Learning Universal Representations of Intermolecular Interactions with ATOMICA

Ada Fang[1,2,3], Michael Desgagné[4], Zaixi Zhang[3], Andrew Zhou[3,5], Joseph Loscalzo[6,7], Bradley L. Pentelute[4,8,9,10], and Marinka Zitnik[2,3,10,11,‡]

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA
[2]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, MA, USA
[3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[4]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA
[5]Program in Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, USA
[6]Department of Medicine, Harvard Medical School, Boston, MA, USA
[7]Brigham and Women's Hospital, Boston, MA, USA
[8]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA
[9]Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
[10]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[11]Harvard Data Science Initiative, Cambridge, MA, USA

‡Correspondence: marinka@hms.harvard.edu

Abstract

Molecular interactions underlie nearly all biological processes, but many representation learning models either focus on single entities or are trained for a narrow set of interaction settings. Here, we introduce ATOMICA, a geometric deep learning model that learns atomic-scale representations of intermolecular interfaces across five modalities, including proteins, small molecules, metal ions, lipids, and nucleic acids. ATOMICA is trained on 2,037,972 interaction complexes to generate embeddings of interaction interfaces at the levels of atoms, chemical blocks, and molecular interfaces. The latent space is multiscale and reflects physicochemical features shared across molecular classes. On the RNAGlib 3D structure-function benchmark, ATOMICA attains the best performance across four tasks, and in protein pocket ligand classification, ATOMICA improves upon established protein pocket encoders and is comparable to protein language models. Using the shared embedding space, we embed orthosteric PPI inhibitors and find inhibitor embeddings are more similar to interface embeddings proximal to the native binding site across protein–peptide and protein–protein complexes. We use ATOMICA to suggest putative ligands to pockets in the dark proteome, which are proteins lacking known function. In total, ligands are predicted for 2,646 dark protein pockets and heme binding is experimentally confirmed for five ATOMICA predictions. ATOMICA opens new avenues for learning representations of intermolecular interactions.

Main

Molecular interactions underpin essentially every chemical and biological process, from enzymatic catalysis and signaling to gene regulation and drug binding. Generative complex predictors such as AlphaFold3 [1] and RoseTTAFold All-Atom [2] show that a single model can model structures across multiple classes of interacting biomolecules. These models are trained for complex structure prediction, not for learning a general representation of intermolecular interfaces across tasks.

In parallel, unimolecular foundation models, including protein language models [3–7], nucleic-acid language models [8–11], and small-molecule encoders [12–14], learn strong within-modality priors from large-scale single-entity corpora. These models typically do not incorporate paired interacting partners as a primary pretraining signal and therefore do not necessarily align interaction features across molecular modalities.

Complementing these efforts, a large body of work learns interaction representations for specific modality pairs, such as protein-protein [15–21], protein-ligand [22–25], protein-RNA [26–31], protein–ion [32–34], and RNA-ligand [35–37]. Because these approaches are typically built around particular entity types, featurizations, and datasets, extending them to additional modalities often requires new encoders, new supervision, or re-specifying the interaction representation.

These limitations motivate an interaction-centered representation model grounded in 3D interface geometry and local chemistry, since non-covalent interactions are governed by the spatial arrangement of atoms [38]. Interface geometry varies across molecular modalities [39–41], and a shared model can provide a common representation across interface types, allowing features learned in one setting to transfer to others.

We introduce ATOMICA, an all-atom geometric deep learning model that learns representations of intermolecular complexes (Fig. 1a) across five molecular modalities: proteins, small molecules, metal ions, lipids, and nucleic acids, encompassing eight interaction types. ATOMICA is trained on a dataset of 2,037,972 interaction complexes (Fig. 1b) from the Cambridge Structural Database (CSD) [42] and the Protein Data Bank (PDB) [43]. The model uses a geometric, SE(3)-equivariant architecture and is trained with a self-supervised objective that combines denoising of atomic coordinates and masked block identity prediction. The latent space groups atoms, blocks, and complexes reflecting their chemical similarity and physicochemical properties. Residues ranked highly by ATOMICAScore at interaction interfaces are overrepresented with

residues involved in non-covalent contacts. ATOMICA pretraining outperforms models trained on a single pair of interacting modalities, with the largest improvements observed for low-data modalities, such as protein-nucleic acid complexes. On RNAGlib 3D structure-function benchmarks, ATOMICA outperforms existing RNA structure encoders and RNA language models. On protein pocket tasks, ATOMICA surpasses specialized pocket encoders and matches the performance of protein language models. The unified ATOMICA embedding space also supports cross-modality interface comparison by relating orthosteric protein-protein interaction (PPI) inhibitors to the corresponding native protein interfaces for protein-protein and protein-peptide interactions. Finally, we apply ATOMICA to the dark proteome, which are regions of the proteome without functional annotation [44–46]. ATOMICA assigns putative ions and cofactors to 2,646 ligand-binding sites. We experimentally confirm heme binding for five dark proteins using recombinant expression and spectroscopic assays.

Results

Dataset of molecular interactions for proteins, small molecules, ions, lipids, and nucleic acids.

We assembled a dataset of interacting molecular entities from the Cambridge Structural Database (CSD) v2022.3.0 [42] and Q-BioLiP [47, 48] which contains intermolecular interactions across all modalities available in the Protein Data Bank (PDB). From the CSD, we filtered for entries of small molecule crystals and extracted unique pairs of conformers from the unit cell (Methods 1), resulting in 1,767,710 interacting pairs of small molecules, of which 1,747,710 are used for training. We also processed 337,993 interaction complexes from Q-BioLiP of which 290,262 are used for training. This includes structures of protein complexes with proteins, DNA, RNA, peptides, small molecules, and metal ions, as well as nucleic acid ligand structures from the PDB. The interaction interface between two entities is defined by atoms within an 8 Å distance to the other molecule to capture atoms in their surrounding molecular context. In Fig. 1b, we show examples of the interaction interface for various complexes and their relative size.

We represent interaction complexes using two-level graphs to capture atomic-level details and higher-order chemical structures (Fig. 1c), which has theoretically higher expressive power than atom-level graphs alone [49]. At the first level, nodes in the graph represent atoms, each defined by its element type and 3D spatial coordinates. At the second level, atoms are grouped

into chemically meaningful blocks, such as amino acids in proteins, nucleotides in nucleic acids, and functional moieties in small molecules, to form a block-level graph [50–52] (Methods 1.3). Within each graph level, we define two types of edges: intramolecular and intermolecular edges which connect the k nearest nodes in Euclidean space within and on the interface of interacting molecules, respectively.

Geometric deep learning model of intermolecular interactions at atomic scale.

ATOMICA is a self-supervised geometric graph neural network that generates multi-scale embeddings at the atom, block, and graph-level from the structure of two interacting molecules (refer to Fig. 1c for embedding notation used throughout). ATOMICA can be finetuned and adapted with task specific heads for predictive tasks such as binding site annotation. Embeddings are learned with SE(3)-equivariant tensor field networks that pass messages over intra- and intermolecular edges (Methods 2), a framework used for interatomic potentials, docking, and RNA structure scoring [53–56].

To train ATOMICA, we use self-supervised pretraining objectives that combine denoising and masked block prediction. Denoising is an effective pretraining strategy for learning molecular representations [24, 57–60]. In ATOMICA, denoising requires the model to reconstruct an interaction complex graph after it is perturbed by a rigid SE(3) transformation and random torsion angle rotations of one molecular entity at the interface (Fig. 1d). This objective encourages the model to use relative spatial relationships and local chemical context at the interface rather than relying on absolute coordinates. Inspired by masked modeling objectives used to learn protein [5] and nucleic-acid [11] representations, ATOMICA is also trained to predict the identity of randomly masked chemical blocks at the interaction interface. Together, these objectives encourage the model to learn geometric and chemical features that are relevant to intermolecular interfaces.

ATOMICA reflects multi-scale organization of interacting molecules.

Visualizing ATOMICA embeddings at the graph level in 2D with UMAP for all interaction complexes suggests that the embeddings reflect relative chemical similarity of molecular interactions (Fig. 1e). We expect the protein-protein and protein-peptide graph embeddings to be relatively close in the latent space as they share similar intermolecular interactions. This is confirmed by observing that the distribution of pairwise cosine similarity between $\mathbf{h}^{\text{graph}}_{\text{protein-peptide}}$ and $\mathbf{h}^{\text{graph}}_{\text{protein-protein}}$ is significantly higher than that between $\mathbf{h}^{\text{graph}}_{\text{protein-peptide}}$ and $\mathbf{h}^{\text{graph}}_{\text{other modality pairs}}$ (KS-

statistic = 0.100, p-value < 0.001). For the latent space of atom embeddings, PCA shows coarse grouping by elemental category (Fig. 2a), with nonmetals tending to separate from metals. For block embeddings, PCA suggests separation between amino-acid blocks and nucleotide blocks (DNA/RNA), with within-group trends that are broadly consistent with residue physicochemical properties (Fig. 2b).

ATOMICASCORE highlights residues most important to the interface representation.

Here we distinguish between the *interface region* (all residues within a certain distance neighborhood of the partner) and *contact residues*, defined here as residues participating in annotated non-covalent interactions such as hydrogen bonds, hydrophobic contacts, and aromatic interactions, based on established geometric criteria [61] (Fig. 2c).

To probe the model in a zero-shot setting, we quantify the importance of block i by masking it and measuring the resulting change in the complex embedding. Specifically, we define ATOMICASCORE $(i) = \text{sim}(\mathbf{h}_G^{\text{graph}}, \mathbf{h}_{G\setminus i}^{\text{graph}})$, where $G \setminus i$ denotes masking block i in graph G (Fig. 2d) and sim is the cosine similarity function. A lower ATOMICASCORE indicates a larger representational change and therefore higher block importance (Methods 7).

The contact categories (hydrogen-bonding, hydrophobic, and aromatic contacts) are derived from standard geometric interaction annotations from PLIP [61]. We compare ATOMICASCORE to a unimolecular baseline (ESM-2 3B) that provides zero-shot residue importance estimates via mutation-effect scoring [62]. We evaluate precision@10 as the proportion of blocks ranked by ATOMICASCORE in the top 10 that are involved in a specific contact (Fig. 2e). On an unseen test set of protein–small-molecule complexes, ATOMICASCORE achieves the best precision@10 (Fig. 2f), retrieving on average 2.7 amino-acid blocks involved in annotated contacts, compared to 2.4 for ESM-2 (3B) and 2.0 for a random reference (Fig. 2f). ATOMICA retrieves residues across hydrogen-bonding, hydrophobic, and aromatic contact types (Fig. S2).

Scaling laws in the generalizability of ATOMICA across molecular modalities.

To quantify the benefit of pretraining ATOMICA on multiple molecular modalities, we compare its performance to otherwise identical models pretrained exclusively on single pairs of interacting modalities (Fig. 2g). We evaluate each model using masked block identity prediction on a test set of interface complexes selected to have low sequence similarity and low chemical similarity to any training or validation examples (Methods 7). Pretraining ATOMICA across all

modalities improves prediction accuracy for masked blocks in protein-DNA interfaces from 0.24 to 0.71 AUPRC (a 190% increase, with 2,253 protein-DNA training complexes, Fig. 2h). Similar improvements are observed for protein-RNA interfaces, where performance increases from 0.19 to 0.55 AUPRC using 2,975 protein-RNA training complexes. For protein-peptide interfaces, performance rises from 0.32 to 0.67 AUPRC with 7,346 training complexes. The improvements in AUPRC scale with dataset size (Fig. 2i). These results illustrate a scaling law: as the diversity and volume of interaction data increase, representation quality and predictive performance consistently improve.

