antibody design, demonstrated by RFantibody (39), which is modified from RFdiffusion, as well as Chai-2 (48), Germinal (49), and mBER (50). From the approach used by AlphaFold to that of RFdiffusion, protein structures are used for learning and generating protein–protein complex structures with atomic details. Despite the small data size [~240,000 experimental structures in the Protein Data Bank (PDB)] compared to those typically used by PLMs (~250 millions to ~3 billions), these methods have achieved encouraging success rate, suggesting the remarkable capacity of AI to learn highly convoluted patterns in complex systems.

Inspired by the above generative AI methods, we here explore a different approach in which an atomic filling model (AFM) was developed for learning atomic-level interactions and filling atomic voids in protein interaction interfaces. This model, named Void-X, is a masked discrete diffusion model based on the D3PM framework (51) and built upon the diffusion transformer (DiT) architecture (52). The training of a large AFM is based on the postulate that optimal atomic packing in the context of a stable macromolecular complex depends not only on the local interactions between neighboring atoms but also high-order couplings with distant atoms. For model training, we extracted more than 8 million spherical clusters of atoms from the PDB serving as the learning data. In each of the spherical clusters, ~30% of the peripheral and spatially contiguous atoms are masked for generation while the remaining atoms are regarded as the context. By training a model with 172 million parameters, Void-X achieves an overall accuracy of 78.3% and 68.2% for intra- and interchain spherical clusters, respectively. This level of performance allows de novo generation of local molecular interactions at the atomic level.

Results

The Design of an Atomic Void-Filling Model for Generating Molecular Interactions. An underlying and common feature of intra- and intermolecular interactions in protein is the exclusion of solvent resulting from intimate atomic packing at the sites of interactions. As discussed above, optimal atomic packing depends not only on interactions between neighboring atoms but also on long-range couplings between distant atoms. The multilayered, high-order atomic interactions involve conformational dynamics and/or allostery that often cannot be modeled reliably.

Generative AI models have achieved notable success in domains such as natural language processing and computer vision, where they are trained to model extremely complex data distributions and sample novel instances (53, 54). We hypothesized that this paradigm can be adapted to train a prompt-answer model for predicting atomic packing at protein interaction interfaces. Specifically, we trained the model using spherical clusters of atoms (Fig. 1 *A* and *B*) from experimental structures in the PDB comprising typically of 587 atoms (heavy and hydrogen), in which ~70% are used as context (or prompt) and ~30% are masked for information recovery (or answer). This type of prompt-answer model can be thought of as filling the atomic voids in a network of atomic interactions, and hence given the name Void-X. The void filling can be performed within a structure formed by a single polypeptide chain (as in Fig. 1*A*) or at the interaction interface between two separate polypeptide chains (as in Fig. 1*B*).

Upon successful training of the Void-X model, it would be possible to generate type and spatial distribution of atoms for

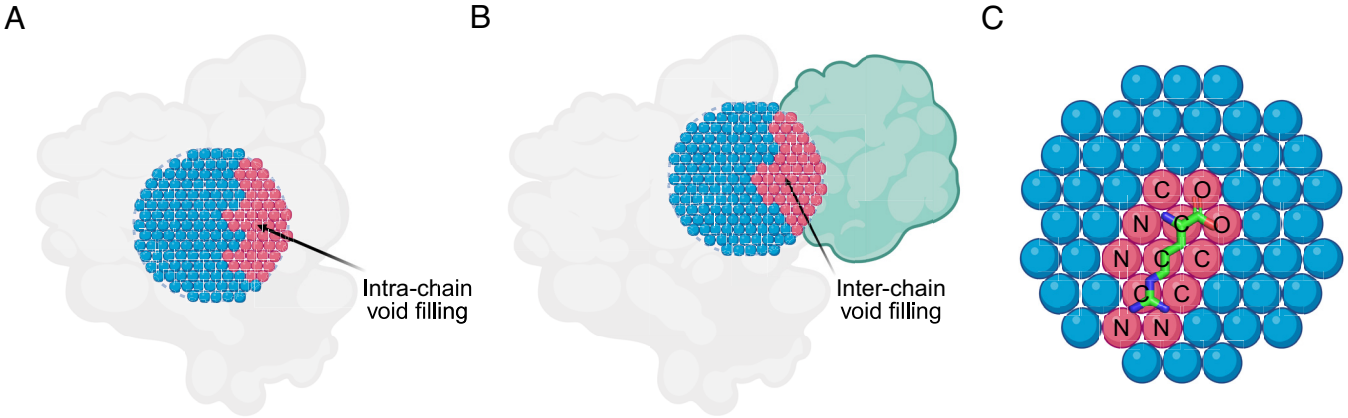


Fig. 1. Schematic illustration of spherical cluster sampling and void filling of atomic interactions. (*A*) Recovery of masked atoms (red) given the context atoms (blue) of a spherical cluster of atoms representing intramolecular interaction patterns in structures formed by a single polypeptide chain. Sampling of the spherical cluster is centered on the C_α atom of a given residue, encompassing all atoms (including hydrogen atoms) within 12 Å radius of the center. Each cluster is constrained to contain 512 to 768 atoms, and the sampled clusters within one protein share no more than 70% overlap. (*B*) Recovery of masked atoms from one polypeptide chain (red) given the context atoms from another polypeptide chain (blue) in a spherical cluster representing intermolecular interaction. The center C_α atom of the cluster is chosen such that the masked atoms account for no more than 30% of the spherical cluster. Clusters are extracted only if they contain atoms from two separate polypeptide chains and at least 96 atoms are from the interacting partner chain. All other sampling parameters are the same as in (*A*). (*C*) A two-dimensional illustration of the uniformization for transforming the atom cluster from real space to the uniform lattice space with hexagonal close-packing. Example in red illustrates uniformization of arginine atoms.

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

optimal binding against a designated structural epitope of a protein. A central challenge to such application, however, arises from the inherently nonuniform spatial distribution of atoms characterized by protein backbone and sidechain geometry, which complicates the initial placement of atoms at the interaction interface. For example, when placing void atoms for packing against a given structural site, consideration of polypeptide bonding geometry inevitably introduces bias in predicting atom types at these voids. To address this problem, we introduced a uniformization step whereby all atoms in a spherical cluster are repositioned to the centers of uniformly sized balls (radius of 1.1 Å) that pack the sphere in a hexagonal close-packed (HCP) mode. The uniformization step transforms the atom cluster in real space to the uniform lattice space (Fig. 1*C* and *SI Appendix*, Fig. S1 *A* and *B*), and it was performed such that the total deviation from the original atomic coordinates is minimized (see *Method* for algorithm). The average RMSD between native and uniformized clusters is 1.2 Å.

Operation in the uniform lattice space allows standardized placement of void atoms against known context atoms in the HCP packing mode. Then prior to model training, the context atoms in the spherical cluster were returned to the original positions (*SI Appendix*, Fig. S1*C*) as their true coordinates better retain protein structural properties that could benefit learning. This procedure provides a practical solution to the problem of atom placement, and hence the application of Void-X to generate protein interaction interface reduces to filling the atomic voids with chemically and structurally probable atoms.

Training the Void-X Model. A key challenge in training a generative model of protein–protein interaction is overcoming the very limited structural data available in the database with only ~160,000 experimental structures of protein–protein interactions. The local spherical sampling strategy partly addresses such limitation by allowing us to extract multiple clusters per protein structure and to further augment sampling by generating multiple context-mask combinations per cluster. Using spherical clusters with a radius of 12 Å, we could extract about 8.7 million clusters of atoms from 189,377 structures of protein and protein–protein complexes (*Method*). In this dataset, the ratio of intra- to interchain clusters is approximately 5:1 (*SI Appendix*, Table S1). For each of the intrachain clusters, training samples were prepared by randomly masking ~30% of the cluster atoms that are peripheral and spatially contiguous at the residue level, yielding an average of 18 augmented samples per cluster (*SI Appendix*, Fig. S2). For interchain clusters, the context and masked atoms were from two separate polypeptide chains, respectively, with masked atoms accounting for no more than 30% of the spherical cluster. Although interchain clusters are less frequent, about 17% of all cluster dataset, Void-X significantly benefited from training on the more abundant intrachain clusters, as described in the following section. All together, 133.8 million context-mask clusters were used for model training.

The Void-X algorithm (Fig. 2) begins with the uniformization of each spherical cluster of atoms, followed by the selection of the masked region and subsequent restoration of the atom positions of the context region. The preprocessed clusters are then passed to the AFM for training atomic void filling of the masked region given a specific atom arrangement of the context region. In contrast to existing diffusion-based protein design methods (36, 37, 39–41), which primarily operate in continuous coordinate space, the AFM denoises atom types in a discrete state space (51), where the DiT architecture (52) was adapted to learn the reverse process (*SI Appendix*, Fig. S3*A* and Algorithm S1).

Void-X employs atomic-level tokens representing 37 heavy atoms and 13 hydrogen atoms (*SI Appendix*, Table S2) to enable all-atom interaction modeling. To compromise computational efficiency with the capacity for global attention over distant atoms, the maximum size of spherical clusters was limited to 768 atoms during training. At each training step, native atomic packing within the masked region were corrupted for up to 500 discrete diffusion steps using a uniform transition matrix, and the AFM with 172 million parameters was trained to recover the original uncorrupted atom

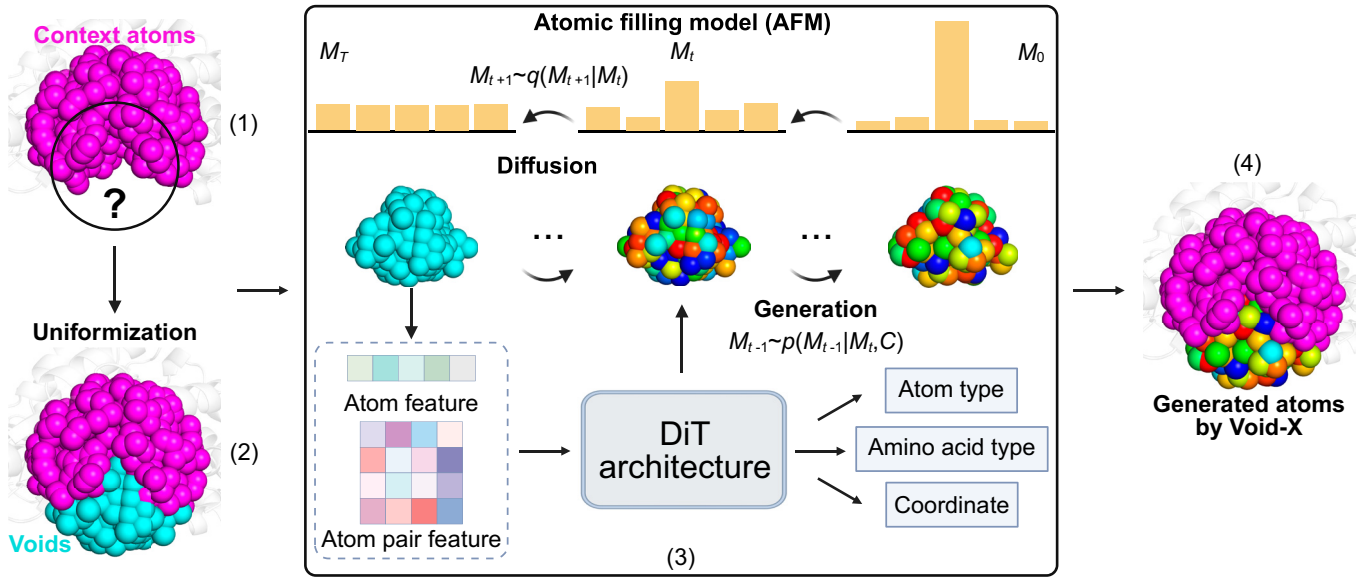


Fig. 2. Flow diagram of Void-X model training for atomic void filling. (1) Within a spherical cluster of atoms in the training dataset, a contiguous region is masked (*M*, cyan), and the remaining part serves as the structural context (*C*, magenta). (2) Perform uniformization that transforms the atom cluster from real space to the uniform lattice space wherein the atoms are represented as uniform spheres (radius 1.1 Å) and packed in the HCP mode; return the atoms in the context region to their real space positions. (3) Perform the diffusion algorithm for training the AFM based on the D3PM framework (51) and built upon the DiT architecture (52). During the diffusion (noising) process (right to left), noise is progressively added to the atoms in the masked region, driving their type distribution toward a uniform distribution over 50 discrete atom types. During generation, atoms are uniformly sampled across these categories, and then iteratively denoised (left to right) to recover atom information of the masked region, including atom type, amino acid type, and coordinate. (4) Once trained, Void-X can generate atoms given a prompt of known structural epitope for protein–protein interaction design.

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

types and corresponding coordinates in real space, as well as the amino acid types of all the masked atoms. The losses considered during the optimization include those for atom type, amino acid type, interatomic distance, atomic coordinate, and the Kullback–Leibler (KL) divergence between the learned generative distribution of denoising process and the categorical posterior distribution of noising process (*Method*). For atom type, amino acid type, and interatomic distance (assigned to 64 bins of 0.375 Å in width), we used categorical cross-entropy loss to measure the difference between prediction and ground truth. For atomic coordinate, the smooth L1 loss (55) was applied to penalize displacement errors. The total loss was computed as the sum of all individual loss components, and the training process reached 4.1 million steps (30 epochs).

The trained neural network comprises branches that can be used to generate atom type, residue type, and their corresponding coordinates in real space (*SI Appendix*, Fig. S3 *B–E*). Within the DiT architecture, we incorporated a multilayer perceptron (MLP) to learn complex transformation patterns between the uniform lattice space and real space. Thus, as the last component of Void-X, the generated atoms in the uniform lattice space were transformed to the real space compatible with protein structure geometry, a step referred to here as coordinate refinement.

Evaluation of the Atomic Void-Filling Model. To evaluate the performance of the trained model in predicting the masked atoms that pack against the context atoms, we first tested the model on 2,893 intraprotein spherical clusters not included for training (Fig. 3*A* and *SI Appendix*, Algorithm S2). The most direct accuracy metric for a discrete atom generation model is the fraction of atoms in a masked region (R) with correctly assigned atom type (AT), formulated below in Eq. **1** as

$$ACC_R^{AT} = n_{AT}/N, \quad [1]$$

where n_{AT} is the number of generated atoms in the masked region with true atom type and N is the total number of generated atoms. Similarly, the amino acid (AA)-type accuracy was computed as $ACC_R^{AA} = n_{AA}/N$, where n_{AA} is the number of generated atoms in the masked region with true AA type. Void-X achieves an overall atom-type accuracy of 78.3% (Fig. 3*B*) and a corresponding AA-type accuracy of 72.9% (Fig. 3*C*).

Furthermore, we introduced a coordinate accuracy metric in the form of spatial deviation of generated atoms from their ground truth positions. Specifically, spatial deviation for a masked region is quantified by RMSD, defined in Eq. **2** as