ATOMICA representations facilitate accurate RNA interface annotation.

We evaluate ATOMICA on four RNAglib tasks [63, 64] where the input is the RNA 3D structure and the predictive tasks are: protein-binding site prediction, small-molecule binding site prediction, RNA functional annotation, and pocket ligand identification (Fig. 3a,c,e,g). We compare to structure-based baselines (RNAglib/RGCN [64, 65], gRNAde [66]) and sequence-based RNA language models (RNAErnie [67], RNA-FM [10, 68], RiNALMo [9]). For each task, we fine-tune ATOMICA with a task-specific head using the benchmark training splits and evaluate on held-out test sets which contain RNA that are distinct in structure and sequence (Methods 4).

The first task is residue-level binary classification of protein-binding sites (Fig. 3a, dataset size: 74,089 residues from 1,138 structures). ATOMICA achieves an AUPRC of 0.604, exceeding the best baseline, gRNAde, by +0.118 (Fig. 3b). We next evaluate RNA functional annotation using Gene Ontology (GO) terms curated from Rfam [69] (Fig. 3c, dataset size: 499 structures). This is a multi-label classification task with five different GO terms. ATOMICA achieves an F1 macro of 0.951 and F1 micro of 0.920, exceeding the best baseline (RiNALMo) by +0.105 (F1 macro) and +0.053 (F1 micro) (Fig. 3d). For residue-level binary classification of small-molecule binding sites (Fig. 3e, dataset size: 14,412 residues from 224 structures), ATOMICA achieves an AUPRC of 0.220 and shows higher performance than the best baseline (RiNALMo) by +0.023 (Fig. 3f). Finally, we consider binding-pocket–level multi-class classification for identifying the ligand bound to an RNA pocket (Fig. 3g, dataset size: 290 pockets). Labels correspond to the three most frequent RNA-bound ligands: paromomycin (PAR), gentamycin C1A (LLL), and aminoglycoside TC007 (8UZ). ATOMICA attains an F1 macro of 0.721 and F1 micro of 0.795, exceeding the best baseline (RiNALMo) by +0.244 on F1 macro and +0.068 on F1 micro (Fig. 3h).

ATOMICA representations facilitate accurate annotation of ligands to protein pockets.

To evaluate ATOMICA on protein pocket annotation, we use the MaSIF-ligand task, which predicts the identity of the bound small molecule from the structure of a protein binding pocket across 2,509 pockets (ADP, CoA, FAD, heme, NAD, NAP, or SAM; Fig. 3i). Compared to established pocket descriptors [16, 23, 70–72], ATOMICA achieves higher performance with a F1 macro of 0.834 and F1 micro of 0.859, outperforming MASIF by +0.096 on F1 macro and +0.072 on F1 micro (Fig. 3j).

We also compare to protein language models (ESM-2 [4], ESM-C [73], ESM-3 [74], SaProt [6], ProstT5 [7]). ProstT5 achieves the highest performance with an F1 macro of 0.858 and an F1 micro of 0.892, outperforming ATOMICA by +0.024 on F1 macro and +0.033 on F1 micro (Fig. 3k). ATOMICA is also computationally efficient, with 7.5M parameters, and achieves performance comparable to protein language models with at least 600M parameters. It exceeds ESM-C by 0.036 in F1 macro and 0.026 in F1 micro, and SaProt by 0.167 in F1 macro and 0.059 in F1 micro.

Cross-modality interface comparison for orthosteric PPI inhibitors.

Orthosteric inhibitors of PPIs include small molecules that bind at a protein-protein interface and competitively block partner binding. Because orthosteric inhibitors engage the same surface features as the native partner, we compare protein-inhibitor and PPI embeddings in the ATOMICA shared latent space and test whether embedding similarity localizes to the inhibited PPI region.

We use the curated 2P2Idb database [75], which contains matched PPI structures and their protein-inhibitor PDB structures. After quality filtering, we examined 18 protein-peptide complexes with 268 protein-inhibitor complexes and six protein-protein complexes with 187 protein-inhibitor complexes (Methods 5). Using the ATOMICA embedding notation in Fig. 1c, for each protein-peptide complex and their matched protein-inhibitor ligand complex, we embed distances between peptide and inhibitor blocks $\text{dist}_{\text{ATOMICA}}(\mathbf{h}^{\text{block}}_{\text{peptide}}, \mathbf{h}^{\text{block}}_{\text{inhibitor}})$ and the spatial distance between the aligned coordinates, $\mathbf{x}$, between the blocks $\text{dist}_{\mathbb{R}^3}(\mathbf{x}^{\text{block}}_{\text{peptide}}, \mathbf{x}^{\text{block}}_{\text{inhibitor}})$. Across 268 matched pairs, each system contained a median of 60 inhibitor-peptide block pairs (IQR: 40-79). We quantify localization with Fold Change@10: the proportion of inhibitor-peptide block pairs with $\text{dist}_{\mathbb{R}^3} < 4$ Å in the top-10 pairs with the lowest $\text{dist}_{\text{ATOMICA}}$ relative to the reference proportion among all inhibitor-peptide block pairs in the matched pair (additional results for Fold Change values are

in Fig. S4 and Fig. S5). We observe overrepresentation above reference in 14/18 protein-peptide complexes (78%), spanning 161/268 matched protein-inhibitor complexes (Fig. 4d). Consistently, $\text{dist}_{\text{ATOMICA}}$ correlates with $\text{dist}_{\mathbb{R}^3}$ across matched structures. Nine out of 18 exhibit a significant positive correlation (FDR q < 0.05, Table S3), exceeding the number of protein-peptide complexes expected by chance (binomial test p=6.28×10^{-8}).

In Fig. 4b, we explore the menin and mixed lineage leukemia (MLL) protein-peptide interaction, which plays a central role in acute leukemias [76], and is inhibited by the ligand MIV-7 (PDB ligand code 2SE) [77]. We compare all 144 inhibitor-peptide block pairs and rank them by ATOMICA embedding similarity. The 10 most similar pairs are closer in the aligned 3D structures than the 10 least similar pairs, indicating that blocks with similar embeddings tend to localize to corresponding regions of the native interface.

For protein-protein (A-B) complexes, we compare the inhibitor interface embedding $\mathbf{h}_{\text{inhibitor}}^{\text{interface}}$ to embeddings $\mathbf{h}_{\text{protein B}}^{\text{interface},i}$ of 1,000 local surface patches sampled on protein B. We rank these patches by $\text{dist}_{\text{ATOMICA}}(\mathbf{h}_{\text{inhibitor}}^{\text{interface}}, \mathbf{h}_{\text{protein B}}^{\text{interface},i})$ and test whether the most similar patches are located near the native A-B binding site on protein B. We evaluate retrieval using Fold Change@10, where positives are sampled surface patches on protein B whose centers lie within 12 Å of the native A-B binding site on B. We set $\text{dist}_{\mathbb{R}^3} < 12$ Å threshold to the 25th percentile of sampled distances and compare the fraction of positives among the top 10 retrieved patches to the fraction among all sampled patches. Fold Change@10 is greater than 1 in all six protein-protein complexes, covering 186 of 187 matched protein-inhibitor complexes (Fig. 4f). In addition, after multiple-testing correction, 5 of 6 complexes show a significant positive correlation between embedding similarity and proximity to the native site (FDR q < 0.05; Table S4). These results indicate that lower $\text{dist}_{\text{ATOMICA}}$ is associated with closer proximity to the native protein-protein binding site across distinct protein-protein interactions.

For the KRAS-SOS1 (Fig. 4g,h) and HRAS-SOS1 (Fig. 4i,j) complexes, we rank 1,000 sampled surface patches on SOS1 by their similarity to the inhibitor embedding and compare the top 10 and bottom 10 patches. For a covalent KRAS inhibitor [78] and a small-molecule HRAS inhibitor [79], the top-ranked patch is closer to the native RAS-SOS1 interface, whereas the bottom-ranked patch is farther away. These examples show that inhibitor embedding similarity is highest near the native protein-protein interface.

Structure-based nomination of ion and cofactor identities in dark proteome pockets.

Ion and cofactor binding sites can recur across diverse folds and sequences [80, 81]. We use ATOMICA to generate structure-based hypotheses about ligand identity for predicted pockets in Foldseek-defined dark clusters [44–46]. In AFDB Foldseek, 711,705 clusters (30.9%) are dark clusters (Fig. 4a) [45, 82] that are enriched for proteins with limited functional annotation [45, 83]. We treat predicted ligand identities as a functional signal to prioritize biochemical hypotheses and guide follow-up experiments.

We finetune ATOMICA-Ligand to predict ligand identity from pocket geometry and apply it to high-confidence AF2 representatives (pLDDT>90) from 33,482 dark clusters. We apply PeSTo [84] to predict candidate ion and small-molecule pockets, yielding 2,851 proteins with ion-binding sites and 969 with small-molecule sites. ATOMICA-Ligand predicts identities for 9 metal ions and 12 common cofactors for protein pockets (Fig. 4b, Methods 6), with predictions for 2,565/2,851 proteins and 81/969 proteins, respectively.

As orthogonal computational support, we score protein–ligand complexes formed by pairing each predicted pocket with the ligand assigned by ATOMICA-Ligand using AlphaFold3 ipTM [85]. ATOMICA-Ligand-assigned ligands yield higher ipTM than random reference assignments (ions: KS=0.11, $p < 0.001$; small molecules: KS=0.54, $p < 0.001$; Fig. 4c). In Section S1, we use these pocket-level predictions to interpret putative biochemical roles.

Experimental validation of heme binders predicted by ATOMICA.

Heme is a cofactor involved in electron transfer, oxygen transport, enzymatic catalysis, and cellular signaling, and it accounts for 3.4% of protein-small molecule interaction complexes in the ATOMICA dataset. We apply ATOMICA-Ligand to nominate heme binders in the dark proteome, selecting nine candidates, and six were produced at sufficient yield for experimental testing (Methods 8.1).

We assess binding via red-shifts in the Soret band relative to free heme (390 nm), using BSA as a negative control (Fig. 5a). Five of six candidates show experimental evidence of heme binding (Methods 8.2). Four confirmed binders are predicted by ATOMICA-Ligand to bind covalently: A0A7W1B5T5 (ATOMICA-Ligand score=0.997), A0A2V6P8N7 (ATOMICA-Ligand score=0.690), A0A7W0X6V6 (ATOMICA-Ligand score=0.992), and A0A1T4N4K0 (ATOMICA-Ligand score=0.874) (Fig. 5b-e). UniProt assigns "cytochrome c"-related function and heme-binding terms to these candidates via automatic rule-based assertion (ARBA/UniRule) that is

driven primarily by sequence-derived family/domain signatures and taxonomy. ATOMICA complements these assignments by predicting heme binding from 3D pocket geometry.

To test whether ATOMICA-Ligand is capturing information beyond the canonical CXXCH motif found in many cytochrome c proteins [86], we performed a motif-controlled evaluation. We keep only proteins that contain CXXCH such that the motif alone provides no discrimination, and ATOMICA-Ligand still separates heme binders from non-binders with AUROC 0.849 (vs. 0.500 for a constant CXXCH-motif-only predictor) (Fig. S6a).

For dark protein V5BF69 (ATOMICA-Ligand score=0.663), which does not contain the canonical CXXCH heme binding motif, we show that it does bind heme experimentally (Fig. 5f). The sequence around the axial cysteine (C) coordinated to heme, AHGVCAG, is not captured by other common heme sequence motifs (GX[HR]XC[PLAV]G [86], FXXGXRXCXG [87]). Although V5BF69 contains partial overlap with motif families through the CXG pattern, a CXG-motif-controlled analysis shows that among proteins containing CXG, ATOMICA-Ligand still separates heme binders from non-binders with AUROC 0.990 (vs. 0.500 for a constant CXG-motif-only predictor) (Fig. S6d).

Discussion

ATOMICA is based on the premise that intermolecular interfaces, rather than isolated molecules, provide a transferable basis for representation learning across biomolecular interaction settings. Pretrained on more than two million complexes spanning proteins, small molecules, metal ions, amino acids, and nucleic acids, ATOMICA learns an embedding space that captures interface-level structure, reflects physicochemical similarities across molecular modalities, and enables zero-shot prioritization of blocks involved in annotated contacts. A central motivation for a shared interaction space is knowledge transfer across interaction types that are unevenly represented in structural databases. Consistent with this goal, multimodal pretraining improves masked block recovery for lower-data interfaces compared with training on individual modality pairs, suggesting that increased modality diversity can improve generalization. Across RNA 3D structure-function and protein pocket ligand classification benchmarks, ATOMICA outperforms several RNA foundation models and protein-pocket encoders and performs comparably to much larger protein language models. The shared embedding space also enables cross-modality interface comparison in

PPI inhibitor systems, where inhibitor embeddings localize to regions near the corresponding native protein-protein interfaces. Finally, we apply ATOMICA-Ligand to nominate ion and cofactor identities for predicted pockets in structurally uncharacterized protein families and experimentally validate heme binding for five candidates.

ATOMICA depends on access to accurate molecular structures, which remain unavailable for many complexes because of crystallization challenges, conformational heterogeneity, and disorder. Recent advances in structure prediction, including AlphaFold3 [1] and RoseTTAFold [2], may extend the use of ATOMICA to complexes that have not been solved experimentally. The cross-modality PPI analyses are also limited to curated systems with matched native and inhibitor-bound structures, which enables controlled geometric evaluation but restricts coverage. For the dark proteome analysis, we use AlphaFold3 ipTM scores as an orthogonal computational check for ATOMICA-Ligand predictions. These scores reflect confidence in predicted complex geometry rather than experimental binding. We prioritized heme for initial validation because binding can be measured sensitively by UV–vis spectroscopy using limited material. Extending validation to additional ions and cofactors will require further experiments.

While ATOMICA performs well on several structure-based benchmarks, protein language models remain competitive and can outperform ATOMICA on protein pocket ligand annotation. This likely reflects the value of evolutionary information captured from sequence, including conservation and covariation, which is not fully available to a structure-only model. These results suggest that interface-centered structural pretraining is not uniformly optimal across all prediction settings. An important next step is to develop joint sequence-structure representations that connect ATOMICA's interaction-centered structural embedding space with protein and RNA language-model embeddings.

Understanding molecular interactions requires representations that capture local geometry and chemistry across diverse molecular partners. Here we introduce ATOMICA, an interaction-centered geometric model pretrained on a large multimodal dataset, and show that its shared embedding space supports cross-modality comparison and fine-tuning across modalities. These results establish that multimodal interface pretraining can support more unified representations of biomolecular interactions across molecular classes.

Data availability. Datasets are available on Harvard Dataverse at https://doi.org/10.7910/DVN/4DUBJX.

Code availability. The code to reproduce the results, documentation and usage examples are available at https://github.com/mims-harvard/ATOMICA. The model weights are available at https://huggingface.co/ada-f/ATOMICA. Details are also available at https://zitniklab.hms.harvard.edu/projects/ATOMICA.

Acknowledgements. A.F. is supported by the Kempner Graduate Fellowship at Harvard University. M.D. is supported by an NSERC/FRQNT Postdoctoral Fellowship at the Massachusetts Institute of Technology. We gratefully acknowledge the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, awards from Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists, Sanofi iDEA-iTECH, Pfizer Research, Chan Zuckerberg Initiative, John and Virginia Kaneb Fellowship at Harvard Medical School, Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence, and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. J.L. is supported, in part, by NIH grants HL155107, HL166137, and HG007691; by AHA grants AHA957729 and AHA24MERIT 1185447; and EU HorizonHealth 2021 grant 101057619 (REPO4EU). M.Z., B.L.P., and M.D. are supported, in part, by the Biswas Computational Biology Initiative in partnership with the Milken Institute. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Authors contributions. A.F. developed and implemented the ATOMICA model, conducted extensive benchmarking, and performed detailed analyses of its performance and applications. Z.Z. contributed to the interpretability framework and developed the ATOMICAScore analysis. A.Z. assisted with the analysis of disease-associated proteins and the evaluation of ATOMICANet. All authors contributed to writing and editing the manuscript. J.L. provided guidance on disease pathway analysis and interpretation of physicochemical features of molecular interactions. M.D. and B.L.D. performed experiments to validate heme binding for ATOMICA predictions. A.F. and M.Z. designed the study.

Competing interests. The authors declare no competing interests.

Figures

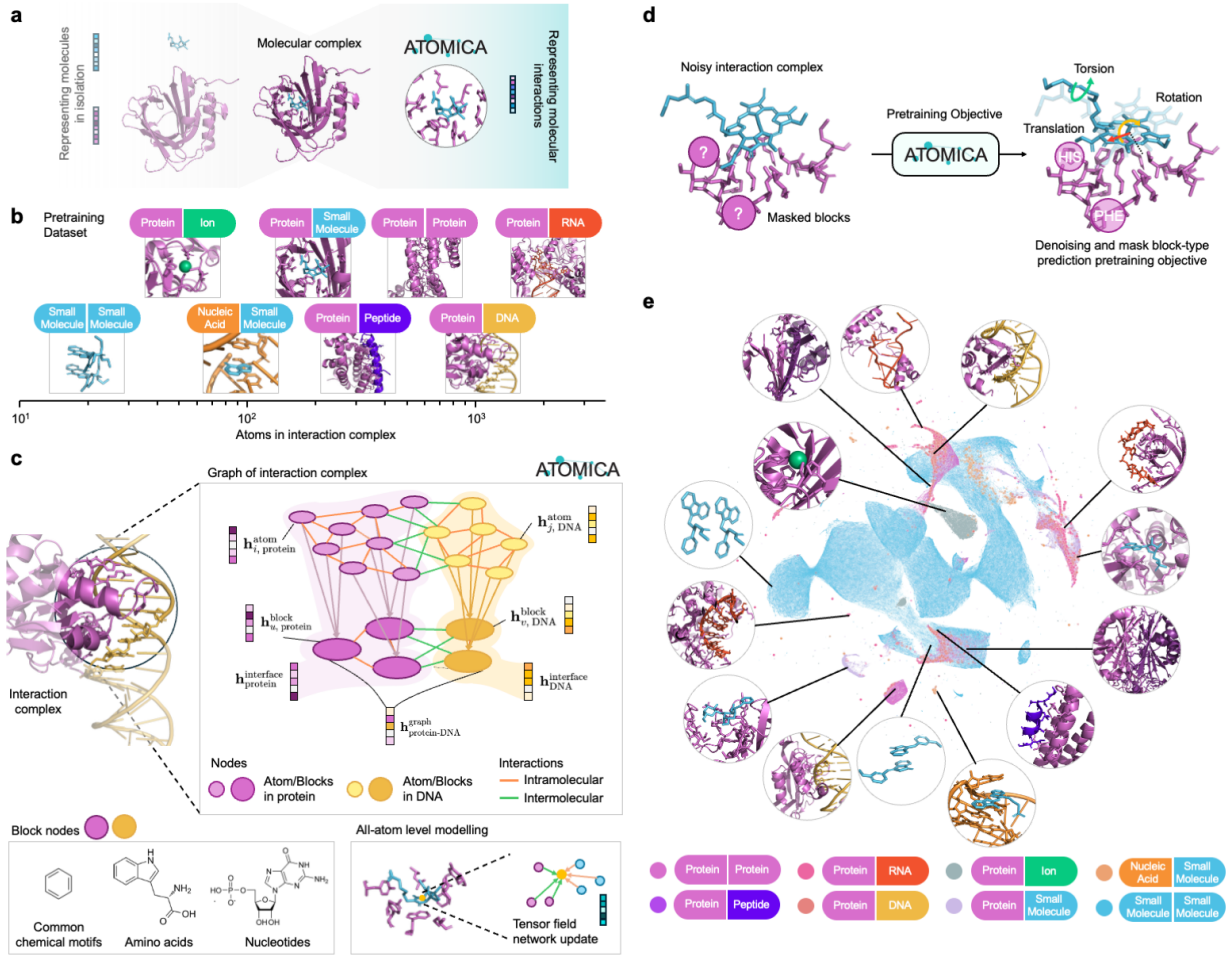

Figure 1: Overview of ATOMICA pretraining data, architecture, and latent space. **a** Complementary to models learning representations of individual molecules, ATOMICA focuses on representations of intermolecular interfaces within complexes, providing a shared interaction-centric representation for interaction tasks. **b** Input modalities and the number of atoms in the interaction complexes that ATOMICA are pretrained on. **c** Overview of ATOMICA architecture, interaction complexes are modeled at the atom and block level. Message passing between nodes at each level is done via intermolecular and intramolecular edges. All embeddings that can be derived from ATOMICA are also shown. **d** Self-supervised pretraining objectives for generating ATOMICA embeddings. **e** UMAP of latent space of all interaction complexes seen during pretraining.