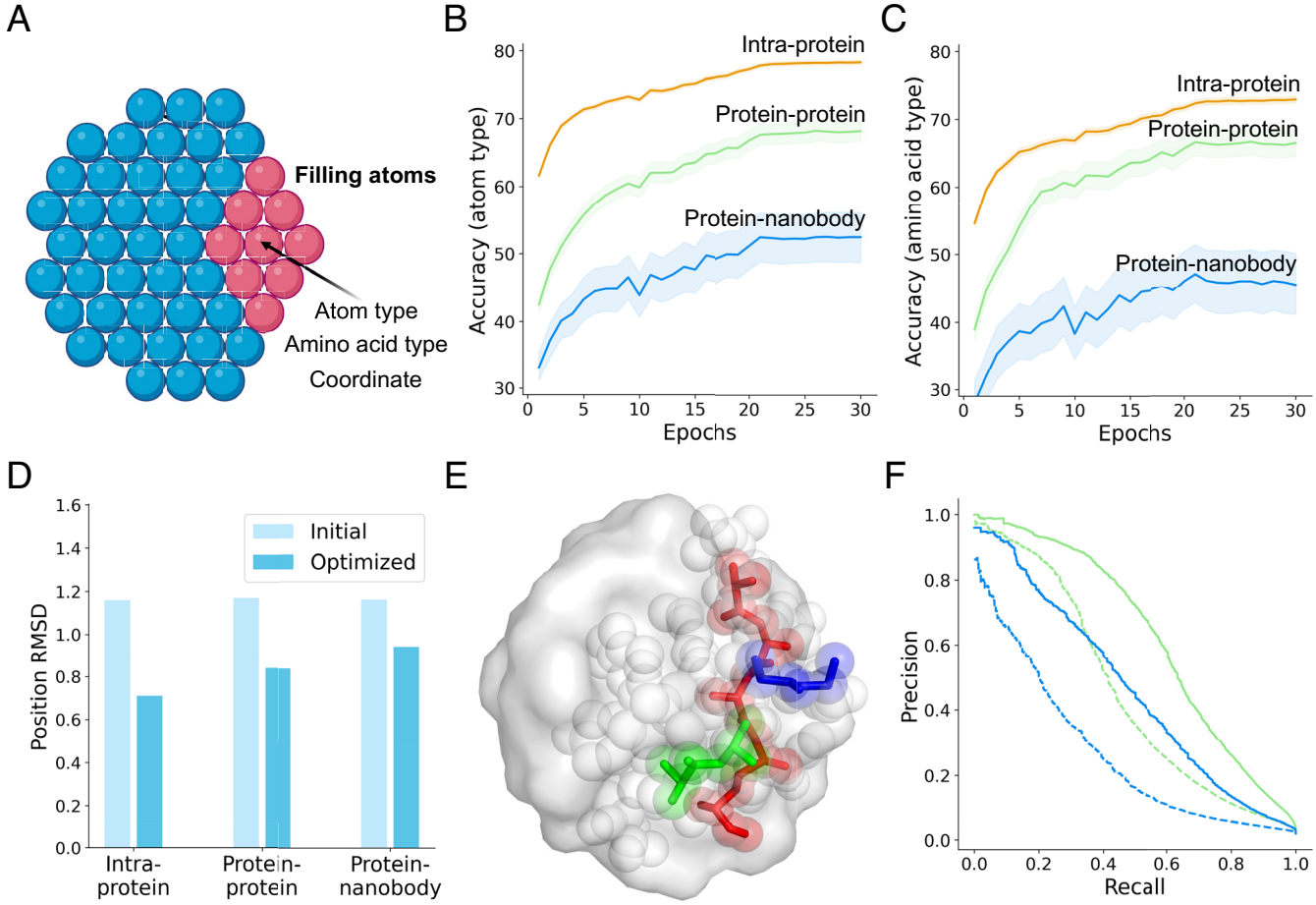


Fig. 3. Evaluation of Void-X for filling atomic interactions in proteins and protein–protein interfaces. (*A*) Each atomic void filled by Void-X contains information about atom type, amino acid type, and coordinate, which are evaluated against the ground truth. (*B*) Overall prediction accuracy (defined in Eq. **1**) for atom type at different training epochs. Shadow regions indicate 95% CI estimated from 10,000 bootstrap samples. (*C*) Same as in (*B*) except the y-axis represents the overall prediction accuracy for amino acid type to which the atoms belong. (*D*) Overall position RMS deviation of the recovered atoms in the uniform lattice space (initial) or after coordinate refinement (optimized) from the ground truth atoms in real space of the masked regions. (*E*) An example of atomic void filling of a protein–nanobody interface (PDB ID: 8OZB) showing atoms assigned by Void-X form 3 AA fragments including HVYDH (red), VE (green), and ES (blue). (*F*) Precision–recall curves for Void-X performance on test datasets of protein–protein clusters (green) and protein–nanobody clusters (blue). The solid and dotted lines represent models trained with dataset that, respectively, included and excluded intrachain clusters. Each precision–recall curve represents the average performance across all atom types, with precision and recall computed independently for each type prior to averaging.

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

$$\mathrm{RMSD} = \left(\frac{1}{N} \sum_{i=1}^{N} \left\| \vec{\boldsymbol{r}}_i - \vec{\boldsymbol{r}}_i^{\,0} \right\|^2 \right)^{1/2}, \quad [2]$$

where $\vec{\boldsymbol{r}}_i$ and $\vec{\boldsymbol{r}}_i^{\,0}$ are positions of the generated atom i after coordinate refinement and its corresponding ground truth, respectively, and N is the total number of atoms in the masked region. We found that the overall position RMSD of the generated atoms before and after coordinate refinement are 1.16 Å and 0.71 Å, respectively (Fig. 3*D*), demonstrating the ability of the model to perform coordinate transformation from the HCP lattice to real space without using any known structural information.

In addition to accuracy calculation, we computed the mean average precision (mAP) by weighing each atom type equally, where average precision (AP) is the area under the precision–recall (PR) curve (56). A larger mAP value indicates both high precision and high recall. In the intramolecular setting, the overall mAP is 0.70, with better performance on backbone atoms than on sidechain atoms (*SI Appendix*, Table S3), consistent with the greater conformational flexibility of side chains in protein structures. The relatively high accuracy of Void-X in intramolecular atom completion suggests its potential for applications such as filling structural voids or unresolved regions in electron density maps during protein structure determination.

We then tested the model performance for predicting the missing atoms that define interchain molecular interactions using 560 protein–protein spherical clusters not included in the training data. In this case, Void-X achieves an overall atom-type accuracy of 68.2%, with a corresponding AA-type accuracy of 66.5%. The overall position RMSD upon coordinate refinement decreased from 1.17 Å to 0.84 Å, and the mAP is 0.63 (Fig. 3). The substantially lower accuracy is likely due to the smaller number of interchain spherical clusters available for training than the intrachain clusters. Consistent with previous results, the mAP for backbone atoms is higher than that for sidechain atoms (*SI Appendix*, Table S3).

A subset of interchain clusters used for model training were from antigen–antibody complex structures. We thus also examined the potential of Void-X in generating protein-antibody binding interfaces by testing atom filling for 87 protein–nanobody spherical clusters not included in the training data. The void-filling accuracy further decreased to 52.6% for atom type and 45.6% for AA type. The overall position RMSD upon coordinate refinement decreased from 1.16 Å to 0.94 Å, and the mAP is 0.46 (Fig. 3 and *SI Appendix*, Table S3). The lower prediction success rate is consistent with the performance of general protein structure prediction methods (33, 34), for which antigen–antibody interface prediction remains particularly challenging due to the limited number of distinct protein–antibody complex structures available for training. Nonetheless, at ~50% prediction accuracy of atom and AA types, Void-X could generate interacting atoms for designing antibody complementarity-determining regions (CDRs) for specific binding to designated structural epitopes of antigens (an example shown in Fig. 3*E*).

Although Void-X was primarily trained on intrachain spherical clusters, its performance for interchain clusters is also quite strong. To estimate the contribution of intrachain data, we trained an alternative AFM with intrachain spherical clusters excluded from the training data. This ablated model achieved an interchain atom-type and AA-type accuracies of 58.3% and 51.8%, respectively, about 10% lower than that of the AFM trained with intra- and interchain data combined. On the protein–nanobody test dataset, the overall accuracy also declined to 38.2% for atom type and 29.1% for AA type when the intrachain data were excluded. Correspondingly, the mAP dropped from 0.63 to 0.45 for the protein–protein test dataset and from 0.46 to 0.26 for the protein–nanobody test dataset (Fig. 3*F* and *SI Appendix*, Table S4). The significant performance gap between the ablated and fully trained models demonstrates that incorporating sufficient intrachain spherical clusters during training improves the model's ability to predict interchain interaction interfaces. This improvement could be attributed to the shared physical principles that govern atomic-level interactions in both intra- and interchain contexts.

To further evaluate the quality of atomic packings generated by Void-X, we present illustrative cases in *SI Appendix*, Fig. S4. These examples show a few continuous fragments of protein backbone as well as residues with complete sidechain structures. Notably, aromatic residues are frequently recovered with full sidechain structures. This level of atomic detail is critical for downstream protein design applications, as accurate sidechain packing at protein binding interfaces directly informs the design of specific and functional molecular interactions.

Indicator of Void-X Prediction Performance. We next investigated the confidence level of Void-X in de novo generation of atoms at the protein–protein interaction interface with the aim of identifying a reliable indicator of prediction accuracy. To this end, we employed information entropy as a confidence measure for the atomic packing generated by Void-X. For each spherical cluster from the test dataset containing both context and masked atoms, we performed 100 independent atomic void fillings of the masked region (*SI Appendix*, Algorithm S2). One inference typically takes 10 s on an H100 GPU. The information entropy in AT assignment for each of the sites i of the masked region is defined in Eq. **3** as

$$S_i^{AT} = -\sum_{j=1}^{K} p\left(x_j\right) ln p\left(x_j\right), \quad [3]$$

where $p\left(x_j\right)$ is the fraction of assignments at the void site i with atom type j in the 100 generated samples and $p\left(x_j\right) ln p\left(x_j\right)$ is summed over $K = 50$ unique atomic-level tokens. The corresponding prediction accuracy of site i, A_i^{AT}, is the fraction of the 100 samples with correct atom type, formulated below in Eq. **4** as

$$A_i^{AT} = N_{AT}/100, \quad [4]$$

where N_{AT} is the number of samples with true atom type. In general, an atom site with lower S_i^{AT} shows higher prediction accuracy and this trend is consistent throughout intra- and interchain clusters from the test dataset (a few samples shown in Fig. 4 *A*–*C*). Meanwhile, an atom site in proximity to context atoms has lower S_i^{AT} (*SI Appendix*, Fig. S5). The generated atoms in direct contact with the context region generally exhibit low entropy and high accuracy, which is expected based on the more stringent rules of packing between neighboring atoms. This is, however, not always the case as some of the masked region sites distant from the context region also exhibited low entropy, suggesting the presence of long-range interactions in atomic packing.