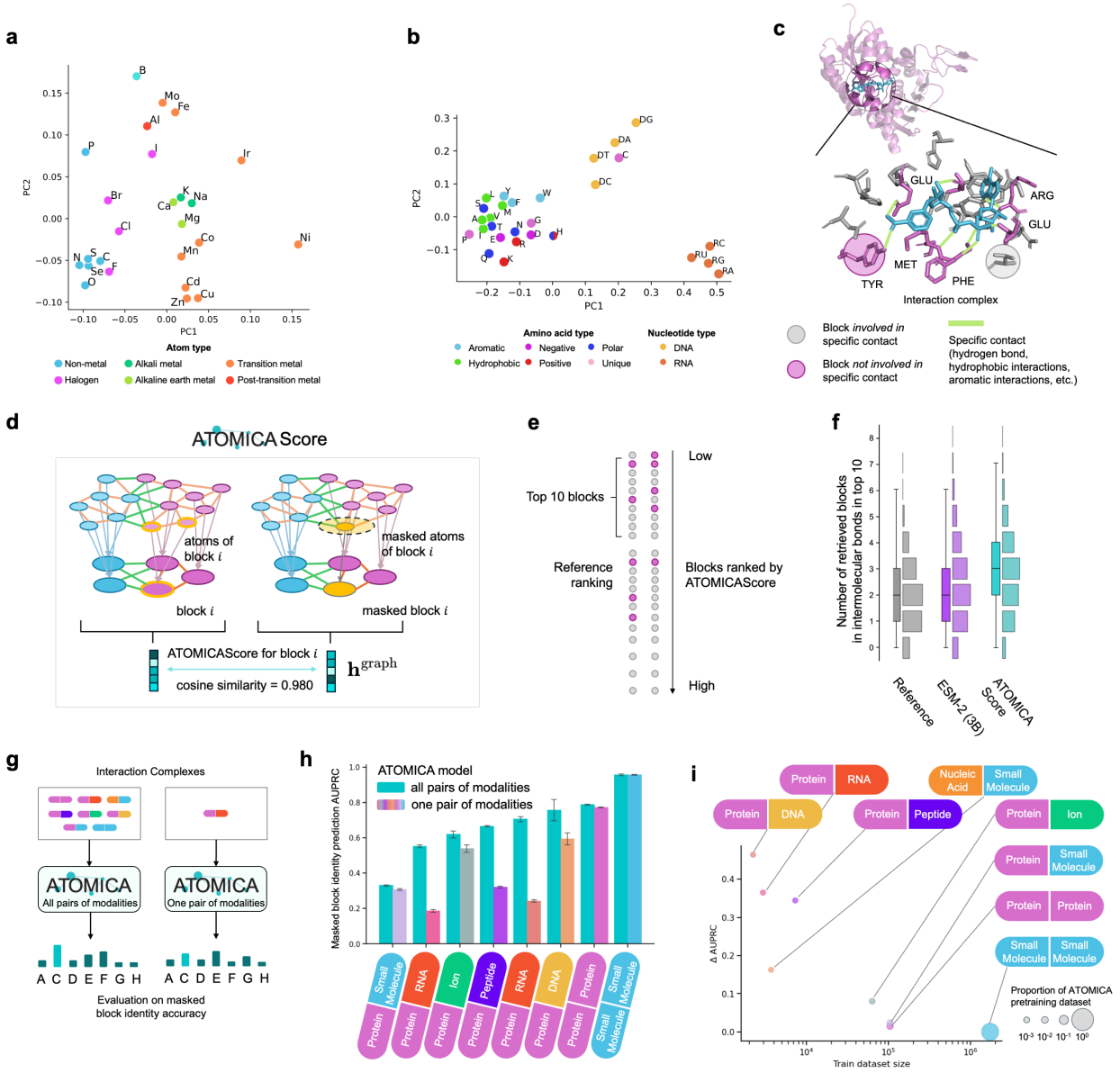

Figure 2: Analysing latent space of ATOMICA. a Principle components of mean embedding of elements in the pretraining validation and test set. **b** Principle components of mean embedding of amino acids and nucleotides in the pretraining validation and test set. **c** Some blocks in the interaction complex are involved in specific contacts (hydrogen bonds, hydrophobic interactions, pi-stacking, etc.). **d** Definition of ATOMICAScore for block i, cosine similarity between interaction graph with block i masked and unmasked. **e** Nomination of blocks involved in specific contacts at interfaces based on ATOMICAScore. The reference ranking is determined by random ordering of the blocks at the interaction interface. **f** Number of blocks involved in specific contacts in the top 10 nominated blocks of ATOMICAScore, ESM-2 (3B parameters), and reference for protein-small molecule complexes in the pretraining test set. ATOMICA nominates more amino acids involved in specific contacts than ESM-2 and the reference. **g** Schema to test generalizability of representations learned by ATOMICA trained on all pairs of modalities compared to models trained on one pair of modalities. We evaluate quality of representations based on masked block identity accuracy. **h** Masked block identity AUPRC for ATOMICA trained on all pairs of modalities and ATOMICA trained on one pair of modalities. ATOMICA trained on all pairs of modalities outperforms ATOMICA trained on one pair of modalities. Error bars represent standard deviation in AUPRC across 5 seeds for randomly masking blocks. **i** Increase in AUPRC between training on one pair of modalities and all pairs of modalities compared to dataset size of the one pair of modalities. Performance gains scale with dataset size increase, with the largest improvements in AUPRC observed for pairs of modalities with the least data.

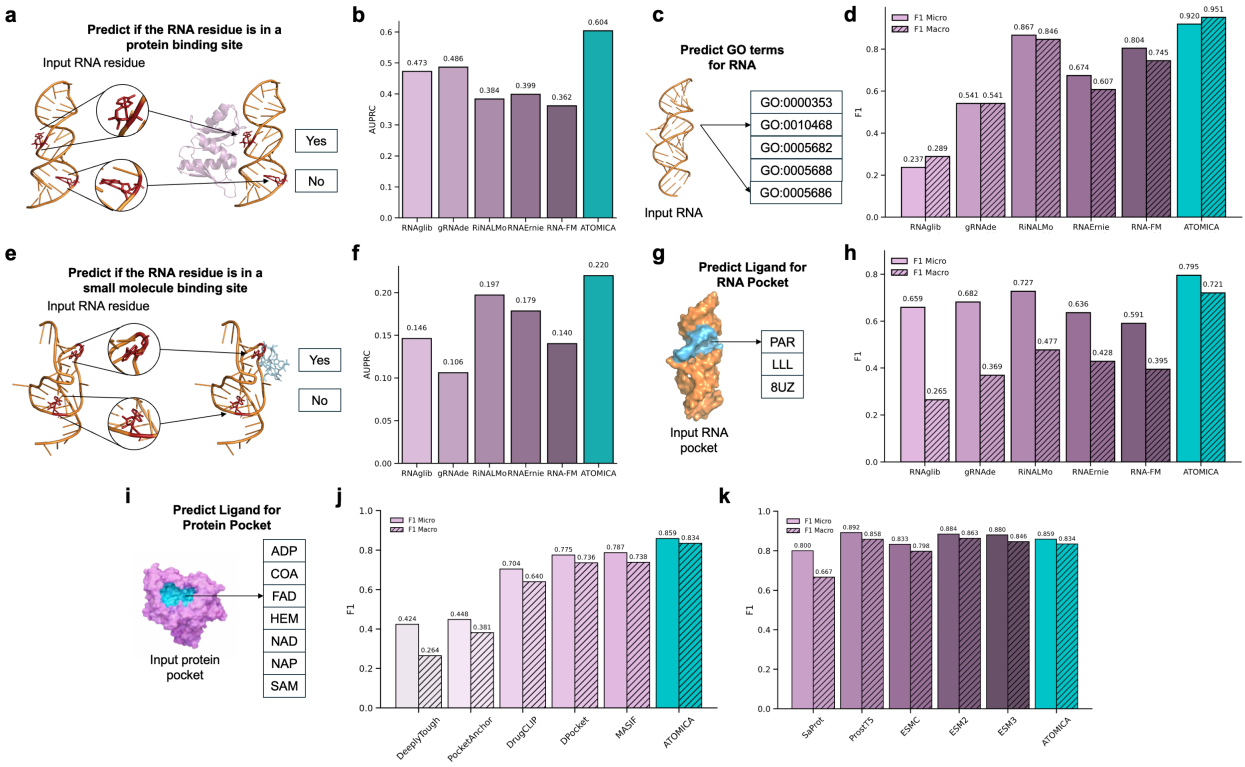

Figure 3: Performance of ATOMICA on protein and RNA benchmarks. Binary classification task for prediction if the RNA residue is in a protein binding site, **a** Schema of the task and **b** Evaluation of models with AUPRC. RNA functional annotation from 3D structure using Gene Ontology (GO) terms, **c** Schema of the task and **d** Evaluation of models with F1 micro and F1 macro. Binary classification task for prediction if the RNA residue is in a ligand binding site, **e** Schema of the task and **f** Evaluation of models with AUPRC. Binding-pocket–level multi-class classification for identifying the ligand bound to an RNA pocket, **g** Schema of the task and **h** Evaluation of models with F1 micro and F1 macro. Binding-pocket–level multi-class classification for identifying the ligand bound to an protein pocket, **i** Schema of the task, evaluation of ATOMICA with F1 micro and F1 macro **j** Compared to protein-pocket encoders and **k** Compared to protein language models. Additional results are available in Fig. S3.

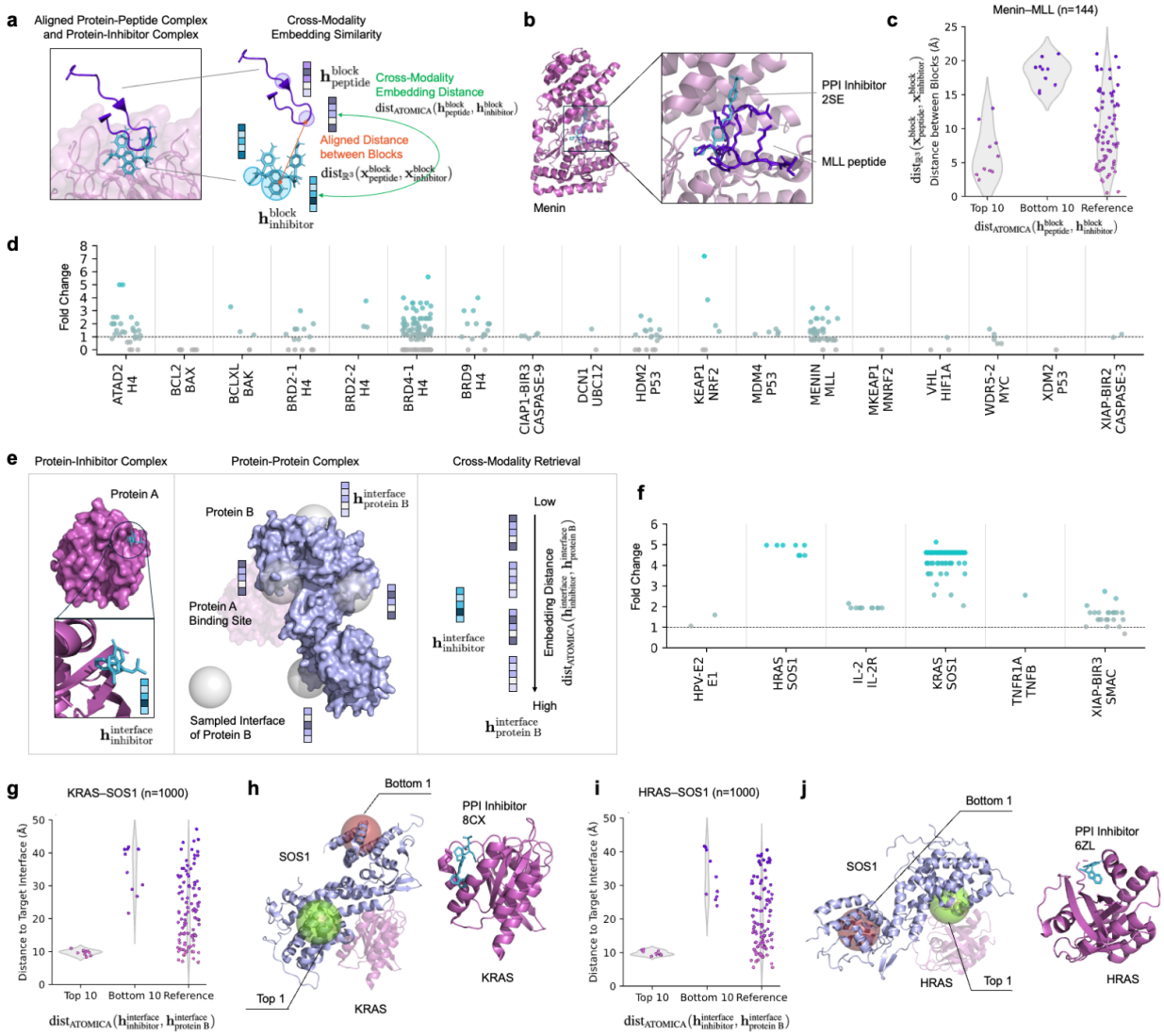

Figure 4: Cross-modality interface comparison of orthosteric PPI inhibitors. **a** Alignment of matched protein–peptide and protein–inhibitor structures. We compare cross-modality ATOMICA embedding distance, $\text{dist}_{\text{ATOMICA}}(\mathbf{h}^{\text{block}}_{\text{inhibitor}}, \mathbf{h}^{\text{block}}_{\text{peptide}})$, where $\mathbf{h}^{\text{block}}_{\text{inhibitor}}$ are inhibitor blocks and $\mathbf{h}^{\text{block}}_{\text{peptide}}$ are peptide blocks, to the aligned 3D distance between the corresponding blocks, $\text{dist}_{\mathbb{R}^3}(\mathbf{x}^{\text{block}}_{\text{inhibitor}}, \mathbf{x}^{\text{block}}_{\text{peptide}})$. **b** Example alignment for the Menin–MLL peptide interaction and the Menin–MIV-7 (ligand code 2SE) inhibitor complex. **c** For the Menin–MLL complex, the top-10 (lowest $\text{dist}_{\text{ATOMICA}}$) peptide–inhibitor block pairs are spatially closer in the aligned structures than the bottom-10 pairs of 144 total pairs. The reference shows the distribution over all peptide–inhibitor block pairs. **d** Fold Change@10 for peptide–inhibitor block pairs with an aligned distance ≤ 4 Å, shown across protein–peptide complexes and their matched protein–inhibitor complexes. Additional results @5 and @20 are available in Fig. S4. **e** Cross-modality retrieval for protein–protein (A–B) interactions: we compare the inhibitor interface embedding, $\mathbf{h}^{\text{interface}}_{\text{inhibitor}}$, to ATOMICA interface embeddings, $\mathbf{h}^{\text{interface}}_{\text{protein B}}$, of sampled interfaces on protein B and rank sampled interfaces by embedding distance to the inhibitor, $\text{dist}_{\text{ATOMICA}}(\mathbf{h}^{\text{interface}}_{\text{inhibitor}}, \mathbf{h}^{\text{interface}}_{\text{protein B}})$. **f** Fold Change@10 for retrieval on protein B for interfaces whose centers lie within 12 Å of the native A–B interface on B (12 Å corresponds to the 25th percentile of sampled-interface distances to A). Additional results @5 and @20 are available in Fig. S5. **g,h** KRAS–SOS1: distances to the native interface for the top-10 (lowest $\text{dist}_{\text{ATOMICA}}$) and bottom-10 (highest $\text{dist}_{\text{ATOMICA}}$) retrieved from 1,000 SOS1 interfaces (**g**), and visualization of the top-1 and bottom-1 retrieved interfaces for SOS1 (**h**). **i,j** HRAS–SOS1: analogous results showing top-10 vs. bottom-10 retrieved from 1,000 interfaces (**i**) and visualization of the top-1 and bottom-1 interfaces (**j**).

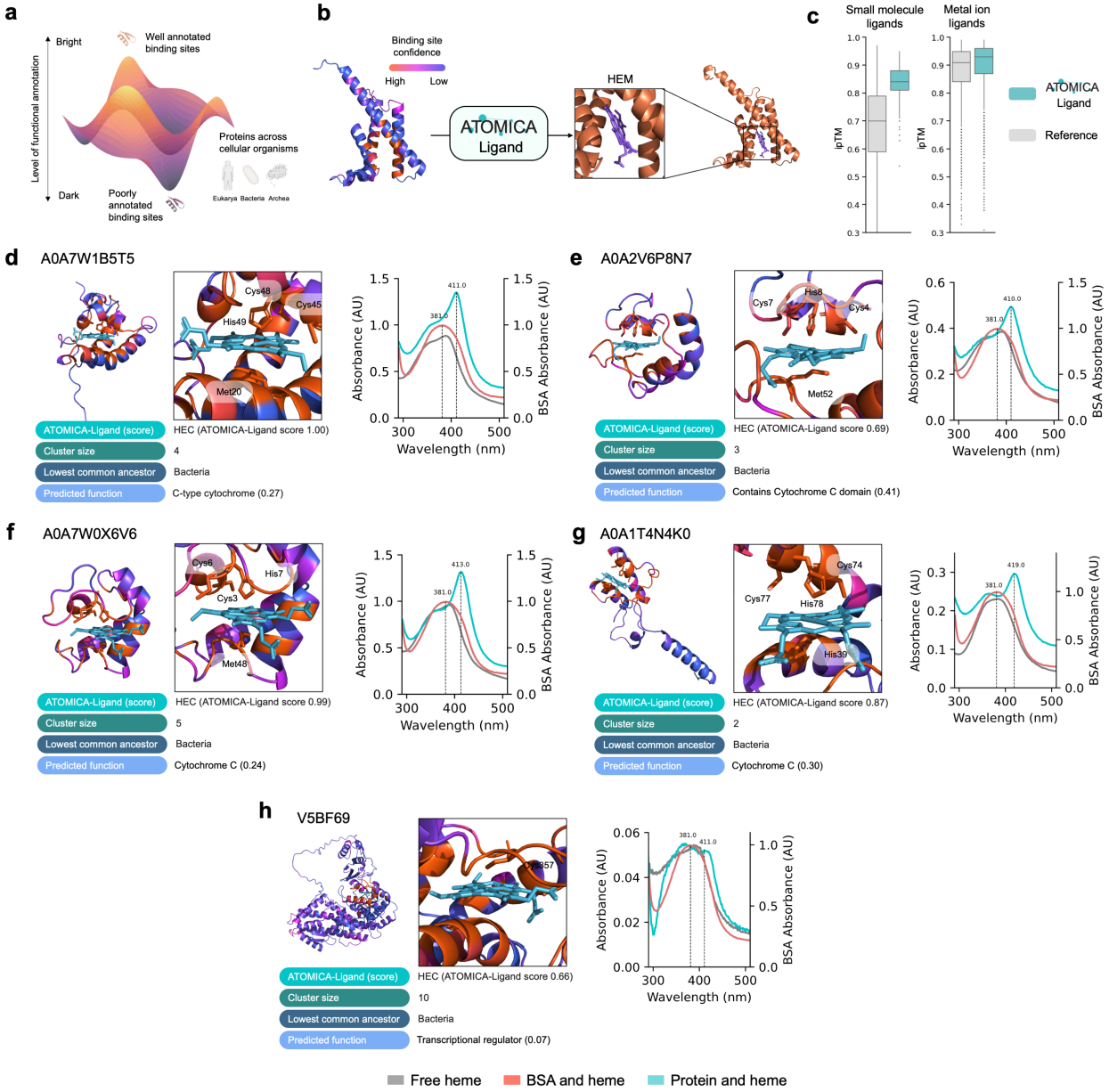

Figure 5: Dark proteome annotation with ATOMICA and experimental validation of dark protein heme binders. **a** Annotation of binding sites for proteins across cellular organisms. **b** Prediction of ligands for metal ion and small molecule binding sites of proteins in the dark proteome. **c** AlphaFold3 ipTM scores of complexes from ATOMICA-Ligand annotated small molecule and metal ion compared to reference. Visualization of AlphaFold3-predicted structures for experimentally validated protein-heme annotations made by ATOMICA-Ligand. Cluster size and lowest common ancestor are derived from Foldseek clustering of the AlphaFold Protein Structure Database [45]. Predicted functions are obtained from ProtNLM, a sequence-based protein function prediction model [88]. To confirm protein binding to heme, we use UV-Vis spectroscopy and the red-shifting of the Soret band at ∼390 nm for free heme. The Soret band of heme and bovine serum albumin (BSA) is also shown as a negative control. The following dark proteins are confirmed to bind to heme **d** A0A7W1B5T5, **e** A0A2V6P8N7, **f** A0A7W0X6V6, **g** A0A1T4N4K0, and **h** V5BF69.

Methods

The Methods describe (1) the curation of datasets, (2) hierarchical all-atom graph neural network architecture of ATOMICA, (3) training ATOMICA, (4) evaluating ATOMICA on RNA and protein annotation, (5) cross-modality interface comparison for orthosteric PPI inhibitors, (6) dark proteome binding site analysis with ATOMICA-Ligand, (7) metrics and statistical analyses, and (8) details on experimental validation of ATOMICA-predicted heme binders.

1 Curation of datasets

1.1 Small molecule structures

We extract structures of small molecule interactions from the Cambridge Structural Database (CSD) v2023.2.0. The database was filtered for all CSD entries that satisfied the following criteria: organic, not polymeric, has 3D coordinates, no disorder, no errors, no metals, had only one SMILES string describing the crystal entry (in other words, each crystal is comprised of only one chemical compound), and molecules with 6-50 heavy atoms. CSD entries are unit cells of infinitely repeating crystal lattices. For our purposes of learning intermolecular interactions, we sampled many pairs of intermolecular interactions to represent all examples of intermolecular interactions in a given unit cell. Given an entry of the CSD, we iterate through each unique conformer in the unit cell and extract all pairs of interactions with neighboring peripheral conformers that are within 4 Å to the central conformer using the CSD Python API. In total, there are 1,767,710 structures of molecular pairs from 375,941 CSD entries. Inspired by fingerprint-based similarity measures used in chemistry [89], we use a one-hot encoding of the molecular complex from a vocabulary of 290 common chemical motifs [52] (which is later used in our assignment of blocks to small molecules). We use Manhattan distance between the embeddings to sample 1,000 molecular complexes and their 100 nearest neighbors, giving a total of 10,000 molecular complexes for validation and test splits, respectively, that are distinct from the training set. Manhattan distance is applied to split small molecules into train/validation/test sets as it measures molecular similarity based on our vocabulary of common chemical motifs, and it is fast to compute at our scale, enabling large-scale nearest-neighbor retrieval for dataset construction.

1.2 General biomolecular structures

We extract the structures of the interacting molecules from QBioLiP (June 2024). This includes structures of proteins that interact with ions, ligands, DNA, RNA, peptides, and proteins, and nucleic acids interacting with ions and ligands from the Protein Data Bank (PDB). For proteins, DNA, and RNA, we crop the complex to keep all residues within 8 Å to any atom, amino acid, or nucleic acid residue in the other molecule. In total, there are 124,541 protein-protein interaction complexes, 119,017 protein-small molecule interaction complexes, 74,514 protein-ion interaction complexes, 8,475 protein-peptide interaction complexes, 5,185 nucleic acid-ligand interaction

complexes, 3,511 protein-RNA interaction complexes, and 2,750 protein-DNA interaction complexes. For protein-ion, protein-small molecule, protein-peptide, and protein-protein molecular complexes, we cluster each set of complexes with 30% protein sequence similarity using MMseqs2 with a coverage of 80%, sensitivity of 8, and cluster mode 1[90]. For protein-protein complexes, we also ensure that for any two complexes in different clusters, there is a maximum of 30% sequence similarity between all chains in the two complexes. For protein-RNA and protein-DNA complexes, we cluster by 30% protein sequence similarity and 30% nucleotide similarity using MMseqs2 with the same settings as above, this ensures that complexes in different clusters have a maximum of 30% protein sequence similarity and 30% nucleotide sequence similarity. For nucleic acid-ligand structures, we cluster based on 30% nucleotide sequence similarity. Finally, we split clusters into train, validation, and test splits using an 8:1:1 ratio.