We next examined entropy versus accuracy for the entire masked regions using 1,000 randomly selected spherical clusters from the intraprotein test dataset, 500 from the protein–protein test dataset, and all 87 available clusters from the protein–nanobody test dataset. As above, for each cluster, the masked region

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

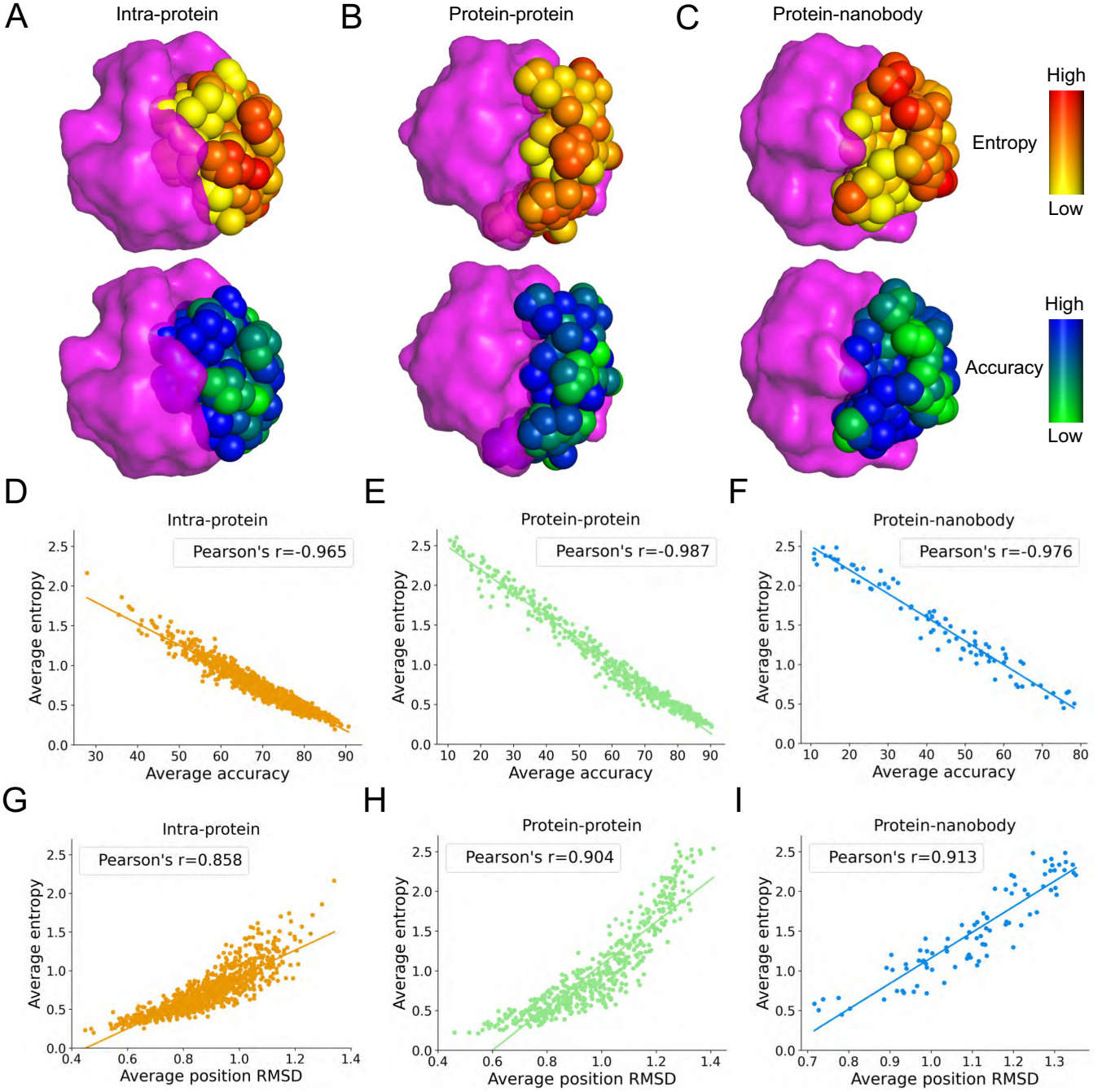


Fig. 4. Correlation between information entropy and prediction performance for Void-X. (*A*) An example of void filling of the masked region (sphere representation) interacting with the context region (surface representation) for an intrachain cluster from a phosphate transporter PHO1 (PDB ID: 9JF8), showing the relationship between entropy calculated in Eq. **3** and accuracy calculated in Eq. **4** at each of the filled atom sites. The atoms in the masked region are shown with the color spectrum yellow to red (representing low to high entropy) or the color spectrum green to blue (representing low to high accuracy). (*B*) The same as in (*A*) for void filling of the masked region for an interchain cluster from the BtpeA CTD–BtpiA complex structure (PDB ID: 9JDI). (*C*) The same analysis as in (*A*) for an interchain cluster from an ADP/ATP carrier bound to nanobody (PDB ID: 9J3L). (*D*) Correlation between information entropy and average accuracy for the entire masked region (defined in the text) for 1,000 intrachain clusters randomly selected from the test dataset. (*E*) Same analysis as in (*D*) for 500 protein–protein clusters randomly selected from the test dataset. (*F*) Same analysis as in (*D*) for all 87 protein–nanobody clusters from the test dataset. (*G*) Correlation between information entropy and the position RMSD for the entire masked region for 1,000 intrachain clusters randomly selected from the test dataset. (*H*) Same analysis as in (*G*) for 500 randomly selected protein–protein clusters. (*I*) Same analysis as in (*G*) for all 87 protein–nanobody clusters in the test dataset.

was generated independently for 100 times. The information entropy of the masked region (S_R^{AT}) was calculated as S_i^{AT} in Eq. **3**, averaged over all atom sites of the masked region. The corresponding average accuracy $\overline{ACC_R^{AT}}$ was computed as ACC_R^{AT} averaged over the 100 samples. Across all three categories (intraprotein, protein–protein, and protein–nanobody), the analysis showed an approximately linear inverse correlation between information entropy and prediction accuracy at the level of prediction of the masked regions (Fig. 4 *D–F*).

In addition to entropy versus atom-type accuracy, we investigated entropy change of the masked region with respect to spatial

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

deviation of generated atoms from their ground truth positions. We used the above randomly selected spherical clusters for this investigation. For each selected cluster, the masked atoms were generated 100 times as above for calculating S_R^{AT}. The corresponding average position RMSD was computed as the RMSD defined in Eq. **2** averaged over the 100 samples. This analysis shows a clear trend that the generated atomic packing of the masked region with smaller information entropy also has lower position RMSD (Fig. 4 *G–I*). The above analyses demonstrate a strong association between information entropy and prediction performance both for individual atom sites and at the level of entire masked regions. Thus, the information entropy is a reliable indicator of prediction confidence for Void-X.

Since information entropy correlates with predictive performance, we selected a few cases with low average entropy for visual inspection to assess the atomic packing quality produced by Void-X (Fig. 5). In general, these local structures contain few isolated residues, and most predicted residues exhibit well-defined sidechain structures. Notably, in the protein–nanobody case (Fig. 5 *E* and *F*), the generated atoms are located at the pocket-like binding interface. This confined localization likely results in lower entropy due to the restrained atomic environment within the binding site. For comparison, more cases with medium (*SI Appendix*, Fig. S6) and high (*SI Appendix*, Fig. S7) average entropy are also provided.

Discussion

We have shown that by learning the intricate patterns of local atom arrangement in protein complex structures using the masked discrete diffusion model, it is possible to recover missing atoms or generate new ones for interaction with a targeted structural region of a protein to a reasonable accuracy. The trained Void-X model is able to achieve a success rate of 53 to 78% depending on different types of test cases, e.g., intra- or intermolecular. For comparison, the chance of correctly assigning 70 out of 100 voids with each void being assigned randomly from the 50 atom types would be 3.47×10^{-94}. The successful training of the Void-X model suggests the existence of highly convoluted but self-consistent rules embedded in the observed protein conformation space for determining the packing of neighboring and distant atoms. In accordance with this notion, inclusion of intrachain clusters during model training resulted in a significant improvement of accuracy (~10%) in filling the atomic voids at intermolecular interfaces, further highlighting that the model has learned elementary properties of atomic interaction shared among the intra- or intermolecular interactions.

The above mentioned accuracy was derived from test cases with known structural answers. However, for de novo applications whereby the ground truth is not available, consistency of the model in generating answers provides a possible measure of

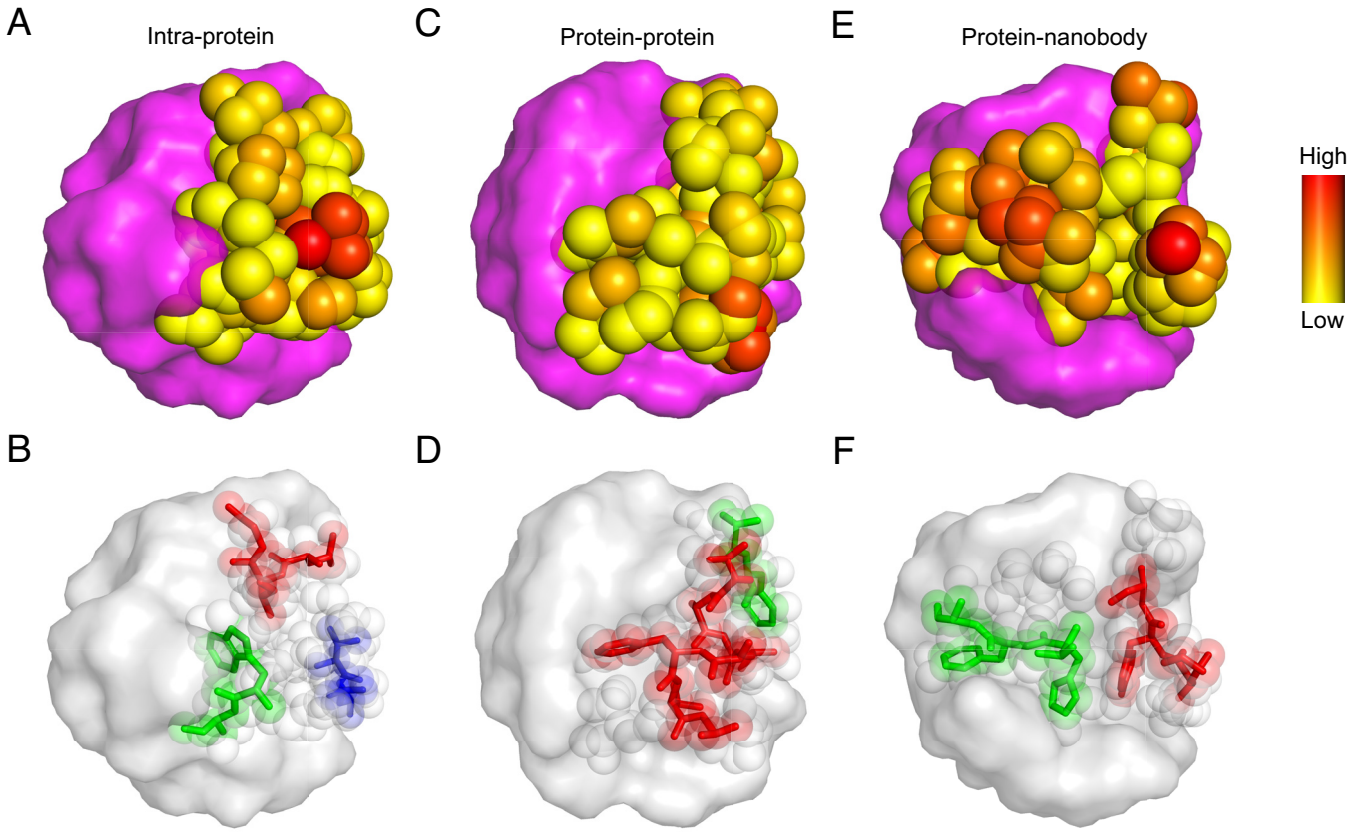


Fig. 5. Examples of Void-X generated atom clusters with low information entropy. (*A* and *B*) Generated atom cluster with low information entropy (0.24) for an intrachain cluster from protein JukB (PDB ID: 9JHT), showing that 3 continuous AA fragments can be reconstructed: LIG (red), WA (green), and VL (blue), with complete side chains for LIG, WA, and valine in the VL fragment. (*C* and *D*) Generated atom cluster with low information entropy (0.24) for a protein–protein cluster from the dimer structure of calcium/manganese antiporter SLC30A10 (PDB ID: 9KVX), showing that two continuous AA fragments can be rebuilt: TALCFTI (red) and VF (green), where the first fragment forms a short helix. All residues except cysteine and isoleucine in the first fragment and valine in the VF fragment contain full side chains. (*E* and *F*) Same analysis as in (*C* and *D*) for a protein–nanobody cluster from a superoxide dismutase bound to nanobody (PDB ID: 8YAT). In this case, the average information entropy for the masked region is 0.53. Two continuous AA fragments can be restored: SPYK (red) and PHFR (green), with tyrosine in SPYK and histidine and phenylalanine in PHFR exhibiting full side chains.

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

confidence level of the predictions. Information entropy has been used to measure perplexity in language models (57, 58). We find Void-X exhibiting remarkably strong inverse correlation between information entropy of the generated atoms and accuracy for the test cases, indicating that information entropy can be a reliable measure of prediction confidence.

A validated measure of prediction confidence offers the opportunity to evaluate whether a given protein site has favorable property for molecular interaction, which we refer to as bindability. For this purpose, one can position a context-void spherical cluster around the C_α atom of each of the surface-proximal residues of a protein such that there is ~70% context atoms and ~30% void atoms. Void-X is then used to fill the voids 100 times to compute information entropy of the generated atoms, with the notion that information entropy is inversely correlated with the intrinsic property of the local structural site to interact with other proteins (see *SI Appendix*, Fig. S8 legend for details). We tested this application for the human epidermal growth factor receptor (EGFR) without ligand binding as several different antibody bound structures are available for this protein. As shown in *SI Appendix*, Fig. S8, known antibody-binding epitopes of EGFR tend to overlap with regions of low information entropy (high bindability score). Protein surface mapping of functional regions has been addressed computationally by finding conserved residues in protein phylogenetic analysis (59). More recently, protein–protein coevolution has been employed to reveal differences between protein surfaces that have evolved to bind to other proteins and those that have not (60). We propose that the Void-X derived bindability metric could be an alternative indicator of site-specific bindability or potentially immunogenicity of protein surfaces.

The Void-X model could also function in synergy with other AI structure prediction models, such as AlphaFold3. To demonstrate this application, we used AlphaFold3 to predict the structures of two antigen–nanobody complexes (PDB IDs: 9HML and 9FZD). In both test cases, the structures predicted by AlphaFold3 were incorrect and did not converge to a unique binding mode. For the protein KIR2DL (9HML), AlphaFold3 provided five different solutions (*SI Appendix*, Fig. S9*A*), all with similar confidence scores (ipTM between 0.10 and 0.12). Likewise, for the OmpA protein (9FZD), AlphaFold3 predicted 5 solutions with different binding modes (*SI Appendix*, Fig. S9*B*). We then used Void-X to measure the information entropy of the ground truth and predicted binding interfaces. Interestingly, in both cases, the ground truth interfaces exhibit substantially lower information entropy than the incorrectly predicted interfaces (see *SI Appendix*, Fig. S9 legend).

Recent advances in PLMs have incorporated structural information to enhance representation learning. For instance, the structure-aware model SaProt (61) extends the standard residue-based vocabulary to include structural tokens derived from Foldseek (62), thereby enabling more informative pretraining. Compared to SaProt, Void-X operates at the atomic level and is trained to generate the structural details of masked regions within protein structures.

The key difference between Void-X and most other AI-based protein design methods, including but not limited to RFdiffusion (36), RFantibody (39), BindCraft (38), and BoltzGen (40), is that the former employs a bottom–up approach of generating atomic packing locally to inform global design whereas the latter use a top–down approach in which the overall protein complexes are first generated at coarse-grained level and the optimal sequences are subsequently designed to optimize the interaction. The top–down approach is attractive in that it provides a design solution to obtain the entire protein–protein complex, although the accuracy is limited by the amount of experimental structures of protein–protein interaction in the database (~160,000). The bottom–up approach benefits from the local sampling method that can drastically augment the training data to millions and potentially billions, but will need another model to translate the generated atom distribution to optimal protein conformation. We believe that as both approaches improve over time, they will converge and become truly complementary.

Finally, the current model was trained using spherical clusters from 189,377 experimentally determined structures. We note that the training dataset could be expanded by at least 1,000 folds by the inclusion of structures from the vast AlphaFold repository (over 200 million structures from 48 different proteomes). It will be interesting to investigate whether augmenting training data with AlphaFold structures can further improve the predictive power of Void-X.

Method

Dataset Construction. Protein structures for model training and evaluation were obtained from the PDB database as of October 30, 2025. The dataset was partitioned chronologically: The training dataset comprises structures deposited before March 31, 2023; the validation dataset includes those deposited between April 1, 2023, and March 31, 2024; and the test dataset consists of structures deposited after March 31, 2024. Data similarity was limited by a 40% sequence identity cutoff, ensuring that no sequences in the validation or test datasets exceed this threshold of similarity with any of sequences in the training dataset (32, 34). Since sequence-based similarity was used to construct the datasets, homologous backbone folds could be represented in both the training and test datasets. However, the corresponding sidechain conformations within these shared folds often exhibit considerable diversity. Training the model on structures having similar folds but varied sidechain geometries enhances its capacity to learn fundamental patterns of atomic packing. Prior to processing, structures were filtered according to the following criteria: a) the structures must have well-defined resolution, including X-ray crystal and cryoelectron microscopy (cryo-EM) structures with resolution better than 5 Å; b) protein chains with fewer than 50 residues are excluded; c) protein chains with more than 50% unassigned residues are removed.

For structures meeting the initial screening criteria, hydrogen atoms were added using PDBFixer (63). From each protein structure, spherical clusters of atoms were collected by centering a 12 Å-radius sphere on the C_α atom of a selected residue and extracting all atoms within this sphere. The selection of a 12 Å radius for generating spherical clusters was guided by a trade-off between the size of the training dataset and the area coverage of the protein–protein interaction (PPI). While larger clusters provide more area coverage of PPI, they yield less training samples. Conversely, smaller clusters allow for a larger training dataset, but with smaller area coverage of PPI. A 12 Å radius was found to be an optimal compromise, yielding a dataset of sufficient size while covering an average interfacial area of ~1,200 $Å^2$. A cluster was retained if it contains 512–768 atoms and at least 30 residues. For clusters containing atoms from two polypeptide chains, a minimum of 96 atoms from the interacting partner chain was required. To maintain dataset diversity, the maximum atom overlap between any two clusters derived from the same protein structure was limited to 70% for the training dataset. For the validation and test datasets, clusters within the same protein structure have no overlapped atom, and atoms from the partner chain were required to constitute 20%–30% of the total cluster atoms. Moreover, data similarity within the validation and test datasets was also limited by a 40% sequence identity cutoff. Following this processing, we collected 8.7 million clusters for training, 3,545 for validation, and 3,453 for final model evaluation. Finally, for further analysis, we constructed an additional test dataset comprising 87 spherical clusters derived from structures deposited after March 31, 2023, representing protein–nanobody interactions. In this test dataset, the complementarity-determining region (CDR) sequences of the nanobodies share no more than 40% sequence identity with those of the antibodies in the training dataset.