1.3 Construction of hierarchical graphs of interacting molecules

Given the atomic structure of two molecules interacting, an atom-level graph is constructed. Each atom in the complex maps to an atom node in the graph with the following features: element and 3D coordinates of the atom. Intramolecular atom edges are defined for each atom to the k nearest atoms in the same molecule. Intermolecular atom edges are defined for each atom to the k nearest atoms in the other molecule. In total, there are 118 atom types based on the elements of the periodic table.

Atom nodes are connected to block nodes. The block nodes have the following features: block type and 3D coordinates of the block given by the mean of the atomic coordinates of the atoms in the block. Each atom is connected to one block node. For proteins, peptides, DNA, and RNA, we define the atoms that belong to a given block by the amino acid and nucleotide residues.

For small molecule ligands, we follow the definition of Kong et al. [52] to group atoms of small molecule ligands into blocks. In their framework, small molecule ligands were tokenized into blocks representing common chemical motifs using a graph-based Byte Pair Encoding algorithm that identifies the most common multi-atom patterns from ligands found in the PDB. This results in 290 common chemical motifs. Atom nodes are then grouped into fragments by matching against the pre-defined vocabulary. The algorithm greedily merges neighboring atoms into the most frequent substructure patterns present in the vocabulary until no further vocabulary matches were found. Each identified fragment is assigned its corresponding vocabulary symbol and its constituent atoms grouped into a single block. Atoms of sections of the molecule that cannot be assigned to a the vocabulary become blocks comprised of one atom. This hierarchical representation enables the model to process ligands as sequences of meaningful chemical building blocks rather than individual atoms and encodes the covalent connectivity of atoms to the model.

Intramolecular block edges are defined for each atom to the k nearest blocks in the same molecule. Intermolecular block edges are defined for each atom to the k nearest blocks in the other molecule. In total, there are the following block types: 20 for canonical amino acids, 4 for DNA nucleotides, 4 for RNA nucleotides, 290 for small molecule fragments, and 118 for elemental blocks.

In addition, there are three special block types: mask, unknown, and global. The mask

node is applied at pretraining for masked identity prediction of blocks. Unknown nodes are used for nodes that do not fall into the defined vocabulary, such as non-canonical amino acids and nucleotides. There are also two atom global-type nodes at the atom and block level. The two global nodes are connected to all nodes in each molecule at their respective level.

2 Hierarchical all-atom geometric deep learning model

Overview. ATOMICA uses a SE(3)-equivariant 3D message passing network on graphs of molecular complexes to learn representations that are informative of the intermolecular interactions between molecules.

Given is a pretraining dataset of graphs of molecular complexes, $\mathcal{D} = \{G^i \mid i = 1, \ldots, N\}$. Our goal is to pretrain a model $\mathcal{F}$ on $\mathcal{D}$ such that it generates representations $\mathbf{h}_i = \mathcal{F}(G^i)$ for every intermolecular complex G^i that are chemically informative. In addition to representation learning, $\mathcal{F}$ can also be finetuned on datasets of labeled graphs of molecular complexes, such as $\mathcal{S} = \{(G^i_{\text{target}}, y_i) \mid i = 1, \ldots, M\}$, where $M \ll N$ to predict y_i for every G^i_{target}.

Atom-level representation learning. Here we outline the SE(3)-equivariant 3D message passing network for ATOMICA on the nodes of the graph G^i. Several rotational equivariant neural networks have been introduced for modeling molecules [53, 91–93]. We build on the E(3)-equivariant neural network layers presented by tensor-field networks implemented in e3nn [94] and DiffDock [55]. Message passing for the intermolecular edges and intramolecular edges is done separately, but the message passing framework for the two edge types is the same.

The feature vector of atom ($\mathbf{h}_a^{\text{atom}}$) node a in G^i is a geometric object comprised of a direct sum of irreducible representations of the O(3) symmetry group. The feature vectors $\mathbf{h}_{a,(\lambda,p)}^{\text{atom}}$ are indexed with λ, p, where $\lambda = 0, 1, 2, \ldots$ is a non-negative integer denoting the rotation order and $p \in \{\text{o}, \text{e}\}$ indicates odd or even parity, which together index the irreducible representations (irreps) of O(3). In ATOMICA model, we set $\lambda_{\max} = 1$ for $\mathbf{h}_a^{\text{atom}}$, and we denote the number of scalar (0e) and pseudoscalar (0o) irrep features in $\mathbf{h}_a^{\text{atom}}$ with ns, and the number of vector (1o) and pseudovector (1e) irrep features in $\mathbf{h}_a^{\text{atom}}$ with nv.

The atom-type of node a, determined by the element of the atom, is embedded with a normal distribution and trainable weights as a scalar $\text{ns} \times 0\text{e}$. There are L_{GNN} layers of message passing between atom nodes. At each layer l, the node updates for node a in the graph of interaction complex G^i are given by:

$$\mathbf{h}_a^{\text{atom}} \leftarrow \mathbf{h}_a^{\text{atom}} + \text{LN}\left(\frac{1}{|\mathcal{N}_a|} \sum_{b \in \mathcal{N}_a} Y\left(\widehat{\mathbf{r}}_{ab}\right) \otimes_{\psi_{ab}} \mathbf{h}_b^{\text{atom}}\right) \tag{1}$$

$$\text{with } \psi_{ab} = \Psi\left(\mathbf{e}_{ab}, \mathbf{t}_{ab}, \mathbf{h}_{a,(0e)}^{\text{atom}}, \mathbf{h}_{b,(0e)}^{\text{atom}}\right). \tag{2}$$

where LN is layer norm. After each layer l of message passing, $\mathbf{h}_a^{\text{atom}}$ is filtered down to irreps with $\lambda_{\max} = 2$. After L layers the $\mathbf{h}_a^{\text{atom}}$ embedding is projected with a 2-layer MLP to a d_{node}-dimension

vector.

Block-level representation learning. The feature vector of block ($\mathbf{h}_b^{\text{block}}$) node b in G^i is also a geometric object defined in the same way as ($\mathbf{h}_a^{\text{atom}}$). We initialize block nodes using a scalar, $\text{ns} \times 0\text{e}$, trainable embedding of block types.

Let d_{node} be the dimension of $\mathbf{h}_b^{\text{block}}$ and n_{heads} be the number of attention heads. We define $d_h = d_{\text{node}}/n_{\text{heads}}$ as the dimension per head. The multi-head cross-attention operation can be expressed as:

$$\mathbf{h}_b^{\text{block}} \leftarrow \mathbf{h}_b^{\text{block}} + \text{MultiHead}(\mathbf{h}_b^{\text{block}}, \{\mathbf{h}_a^{\text{atom}}\}_{a \in A_b}) \tag{3}$$

where A_b is the set of atoms in block b, and MultiHead is defined as:

$$\text{MultiHead}(\mathbf{h}_b^{\text{block}}, \{\mathbf{h}_a^{\text{atom}}\}_{a \in A_b}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_{n_{\text{heads}}})\mathbf{W}_O \tag{4}$$

and each head is computed as:

$$\text{head}_i = \sum_{a \in A_b} \alpha_{ba} \mathbf{v}_a^{\text{atom},(i)} \text{ with } \alpha_{ba} = \frac{\exp\left(\mathbf{q}_b^{\text{block},(i)} \cdot \mathbf{k}_a^{\text{atom},(i)} / \sqrt{d_h}\right)}{\sum_{v \in A_b} \exp\left(\mathbf{q}_b^{\text{block},(i)} \cdot \mathbf{k}_v^{\text{atom},(i)} / \sqrt{d_h}\right)} \tag{5}$$

where $\mathbf{q}_b^{\text{block},(i)} = \mathbf{h}_b^{\text{block}} \mathbf{W}_Q^{(i)}$, $\mathbf{k}_a^{\text{atom},(i)} = \mathbf{h}_a^{\text{atom}} \mathbf{W}_K^{(i)}$, $\mathbf{v}_a^{\text{atom},(i)} = \mathbf{h}_a^{\text{atom}} \mathbf{W}_V^{(i)}$, and $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d_{\text{node}} \times d_h}$ and $\mathbf{W}_O \in \mathbb{R}^{d_{\text{node}} \times d_{\text{node}}}$. Message passing between the block nodes is specified by the same architecture as the atom nodes described in equation (1) and has separate model parameters.

Graph-level representation learning. To pool $\mathbf{h}_b^{\text{block}} \in \mathbb{R}^d$ for $b \in G^i$ for a graph-level representation $\mathbf{h}_i^{\text{graph}} \in \mathbb{R}^d$, we use multi-head self-attention for L_{pool} layers and sum the output $\mathbf{h}_b^{\text{block}}$ for all $b \in G^i$ for $\mathbf{h}_i^{\text{graph}}$.

Interface-level representation learning. To obtain an interface-level representation $\mathbf{h}_{\text{molecule A}}^{\text{interface}}$ of the molecule A interface in the complex A–B, we take the average over all block embeddings of blocks belonging to molecule A $\mathbf{h}_{\text{molecule},A}^{\text{interface}} = \frac{1}{|\text{molecule A}|} \sum_{\text{b} \in \text{molecule A}} \mathbf{h}_b^{\text{block}}$, where $|\text{molecule A}|$ is the number of blocks, b, in molecule A.

Self-supervised learning with denoising and block identity masking. Node-level denoising as an objective function has been useful for pretraining on 3D coordinate molecular datasets from DFT-generated molecules to prevent over-smoothing of GNNs [95], and it has proven that it is related to learning a force field of per-atom forces [96, 97]. In addition, denoising is linked to score-matching which has also been popular in training generative models [55, 98] as well as unsupervised binding affinity prediction [60]. Thus, this motivates the application of denoising as an objective for self-supervised training.

Given $G^i \in \mathcal{D}$, which is comprised of atom and block nodes from two interacting molecules. $\widetilde{G}^i$ is a perturbed graph created by applying two transformations to a molecule in G^i:

- Rigid rotation and translation: A rotation vector is sampled $\boldsymbol{\omega} \sim p(\boldsymbol{\omega}) = \mathcal{N}_{SO(3)}$ and we apply the rotation of all atom and block coordinates about the center of the selected molecule.

A translation vector is sampled $\mathbf{t} \sim p(\mathbf{t}) = \mathcal{N}(0, \sigma_t^2 \mathbf{I})$ and we apply this translation to all atom and block coordinates of the selected molecule.

- Torsion angle noising: Torsion angles are sampled $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) = \mathcal{N}_{SO(2)^m}$ where m is the number of rotatable bonds in the molecule. For peptides, proteins, RNA and DNA we only perturb rotatable bonds in the side chain.

To predict the rotation score $\mathbf{s}_{\boldsymbol{\omega}} \in \mathbb{R}^3$ and the translation score $\mathbf{s}_{\mathbf{t}} \in \mathbb{R}^3$ from $\widetilde{G}^i$, the node representations at the atom and block level are convolved with the center of the graph using a tensor field network [55]:

$$\mathbf{s} \leftarrow \mathrm{LN}\left(\frac{1}{|\mathcal{A}'|} \sum_{a \in \mathcal{A}'} Y\left(\widehat{r}_{ca}\right) \otimes_{\phi_{ca}} \mathbf{h}_a^{\text{atom}}\right) \text{ with } \phi_{ca} = \Phi\left(\mathbf{e}_{ca}, \mathbf{h}_{a,(0e)}^{\text{atom}}\right), \tag{6}$$

where node $a \in \mathcal{A}'$ are the atom nodes in the perturbed molecule and c is the center of the perturbed molecule. This is a weighted tensor product, with the weights given by a 2-layer MLP, Φ, which takes as input the Gaussian smearing d_{edge}-embedding of the Euclidean distance between coordinates of the center c and node a, and the scalar component of $\mathbf{h}_a^{\text{atom}}$.