Uniformization of Atom Spatial Distribution. Protein surfaces display highly diverse geometric and chemical properties, presenting a substantial challenge to the computational design of atomic-level interfaces. A central problem is the lack

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

of a generalizable unbiased protocol for the placement of complementary atoms at an arbitrary structural site. To overcome this limitation, we introduce a cluster preprocessing strategy that uniformizes the spatial occupancies of all atoms in each spherical cluster. To maximize atomic occupancy within a sphere, atoms were arranged in a HCP configuration, which yields the highest average packing density (64). Specifically, a spherical region with 12 Å radius can be densely filled with 955 uniformly sized balls of radius 1.1 Å in the HCP arrangement. By reassigning the spherical cluster of atoms to the uniformly distributed balls, the atom spatial distribution at the context-mask interface is standardized, thus allowing for unbiased positioning of atoms for packing against a designated protein surface.

The task of uniformization is to positionally reassign all atoms in the spherical cluster extracted from the PDB to the uniformly spaced and optimally packed balls such that the total spatial deviation is minimized. This can be formulated as a linear sum assignment problem, in which the objective is to find a minimum-weight matching between the two sets of atoms. Specifically, let $\boldsymbol{C}$ be a cost matrix where $\boldsymbol{C}[i,j]$ represents the Euclidean distance between native atom i and repositioned atom j in the uniform reference. The solution is the assignment matrix $\boldsymbol{X}$ that minimizes the total cost:

$$min \sum_i \sum_j \boldsymbol{C}[i,j]\boldsymbol{X}[i,j] \quad [5]$$

subject to the constraints that $\boldsymbol{X}$ is a binary matrix with each row and column sums to 1, ensuring a one-to-one matching. This optimization is efficiently solved using the algorithm described in ref. 65.

AFM. The proposed Void-X is an AFM designed to complete interfacial voids within protein binding sites. The model operates on a local spherical cluster (SC) centered on a target binding site, which encompasses the surrounding context atoms as well as the atomic voids in the interface region to be generated. Formally, a spherical cluster of N atoms is defined as $SC = (\boldsymbol{at}^c, \boldsymbol{at}^m, \boldsymbol{aa}, \boldsymbol{r})$, where $\boldsymbol{at}^c \in \mathbb{R}^{n_c}$ and $\boldsymbol{at}^m \in \mathbb{R}^{n_m}$ denote atom types (defined in *SI Appendix*, Table S2) for the context and masked regions, respectively, with $n_c + n_m = N$, $\boldsymbol{aa} \in \mathbb{R}^N$ specifies amino acid types, and $\boldsymbol{r} \in \mathbb{R}^{N\times 3}$ represents atomic coordinates. Within the masked void region, amino acid identities are assigned an unknown label (UNK). Atomic coordinates in the spherical cluster are transformed from the real space to the uniform lattice space via the uniformization procedure described above, where the context atoms are returned to their original positions.

The problem of assigning the masked voids given the context atoms in a spherical cluster can be formalized as modeling the conditional distribution $\boldsymbol{p}(\boldsymbol{at}^m|\boldsymbol{at}^c, \boldsymbol{aa}, \boldsymbol{r})$, and a generative masked discrete diffusion model (51) is developed for this purpose. This model is trained on spherical clusters of atoms extracted from experimentally resolved protein structures. During training, a forward diffusion process progressively adds noise to original atom types within the masked region until a uniform distribution is reached. The model is then optimized to reverse this noising process for recovering original atom types. The algorithm of model training is given below.

For an atom with $K = 50$ categories, let $at_t \in \{1, \cdots, K\}$ denote the atom type, and $\boldsymbol{at}_t \in \{0,1\}^K$ represent the corresponding one-hot encoded row vector at diffusion timestep $t \in \{1, \cdots, T\}$, where $T = 500$. In this encoding, the k-th element of $\boldsymbol{at}_t$ is 1 (all others 0) if the atom belongs to the k-th category. Here, $\boldsymbol{at}_0$ is the original data, and $\boldsymbol{at}_T$ approaches a uniform distribution over the K categories.

The noising (forward diffusion) process q is defined as a categorical distribution with probabilities from $\boldsymbol{p}$:

$$q(\boldsymbol{at}_t|\boldsymbol{at}_{t-1}) = \text{Cat}(\boldsymbol{at}_t; \boldsymbol{p} = \boldsymbol{at}_{t-1}\boldsymbol{Q}_t). \quad [6]$$

In Eq. **6**, $\text{Cat}(\boldsymbol{at}_t; \boldsymbol{p})$ is the categorical distribution over the one-hot row vector $\boldsymbol{at}_t$ and $[\boldsymbol{Q}_t]_{ij} = q(at_t = j|at_{t-1} = i)$ is the transition probability matrix given by

$$\boldsymbol{Q}_t = (1-\beta_t)\boldsymbol{I} + \frac{\beta_t}{K}\boldsymbol{1}\boldsymbol{1}^T, \quad [7]$$

where $\boldsymbol{I}$ is the $K \times K$ identity matrix, $\boldsymbol{1}$ is a K-dimensional column vector of ones, and $\beta_t \in [0,1]$ follows a cosine noise schedule (66):

$$f(t) = cos^2\left(\frac{\pi}{2} \bullet \frac{t/T + s}{1+s}\right), \overline{\alpha}_t = \frac{f(t)}{f(0)}, \beta_t = 1 - \frac{\overline{\alpha}_t}{\overline{\alpha}_{t-1}} \quad [8]$$

with $s = 0.008$ to prevent excessively small β_t at the beginning of the noising process. Under these formulations, a small amount of uniform noise is added incrementally as the timestep t increases, and the previous value $\boldsymbol{at}_{t-1}$ is sampled with probability $1 - \beta_t$. The transition matrix $\boldsymbol{Q}_t$ is applied independently to each atom. Because the diffusion process is Markovian, $q(\boldsymbol{at}_{1:T}|\boldsymbol{at}_0) = \prod_{t=1}^{T} q(\boldsymbol{at}_t|\boldsymbol{at}_{t-1})$, $\boldsymbol{at}_t$ can be sampled directly from $\boldsymbol{at}_0$ via:

$$q(\boldsymbol{at}_t|\boldsymbol{at}_0) = \text{Cat}(\boldsymbol{at}_t; \boldsymbol{p} = \boldsymbol{at}_0\overline{\boldsymbol{Q}}_t), \overline{\boldsymbol{Q}}_t = \boldsymbol{Q}_1\boldsymbol{Q}_2 \cdots \boldsymbol{Q}_t, \quad [9]$$

where $\overline{\boldsymbol{Q}}_t$ is the cumulative product. And the categorical posterior distribution at diffusion timestep $t-1$ can be derived in closed form:

$$\begin{aligned} q(\boldsymbol{at}_{t-1}|\boldsymbol{at}_t, \boldsymbol{at}_0) &= \frac{q(\boldsymbol{at}_t|\boldsymbol{at}_{t-1}, \boldsymbol{at}_0)q(\boldsymbol{at}_{t-1}|\boldsymbol{at}_0)}{q(\boldsymbol{at}_t|\boldsymbol{at}_0)} \\ &= \text{Cat}(\boldsymbol{at}_{t-1}; \boldsymbol{p} = \frac{\boldsymbol{at}_t\boldsymbol{Q}_t^{\top} \odot \boldsymbol{at}_0\overline{\boldsymbol{Q}}_{t-1}}{\boldsymbol{at}_0\overline{\boldsymbol{Q}}_t\boldsymbol{at}_t^{\top}}), \end{aligned} \quad [10]$$

where $\top$ represents transpose operation and $\odot$ denotes element-wise multiplication.

In the denoising (reverse diffusion) process, the DiT (52) serves as the backbone of the AFM (*SI Appendix*, Fig. S3*A*), which predicts the clean data $\boldsymbol{at}_0$ from the noisy data $\boldsymbol{at}_t$. The AFM accepts input features describing individual atoms and atom–atom pairs, defined as

$$\boldsymbol{F}_a = Emb(\boldsymbol{at}^n) + Emb(\boldsymbol{aa}), \boldsymbol{F}_{ap} = Emb\left(dist(\boldsymbol{r})\right), \quad [11]$$

where $\boldsymbol{F}_a$ and $\boldsymbol{F}_{ap}$ denote atomic-level and pairwise features respectively, $\boldsymbol{at}^n$ includes context atoms and noised atoms. $Emb(\bullet)$ retrieves learnable embeddings using the input indices, and $dist(\bullet)$ calculates the Euclidean distance between atom pairs. Distances are discretized into 64 bins, each with a width of 0.375 Å, covering a total range of 0 to 24 Å. This range aligns with the diameter of the spherical cluster used to represent the local atomic environment.

The atomic-level and pairwise input features are then processed through a series of transformer-based DiT blocks, as defined by

$$\boldsymbol{F}_a^l, \boldsymbol{F}_{ap}^l = \text{DiTblock}^l(\boldsymbol{F}_a^{l-1}, \boldsymbol{F}_{ap}^{l-1}, t), \quad [12]$$

where l denotes the l-th DiT block. The diffusion timestep t is incorporated via adaptive layer normalization (adaLN) (67), which performs feature-wise normalization with parameters regressed from the embedding of t. From the outputs of the final DiT block, four separate prediction heads are constructed for the masked region: one for atom type (*SI Appendix*, Fig. S3*B*), one for amino acid type (*SI Appendix*, Fig. S3*C*), one for pairwise weight (Eq. **14**) used in coordinate refinement (*SI Appendix*, Fig. S3*D*), and one for interatomic distance (*SI Appendix*, Fig. S3*E*). These predictions are formulated as

$$\hat{\boldsymbol{at}} = \text{MLP}_{at}\left(\boldsymbol{F}_a^L, t\right), \hat{\boldsymbol{aa}} = \text{MLP}_{aa}\left(\boldsymbol{F}_a^L, t\right), \hat{\boldsymbol{d}} = \text{MLP}_d(\boldsymbol{F}_{ap}^L, t) \quad [13]$$

$$\Delta\boldsymbol{F}_{ap} = \boldsymbol{F}_{ap}^L - \boldsymbol{F}_{ap}^0, \hat{\boldsymbol{r}}_i = \boldsymbol{r}_i + \sum_{j=1}^{N} \frac{(\boldsymbol{r}_i - \boldsymbol{r}_j)\text{MLP}_r(\Delta\boldsymbol{F}_{ap}, t)}{N} \quad [14]$$

where each MLP denotes a two-layer feedforward neural network, $\hat{\boldsymbol{at}}$, $\hat{\boldsymbol{aa}}$, and $\hat{\boldsymbol{d}}$ represent the predicted atom type, amino acid type, and interatomic distance (assigned to 64 bins) respectively, N is the number of atoms, and $\hat{\boldsymbol{r}}_i$ is the refined coordinate of atom i, obtained via a weighted sum of displacement vectors relative to all other atoms (68, 69).

The AFM denoising network approximates the true atomic state $\boldsymbol{at}_0$ by predicting a categorical probability distribution $\hat{p}_\theta(\hat{\boldsymbol{at}}_0|\boldsymbol{at}_t)$ at each diffusion timestep.

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

Using the posterior distribution defined in Eq. **10**, the reverse generative distribution $p_\theta(\boldsymbol{at}_{t-1}|\boldsymbol{at}_t)$ is then parameterized by

$$p_\theta(\boldsymbol{at}_{t-1}|\boldsymbol{at}_t) \propto \sum_{\hat{\boldsymbol{at}}_0} q(\boldsymbol{at}_{t-1}|\boldsymbol{at}_t, \hat{\boldsymbol{at}}_0)\hat{p}_\theta(\hat{\boldsymbol{at}}_0|\boldsymbol{at}_t), \quad [15]$$

where the sum is taken over all possible realizations of the predicted state $\hat{\boldsymbol{at}}_0$. To train the reverse denoising process, we minimize the Kullback–Leibler (KL) divergence between the learned generative distribution and the categorical posterior distribution:

$$L_{KL} = D_{KL}(q(\boldsymbol{at}_{t-1}|\boldsymbol{at}_t, \boldsymbol{at}_0)||p_\theta(\boldsymbol{at}_{t-1}|\boldsymbol{at}_t)), \quad [16]$$

where L_{KL} is averaged over all noised atoms. To enhance training stability and prediction accuracy, auxiliary loss terms are incorporated. The cross-entropy loss is applied for atom type (L_{at}), amino acid type (L_{aa}), and interatomic distance (L_{dist}), each measured against the corresponding ground truth categorical labels, defined as

$$L_{at}, L_{aa}, L_{dist} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \quad [17]$$

where y_i is the true label for class i, $\hat{y}_i$ is the predicted probability for class i, C is the number of classes. For atomic coordinate refinement, the smooth L1 loss (L_{coord}) is used to penalize displacement errors, defined as

$$L_{coord} = \begin{cases} 0.5(r-\hat{r})^2, if|r-\hat{r}| < 1.0 \\ |r-\hat{r}| - 0.5, \text{otherwise} \end{cases}, \quad [18]$$

where r is the true coordinate component and $\hat{r}$ is the predicted coordinate component. The complete training objective is therefore:

$$L = L_{at} + L_{aa} + L_{dist} + L_{coord} + L_{KL}. \quad [19]$$

Model Training. In the AFM configuration (*SI Appendix*, Table S1), the architecture comprises 36 DiT blocks with a hidden dimension of 512 and 16 attention heads. The model was trained using the AdamW (70) optimizer with a base learning rate of 1×10^{-4} and a weight decay of 0.01. The learning rate followed a polynomial schedule over 3 million training steps, with a linear warmup from 0 to the base learning rate across the first 2,000 steps, then decaying to 1×10^{-6} with a power factor of 1. A constant learning rate of 1×10^{-6} was maintained for further training. Training was conducted with a global batch size of 64, and gradients were clipped to a maximum norm of 10. The model was trained on 8 NVIDIA H100 GPUs, completed in approximately 3 wk.

Evaluation Metrics. The quality of atomic void filling within the masked region is evaluated by comparing the generated atoms to their corresponding ground truth in the protein structures. For each spherical cluster, the generated atom distributions are evaluated using two metrics: accuracy (ACC) and RMSD. ACC is defined as the fraction of atoms in the masked region with correctly assigned atom type. RMSD quantifies the average spatial deviation from ground truth of all atoms in the masked region filled by the AFM.

In protein structures, the distribution of atom types is imbalanced, as backbone atoms occur with higher frequency than individual sidechain atoms. To enable a detailed assessment of the AFM across all atom types, we compute a precision–recall (PR) curve for each type. Using the backbone alpha carbon (C_α) as an example, precision and recall are defined as follows:

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, \quad [20]$$

where *TP* is the number of correctly assigned C_α atoms (true positive), *FP* is the number of incorrectly assigned C_α atoms (false positive), and false negative (*FN*) is the number of ground truth C_α atoms that are not matched by any generated atoms. The PR curve is constructed by plotting precision against recall at varying thresholds (from 0 to 1). A higher area under the curve reflects better quality of generation, characterized by both high precision and recall. The average precision (AP) approximates the area under the PR curve, formulated as

$$\text{AP} = \sum_i (R_i - R_{i-1})P_i, \quad [21]$$

where P_i and R_i are the precision and recall defined in Eq. **20** at the i-th threshold. We report the mean average precision (mAP) by averaging AP values across all atom types. This metric is calculated using all atoms from all spherical clusters in the test dataset.

For confidence assessment, we employ information entropy as a measure of uncertainty for the atomic packing generated by Void-X. Given an arbitrary spherical cluster with context atoms and masked voids, 100 independent void fillings are sampled for the masked region. For a detailed derivation of the information entropy, please refer to the main text.

Data, Materials, and Software Availability. Training and evaluation structures were downloaded from https://files.wwpdb.org/pub/pdb/data/structures/ (71). Sequence-based clusters are available at https://cdn.rcsb.org/resources/sequence/clusters/clusters-by-entity-40.txt (72). All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. We thank Hang Qiao for helpful discussion and Cheng Xia for helping with access to computation resources. This project was supported by the Science and Technology Commission of Shanghai Municipality (22JC1410500), the National Natural Science Foundation of China-Research Fund for International Scientists (82350710799), the Shanghai Municipal Science and Technology Major Project, the Shanghai Basic Research Pioneer Project, the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB1060000), and the Chinese Academy of Sciences (2023000040, SIOCIRCBC230202) (to J.J.C.); the China National Natural Science Foundation (82188101), the Shanghai Municipal Science and Technology Major Project (Grant No. 2019SHZDZX02), the Shanghai Science and Technology Development Funds (22JC1410400), and the Shanghai Key Laboratory of Aging Studies (19DZ2260400) (to J. Yuan).