Finally, the rotation score is given by the pseudovector irrep component $\mathbf{s}_{\boldsymbol{\omega}} = \Gamma_{\boldsymbol{\omega}}(\mathbf{h}_i^{\text{graph}}) * \mathbf{s}_{(1e)}$ and the translation score is given by the vector irrep component $\mathbf{s}_{\mathbf{t}} = \Gamma_{\mathbf{t}}(\mathbf{h}_i^{\text{graph}}) * \mathbf{s}_{(1o)}$, where $\Gamma_{\boldsymbol{\omega}}$ and $\Gamma_{\mathbf{t}}$ are 2-layer MLPs that project the graph representation of G^i to a single scalar.

To predict the torsion score $\mathbf{s}_{\boldsymbol{\theta}} \in \mathbb{R}^m$, the atom nodes are convolved with the center of the rotatable bonds connecting atoms a_{z0}, a_{z1}. Let z denote the center of one of the rotatable bonds. We connect $\mathcal{N}_z = \{a \mid a \in \mathcal{A}', ||\widehat{r}_{za}|| < 5\mathring{A}\}$ which are all atoms in the perturbed molecule within 5 Å to the center of the bond:

$$\mathbf{h}_z = \frac{1}{|\mathcal{N}_z|} \sum_{a \in \mathcal{N}_z} \left(Y^2(\widehat{\mathbf{r}}_z) \otimes Y(\widehat{\mathbf{r}}_{za})\right) \otimes_{\pi_{za}} \mathbf{h}_a^{\text{atom}} \text{ with } \pi_{za} = \Pi^{(t)}\left(\mathbf{e}_{za}, \mathbf{h}_{a,(0e)}^{\text{atom}}, \mathbf{h}_{a_{z0},(0e)}^{\text{atom}} + \mathbf{h}_{a_{z1},(0e)}^{\text{atom}}\right), \tag{7}$$

The first tensor product is between the second order irreps of the unit direction vector along the two atoms a_{z0}, a_{z1} of the bond z, $Y^2(\widehat{\mathbf{r}}_z)$, and the unit direction vector between the center of the bond and atom a, $Y(\widehat{\mathbf{r}}_{za})$. This is followed by a weighted tensor product with the weights given by a 2-layer MLP, Π, which takes as input the Gaussian smearing d_{edge}-embedding of the Euclidean distance between coordinates of the bond center z and node a, the scalar component of $\mathbf{h}_a^{\text{atom}}$, and the sum of the scalar components of the two atoms in the bond, $\mathbf{h}_{a_{z0}}$ and $\mathbf{h}_{a_{z1}}$. Finally, we sum the scalar and pseudoscalar components of $\mathbf{h}_z$ and project it to a single scalar $\mathbf{s}_{\boldsymbol{\theta}_z}$ using a 2-layer MLP.

We calculate the loss components as follows:

$$l_\omega = ||\mathbf{s}_{\boldsymbol{\omega}} - \nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{\omega})||^2 \tag{8}$$

$$l_t = ||\mathbf{s}_{\mathbf{t}} - \nabla_{\mathbf{t}} \log p(\mathbf{t})||^2 \tag{9}$$

$$l_\theta = \sum_z ||\mathbf{s}_{\boldsymbol{\theta}_z} - \nabla_{\boldsymbol{\theta}_z} \log p(\boldsymbol{\theta}_z)||^2 \tag{10}$$

where $\nabla_{\mathbf{t}} \log p(\mathbf{t}) = -\mathbf{t}/\sigma_t^2$. The values of $\nabla_{\mathbf{t}} \log p(\mathbf{t}), \nabla_{\boldsymbol{\theta}_z} \log p(\boldsymbol{\theta}_z)$ can be calculated by pre-

computing a truncated infinite series following [55, 60].

In addition to denoising, we also pretrain the model by masking out block identities and predicting the masked block identities. For each graph G^i, 10% of blocks are randomly sampled and their block identities are replaced with the special 'mask' block and we denote these blocks as $\mathcal{B}$. For a masked block $b \in \mathcal{B}$, the probability vector of the block identity is predicted with $\hat{\mathbf{y}}_b = \text{Softmax}(\Upsilon(\mathbf{h}_b^{\text{block}}))$, where Υ is a 2-layer MLP. We calculate the masked loss using a cross-entropy loss:

$$l_m = -\frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \mathbf{y}_b \cdot \log(\hat{\mathbf{y}}_b) \tag{11}$$

The pretraining loss is then calculated by a weighted sum of the above loss functions:

$$\mathcal{L} = \beta_\omega l_\omega + \beta_t l_t + \beta_\theta l_\theta + \beta_m l_m$$

3 Training details for ATOMICA

Overview. We pretrain ATOMICA on the training split of biomolecular structures and small molecule structures to generate embeddings of molecular complexes at the atom, block, and graph scales. To learn representations in a self-supervised manner, during training, we apply noise to the atomic coordinates and mask block identities of the input graphs of the molecular complex. At inference time, embeddings from the graphs are generated without noise or masked blocks.

Hyperparameter tuning. We employed a hyperparameter optimization strategy utilizing Ray Tune [99] in conjunction with Optuna [100] and the Asynchronous Successive Halving Algorithm (ASHA) scheduler [101]. The hyperparameter space we search on includes: the number of nearest neighbors to define edges to in the graph $k \in [4, \mathbf{8}, 16]$, dropout in the tensor field network $\in [\mathbf{0.00}, 0.01, 0.05, 0.10]$, edge dimension $d_{\text{edge}} \in [\mathbf{16}, 24, 32]$, node dimension $d_{\text{node}} \in [16, 24, \mathbf{32}]$, and the number of tensor field network layers $L \in [\mathbf{4}, 6, 8]$. The best hyperparameters are shown in bold and chosen based on the lowest validation loss when trained on a random 10% subsample of the training set.

Then for determining the level of noise to apply to the interaction complexes, we conducted a second hyperparameter search on rotation $\sigma_\omega \in [\mathbf{0.25}, 0.5, 1]$, $\omega_{\text{max}} \in [0.25, \mathbf{0.5}, 1]$, translation $\sigma_t \in [0.5, \mathbf{1}, 1.5]$, and torsion $\sigma_\theta \in [0.25, \mathbf{0.5}, 1]$. Again, the best hyperparameters are shown in bold and chosen based on the highest masked block identity prediction accuracy when trained on a random 10% subsample of the training set. For the loss function, we set $\beta_\omega = 1, \beta_t = 1, \beta_m = 0.1$, and the block identities are randomly masked at 10% probability. ATOMICA is trained on the full training set with the above hyperparameters, the learning rate cycles between 1e-4 and 1e-6 using Cosine Annealing Warm Restarts, with a cycle length of 400,000 steps, and the model is trained for 150 epochs.

Implementation. ATOMICA is implemented with PyTorch (Version 2.1.1) [102] and PyTorch Geometric (Version 2.1.1) [103]. Training runs were monitored with Weights and Biases [104]. The final ATOMICA model was trained on 4 NVIDIA H100 Tensor Core GPUs in parallel for 24

days.

4 Evaluation on RNA and protein annotation tasks

We evaluate ATOMICA on benchmark tasks for RNA and protein annotation. Here we describe how we adapt ATOMICA to each task and the baselines used for comparison.

4.1 Datasets

We evaluate ATOMICA on five benchmark tasks from RNAglib [64] and MaSIF [16] spanning RNA and protein structure annotation: RNA-protein binding site prediction, RNA Gene Ontology (GO) term prediction, RNA-small molecule binding site prediction, RNA-ligand pocket classification, and protein-ligand pocket classification.

RNAglib RNA-Protein. This task from RNAglib requires predicting whether each RNA residue is part of a protein binding site. The dataset contains 891 training structures with 52,175 residues of which 26.7% are part of a protein binding site, 191 validation structures with 11,063 residues of which 27.1% are part of a protein binding site, and 190 test structures with 10,851 residues or which 27.4% are part of a protein binding site. Positive labels are assigned to residues within 8 Å of any protein atom in a protein-RNA complex. This is a binary classification task at the residue level. Benchmark dataset splits are defined by clustering with USalign at a similarity threshold of 0.5.

RNAglib RNA-GO. This task involves predicting Gene Ontology (GO) terms associated with RNA molecules. The dataset includes 349 training samples, 75 validation samples, and 75 test samples. We predict membership in 5 GO term classes: GO:0000353 formation of quadruple SL/U4/U5/U6 snRNP (33.5%), GO:0010468 regulation of gene expression (20.0%), GO:0005682 U5 snRNP (16.2%), GO:0005688 U6 snRNP (15.6%), and GO:0005686 U2 snRNP (14.2%). This is a multilabel classification task where each RNA can be assigned to multiple GO terms. Of the 499 samples in this dataset, 161 samples are assigned no GO terms, 179 samples are assigned one GO term, and 159 samples are assigned two GO terms. Benchmark dataset splits follow a 60% sequence-identity split.

RNAglib RNA-Site. This task predicts whether each RNA residue is part of a small molecule binding site. The dataset contains 157 training structures with 10,092 residues of which 7.8% belong to a small molecule binding site, 34 validation structures with 2,162 residues of which 7.8% belong to a small molecule binding site, and 33 test structures with 2,158 residues of which 7.8% belong to a small molecule binding site. A residue is labeled positive if it lies within 8 Å of any atom of a bound ligand. This is a binary classification task at the residue level. Benchmark dataset splits are defined by clustering with USalign at a similarity threshold of 0.5.

RNAglib RNA-Ligand. This task predicts the ligand type for RNA binding pockets. The dataset includes 203 training pockets, 43 validation pockets, and 44 test pockets across 3 ligand classes:

Paromomycin (PAR, 22.3%), Gentamycin C1A (LLL, 67.2%), and Aminoglycoside TC007 (8UZ, 10.0%). Pockets are defined by expanding from residues within 8 Å of ligand atoms. This is a multiclass classification task at the pocket level. Benchmark dataset splits defined by clustering with USalign at a similarity threshold of 0.5.

MaSIF-ligand. This benchmark evaluates protein pocket classification across 7 common small molecule ligands: ADP (28.9%), CoA (12.6%), FAD (16.2%), heme (12.8%), NAD (11.4%), NAP (8.0%), and SAM (10.2%). The dataset contains 2,509 total pockets split into 1,839 training, 203 validation, and 467 test pockets. Binding pockets are defined as residues within 8Å of the ligand heavy atoms.

4.2 Fine-tuning ATOMICA for downstream tasks

We fine-tune ATOMICA for each downstream task using task-specific prediction heads while leveraging the pretrained atomic, block, and graph-level representations. The hyperparameters are selected based on the best validation metric, all models were trained on 1 NVIDIA H100 Tensor Core GPU for up to 12 hours.

RNA-Protein. For RNA-protein binding site prediction, we pass the d_{node}-residue-level embeddings to a residue-level binary classifier with a 3-layer MLP head. Models are trained with an initial learning rate of 5×10^{-5}, a final learning rate of 1×10^{-6}, weight decay of 1×10^{-3}, gradient clipping at 1.0, trained for 400 epochs, and AUPRC as the validation metric. We train 5 models with different random seeds, identical architecture, and hyperparameters, and ensemble the models by taking the mean over the predicted logits.