Author affiliations: [a]Interdisciplinary Research Center on Biology and Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 201203, China; [b]State Key Laboratory of Chemical Biology, Chinese Academy of Sciences, Shanghai 200032, China; and [c]Shanghai Key Laboratory of Aging Studies, Shanghai 201203, China

1. E. Katchalski-Katzir *et al.*, Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 2195–2199 (1992).
2. D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, H. J. Wolfson, PatchDock and SymmDock: Servers for rigid and symmetric docking. *Nucleic Acids Res.* **33**, W363–W367 (2005).
3. T.M.-K. Cheng, T. L. Blundell, J. Fernandez-Recio, pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins* **68**, 503–515 (2007).
4. S. Lyskov, J. J. Gray, The RosettaDock server for local protein–protein docking. *Nucleic Acids Res.* **36**, W233–W238 (2008).
5. B. G. Pierce *et al.*, ZDOCK server: Interactive docking prediction of protein–protein complexes and symmetric multimers. *Bioinformatics* **30**, 1771–1773 (2014).
6. G. C. P. van Zundert *et al.*, The HADDOCK2.2 web server: User-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
7. N. Plattner, S. Doerr, G. De Fabritiis, F. Noé, Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **9**, 1005–1011 (2017).
8. A. C. Pan *et al.*, Atomic-level characterization of protein–protein association. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 4244–4249 (2019).
9. I. A. Vakser, S. Grudinin, N. W. Jenkins, P. J. Kundrotas, E. J. Deeds, Docking-based long timescale simulation of cell-size protein systems at atomic resolution. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2210249119 (2022).
10. J.-N. Chen, F. Jiang, Y.-D. Wu, Accurate prediction for protein–peptide binding based on high-temperature molecular dynamics simulations. *J. Chem. Theory Comput.* **18**, 6386–6395 (2022).
11. S. A. Wankowicz, J. S. Fraser, Advances in uncovering the mechanisms of macromolecular conformational entropy. *Nat. Chem. Biol.* **21**, 623–634 (2025).
12. R. Georgelin, C. J. Jackson, Entropy, enthalpy, and evolution: Adaptive trade-offs in protein binding thermodynamics. *Curr. Opin. Struct. Biol.* **94**, 103080 (2025).
13. C.-Y. Chen, I. Georgiev, A. C. Anderson, B. R. Donald, Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 3764–3769 (2009).
14. Y. Min *et al.*, From static to dynamic structures: Improving binding affinity prediction with graph-based deep learning. *Adv. Sci. (Weinh)* **11**, e2405404 (2024).
15. S. Sledzieski, R. Singh, L. Cowen, B. Berger, D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Syst.* **12**, 969–982.e966 (2021).

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.

16. U. Lupo, D. Sgarbossa, A.-F. Bitbol, Pairing interacting protein sequences using masked language modeling. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2311887121 (2024).
17. Y. S. Ko, J. Parkinson, C. Liu, W. Wang, TUnA: An uncertainty-aware transformer model for sequence-based protein–protein interaction prediction. *Brief. Bioinf.* **25**, bbae359 (2024).
18. D. Liu *et al.*, PLM-interact: Extending protein language models to predict protein-protein interactions. *Nat. Commun.* **16**, 9012 (2025).
19. N. Ferruz, S. Schmidt, B. Höcker, ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
20. A. Madani *et al.*, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
21. A. Rives *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2016239118 (2021).
22. Z. Lin *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
23. T. Hayes *et al.*, Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
24. B. Chen *et al.*, xTrimoPGLM: Unified 100-billion-parameter pretrained transformer for deciphering the language of proteins. *Nat. Methods* **22**, 1028–1039 (2025).
25. The UniProt Consortium, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
26. L. Richardson *et al.*, MGnify: The microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
27. B. L. Hie *et al.*, Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* **42**, 275–283 (2024).
28. V. R. Shanker, T. U. J. Bruun, B. L. Hie, P. S. Kim, Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science* **385**, 46–53 (2024).
29. R. Singh *et al.*, Learning the language of antibody hypervariability. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2418918121 (2025).
30. K. Ding *et al.*, Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nat. Commun.* **15**, 6392 (2024).
31. F. Jiang *et al.*, A general temperature-guided language model to design proteins of enhanced stability and activity. *Sci. Adv.* **10**, eadr2641 (2024).
32. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
33. R. Evans *et al.*, Protein complex prediction with AlphaFold-Multimer. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2021.10.04.463034 (Accessed 10 March 2022).
34. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
35. S. Basu, B. Wallner, DockQ: A quality measure for protein-protein docking models. *PLoS ONE* **11**, e0161879 (2016).
36. J. L. Watson *et al.*, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
37. S. A. Rettie *et al.*, Accurate de novo design of high-affinity protein-binding macrocycles using deep learning. *Nat. Chem. Biol.* **21**, 1948–1956 (2025).
38. M. Pacesa *et al.*, One-shot design of functional protein binders with BindCraft. *Nature* **646**, 483–492 (2025).
39. N. R. Bennett *et al.*, Atomically accurate de novo design of antibodies with RFdiffusion. *Nature* **649**, 183–193 (2025).
40. H. Stark *et al.*, BoltzGen: Toward universal binder design. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.11.20.689494 (Accessed 24 November 2025).
41. W. Ahern *et al.*, Atom-level enzyme active site scaffolding using RFdiffusion2. *Nat. Methods* **23**, 96–105 (2026).
42. P. Gainza *et al.*, De novo design of protein interactions with learned surface fingerprints. *Nature* **617**, 176–184 (2023).
43. V. Zambaldi *et al.*, De novo design of high-affinity protein binders with AlphaProteo. arXiv [Preprint] (2024). https://doi.org/10.48550/arXiv.2409.08022 (Accessed 12 September 2024).
44. A. Vaswani *et al.*, "Attention is all you need" in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds. (Curran Associates, Inc., 2017), pp. 5998–6008.
45. J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics" in *International Conference on Machine Learning* (PMLR, 2015), pp. 2256–2265.
46. J. Ho, A. Jain, P. Abbeel, "Denoising diffusion probabilistic models" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin, Eds. (Curran Associates, Inc., 2020), pp 6840–6851.
47. J. Dauparas *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
48. T. Chai Discovery *et al.*, Zero-shot antibody design in a 24-well plate. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.07.05.663018 (Accessed 6 July 2025).
49. L. S. Mille-Fragoso *et al.*, Efficient generation of epitope-targeted de novo antibodies with Germinal. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.09.19.677421 (Accessed 25 September 2025).
50. E. Swanson, M. Nichols, S. Ravichandran, P. Ogden, mBER: Controllable de novo antibody design with million-scale experimental screening. bioRxiv [Preprint] (2025). https://doi.org/10.1101/2025.09.26.678877 (Accessed 28 September 2025).
51. J. Austin, D. D. Johnson, J. Ho, D. Tarlow, R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces" in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, J. Wortman Vaughan, Eds. (Curran Associates, Inc., 2021), pp 17981–17993.
52. W. Peebles, S. Xie, "Scalable diffusion models with transformers" in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 4195–4205.
53. J. Achiam *et al.*, GPT-4 technical report. arXiv [Preprint] (2023). https://doi.org/10.48550/arXiv.2303.08774 (Accessed 19 December 2023).
54. A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with CLIP latents. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2204.06125 (Accessed 13 April 2022).
55. R. Girshick, "Fast R-CNN" in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2015), pp. 1440–1448.
56. J. Davis, M. Goadrich, "The relationship between precision-recall and ROC curves" in *International Conference on Machine Learning* (PMLR, 2006), pp. 233–240.
57. F. Jelinek, R. L. Mercer, L. R. Bahl, J. K. Baker, Perplexity–A measure of the difficulty of speech recognition tasks. *J. Acoust. Soc. Am.* **62**, S63–S63 (1977).
58. P. F. Brown, V. J. D. Pietra, R. L. Mercer, S. A. D. Pietra, J. C. Lai, An estimate of an upper bound for the entropy of English. *Comput. Linguist.* **18**, 31–40 (1992).
59. A. Armon, D. Graur, N. Ben-Tal, ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.* **307**, 447–463 (2001).
60. A. Yang *et al.*, Structural ontogeny of protein-protein interactions. *Science* **391**, eadx6931 (2026).
61. J. Su *et al.*, "SaProt: Protein language modeling with structure-aware vocabulary" in *International Conference on Learning Representations* (2024).
62. M. van Kempen *et al.*, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
63. P. Eastman *et al.*, OpenMM 8: Molecular dynamics simulation with machine learning potentials. *J. Phys. Chem. B* **128**, 109–116 (2024).
64. T. C. Hales, Historical overview of the Kepler conjecture. *Discrete Comput. Geom.* **36**, 5–20 (2006).
65. D. F. Crouse, On implementing 2D rectangular assignment algorithms. *IEEE Trans. Aerosp. Electron. Syst.* **52**, 1679–1696 (2016).
66. A. Q. Nichol, P. Dhariwal, "Improved denoising diffusion probabilistic models" in *International Conference on Machine Learning* (PMLR, 2021), pp. 8162–8171.
67. E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville, "Film: Visual reasoning with a general conditioning layer" in *Proceedings of the AAAI Conference on Artificial Intelligence* (2018).
68. V. G. Satorras, E. Hoogeboom, M. Welling, "E(n) equivariant graph neural networks" in *International Conference on Machine Learning* (PMLR, 2021), pp. 9323–9332.
69. G. Zhou *et al.*, "Uni-mol: A universal 3D molecular representation learning framework" in *International Conference on Learning Representations* (2023).
70. I. Loshchilov, F. Hutter, "Decoupled weight decay regularization" in *International Conference on Learning Representations* (2017).
71. H. M. Berman, K. Henrick, H. Nakamura, Announcing the worldwide Protein Data Bank. Protein Data Bank. https://files.wwpdb.org/pub/pdb/data/structures/. Accessed 30 October 2025.
72. H. M. Berman *et al.*, The Protein Data Bank. https://cdn.rcsb.org/resources/sequence/clusters/clusters-by-entity-40.txt. Accessed 30 October 2025.

Downloaded from https://www.pnas.org by CAS CONSORTIUM CHINESE ACADEMY OF SCIENCES on June 14, 2026 from IP address 101.230.128.10.