RNA-GO. For RNA gene ontology term prediction, we employ a graph-level multi-label classifier predicting 5 GO term classes. To address class imbalance, we use focal loss with $\gamma = 2.0$: $\text{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$, which down-weights easy examples and focuses learning on hard cases. We use the ATOMICA d_{node}-graph-level embeddings with a 4-layer MLP, with a constant learning rate of 4×10^{-5}, weight decay of 1×10^{-3}, gradient clipping at 1.0, trained for 200 epochs, and F1 macro as the validation metric. The model is also trained with a dataset augmentation of randomly masking 10% of residues for 80% of the training samples. We train 5 models with different random seeds, identical architecture, and hyperparameters, and ensemble the models by taking the mean over the predicted logits.

RNA-Site. For RNA small molecule binding site prediction, we pass the d_{node}-residue-level embeddings to a residue-level binary classifier with a 3-layer MLP head. Models are trained with a constant learning rate of 5×10^{-5}, weight decay of 1×10^{-3}, no gradient clipping, trained for 400 epochs, and AUPRC as the validation metric. We train 5 models with different random seeds, identical architecture, and hyperparameters, and ensemble the models by taking the mean over the predicted logits.

RNA-Ligand. For RNA-ligand pocket classification, we employ a graph-level multi-class classifier predicting one of the three ligand classes (PAR, LLL, 8UZ). With a weighted cross-entropy loss, the weights are given by the inverse of training label frequency. We use the ATOMICA d_{node}-

graph-level embeddings with a 4-layer MLP, with a constant learning rate of 1×10^{-5}, weight decay of 0.1, gradient clipping at 1.0, trained for 400 epochs, and F1 macro as the validation metric. The model is also trained with a dataset augmentation of randomly masking 10% of residues for 80% of the training samples. We train 5 models with different random seeds, identical architecture, and hyperparameters, and ensemble the models by taking the mean over the predicted logits.

MaSIF-ligand. For protein-ligand pocket classification, we use a pocket-level multiclass classifier to predict one of 7 ligand classes (ADP, COA, FAD, HEM, NAD, NAP, SAM). To address class imbalance, we use weighted cross-entropy loss with class weights inversely proportional to their training set frequencies. We use the ATOMICA d_{node}-graph-level embeddings with a 4-layer MLP, with a constant learning rate of 3×10^{-5}, weight decay of 1×10^{-3}, no gradient clipping, trained for 300 epochs, and F1 macro as the validation metric. We train 5 models with different random seeds, identical architecture, and hyperparameters, and ensemble the models by taking the mean over the predicted logits.

4.3 Benchmarking models for RNAglib tasks

We compare ATOMICA against structure and sequence-based baselines on the RNAglib benchmark tasks. For baselines that provide frozen pretrained embeddings, we train a task-specific four-layer MLP classifier on top of the embeddings and perform a hyperparameter sweep over learning rate $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}, 10^{-5}\}$, dropout $\{0.0, 0.1, 0.2, 0.3, 0.4\}$, and MLP hidden dimension $\{64, 128, 256, 512, 1024\}$. For consistency with ATOMICA, we sweep between binary cross-entropy loss and focal loss for RNA-Go. For RNA-Ligand, we sweep between cross-entropy loss and weighted cross-entropy loss, where class weights are set to the inverse of the training-label frequencies. For all models, we report the performance when ensembling the predicted logits of models trained across 5 random seeds.

RNAglib. The RNAglib baseline is a graph neural network trained end-to-end on RNA 2.5D structure graphs, with nucleotides as nodes and edges given by backbone connectivity and base-pair interactions. We use the hyperparameters for the RGCN model specified by RNAglib [64].

RiNALMo, RNA-FM, and RNAErnie. For the RNA language models. We encode the RNA chain and extract residue-level embeddings from the final layer of the pretrained models. We keep the embeddings of residues in the RNA graph. For the RNA-GO and RNA-Ligand task, which are tasks at the RNA subsequence and pocket-level, respectively, we use the mean embedding over residues that are present in the substructure. For residue-level tasks, RNA-Protein and RNA-Site, we use the residue-level embeddings. We then fine-tune these embeddings with a MLP for each task as specified above.

gRNAde. Geometric RNA Design (gRNAde) is a structure-based model for sequence design from an RNA backbone structure using graph vector-perceptron (GVP) layers [105] on 3D RNA graphs. We pass the RNA-structure graph into the pretrained gRNAde backbone encoder and extract the residue-level embeddings. For RNA-GO and RNA-Ligand we take the average over all residue-level embeddings. Finally, we fine-tune these embeddings with a MLP for each task as specified

above.

4.4 Benchmarking models for MaSIF-ligand prediction tasks

We compare ATOMICA to geometry-based, protein language model (PLM), and contrastive-learning baselines on the MaSIF-ligand benchmark. For PLM baselines, we extract pocket embeddings by mean pooling the residue representations for pocket residues. We then train a four-layer MLP classifier on top of these embeddings using cross-entropy loss, and sweep hyperparameters over learning rate $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}, 10^{-5}\}$, dropout $\{0.0, 0.1, 0.2, 0.3, 0.4\}$, and MLP hidden dimension $\{64, 128, 256, 512, 1024\}$. We report the performance when ensembling the predicted logits of models trained across 5 random seeds.

MaSIF. MaSIF (Molecular Surface Interaction Fingerprinting) is a geometry-based approach that characterizes protein surfaces using shape complementarity, electrostatics, hydrophobicity, and chemical features. We use the published MaSIF-ligand implementation.

DPocket. DPocket extracts hand-crafted geometric and physicochemical pocket features, including volume, surface area, depth, hydrophobicity, electrostatic potential, and residue identity. We z-score normalize features using training-set statistics and use these features for the MLP.

ESM-2, ESM-C. We use ESM-2 (3B) and ESM-C (300M) to encode full protein chains and extract pocket residue embeddings from the final layer. We obtain pocket-level embeddings by mean pooling over pocket residues and pass them to the MLP.

ESM-3. We use ESM-3 (1.4B) to encode protein chains using the sequence and structure tracks and extract pocket residue embeddings from the final layer. We obtain pocket-level embeddings by mean pooling over pocket residues and pass them to the MLP.

SaProt. SaProt is a PLM that incorporates structural information via structure-aware tokenization. We extract residue embeddings from the pretrained SaProt model using Foldseek 3Di tokens for protein chains. We obtain pocket-level embeddings by mean pooling over pocket residues and pass them to the MLP.

ProstT5. ProstT5 is a T5-based PLM trained on both sequence and structure. We encode protein chains with sequence tokens and with Foldseek 3Di structure tokens. We mean pool over pocket residues in each track, concatenate the resulting pocket embeddings, and pass the concatenated embedding to the MLP.

DrugCLIP. DrugCLIP is a contrastive framework that learns aligned protein and small-molecule representations. We extract pocket embeddings from the pretrained DrugCLIP protein encoder and feed them to the MLP. We also evaluated the original retrieval setup (encoding ligand conformers and selecting by highest cosine similarity to pocket embeddings), but this underperformed MLP-based fine-tuning.

DeeplyTough. DeeplyTough learns pocket descriptors via metric learning on pocket similarity

tasks. We extract pretrained pocket descriptors from DeeplyTough and classify them with an MLP.

PocketAnchor. PocketAnchor learns pocket representations via contrastive learning by anchoring pockets to their ligands. We extract pretrained pocket embeddings from PocketAnchor and feed them to an MLP.

5 Cross-modality comparison for orthosteric PPI inhibitors

5.1 Dataset

We utilized the 2P2IDB database, which contains experimentally determined structures of protein-protein interactions and their inhibitors from the Protein Data Bank (PDB). For PPI structures, they were processed to keep only the target (A) and partner (B) chain. For protein–inhibitor structures, they were processed to keep only the target (A) and inhibitor (I) ligand.

5.2 Protein-peptide inhibitor analyses

For protein–peptide complexes where peptide B contains ≤ 30 residues, we performed block-level comparisons between inhibitor-bound structures and peptide B to assess whether ATOMICA embeddings capture structural similarities at the interface.

5.2.1 Structure alignment

We aligned inhibitor target chain structures to PPI target chain structures. Initial sequence alignment using BLOSUM62 substitution matrix with gap opening penalty of -10 and gap extension penalty of -1. Superposition of matched Cα atom coordinates using Kabsch algorithm. Iterative outlier rejection was then applied and residue pairs with RMSD > 2.0 Å were removed. The alignment was refined over 5 cycles until convergence.

5.2.2 ATOMICA embedding computation

We computed ATOMICA embeddings for both inhibitors and peptide B using the pretrained ATOMICA model. For the inhibitor embedding, we embedded the inhibitor bound to the target protein pocket and extracted the block embeddings of the inhibitor, $\mathbf{h}^{\text{block}}_{\text{inhibitor}}$. For the protein–peptide embedding, we embedded the peptide bound to the target protein pocket and extracted the block embeddings of the peptide, $\mathbf{h}^{\text{block}}_{\text{peptide}}$.

5.2.3 Block pairwise comparison

For each inhibitor-PPI pair, we performed the following block-level comparison. Applied the rotation matrix R and translation vector t from the alignment step to transform inhibitor block coordinates into the PPI reference frame. Block coordinates, $\mathbf{x}$, are the average of the atom coordinates within each block. We filtered blocks to remove singleton blocks (size = 1) and global-type blocks, retaining only blocks with >1 atoms. Pairwise embedding distances, $\text{dist}_{\text{ATOMICA}}(\mathbf{h}^{\text{block}}_{\text{peptide}}, \mathbf{h}^{\text{block}}_{\text{inhibitor}})$, between all inhibitor blocks and peptide B blocks were calculated as the cosine distance between block embedding vectors. Pairwise spatial distance was given by the Euclidean distance between block center coordinates which we define as $\text{dist}_{\mathbb{R}^3}(\mathbf{x}^{\text{block}}_{\text{peptide}}, \mathbf{x}^{\text{block}}_{\text{inhibitor}})$ (Å). We filter the pairs of inhibitor–PPI pairs to only keep those with at least 10 pairwise blocks for comparison, this results in 18 out of 31 protein-peptide complexes being suitable leaving 268 out of 1848 matched inhibitor structures. The final number of inhibitor structures matched to each PPI complex is available in Table S1.

5.2.4 Statistical analyses

Spearman's rank correlation. For each protein-peptide complex and their matched protein-inhibitor complexes, we computed the Spearman's rank correlation coefficient between $\text{dist}_{\text{ATOMICA}}$ and the spatial distance $\text{dist}_{\mathbb{R}^3}$ for all inhibitor–peptide blocks for all inhibitors that match the protein-peptide complex. We tested for statistical significance using a two-tailed test ($p < 0.05$). Multiple testing correction was performed using the Benjamini-Hochberg false discovery rate (FDR) procedure with $q < 0.05$.

Fold Change analysis. We evaluated whether embedding-based retrieval could identify spatially proximal blocks. For each inhibitor block, we ranked peptide B blocks by embedding distance, $\text{dist}_{\text{ATOMICA}}$, and selected the top-$k$ = 10 lowest distances as retrieved inhibitor-peptide block pairs. Inhibitor-peptide block pairs with $\text{dist}_{\mathbb{R}^3}$ within a geometric threshold of 4.0 Å were considered spatially close. We computed *Precision@10* as the fraction of top-10 retrieved blocks within 4.0 Å and *Fold Change@10*, which is Precision@10 divided by the baseline rate (overall fraction of blocks within the threshold $\text{dist}_{\mathbb{R}^3} < 4$ Å).

5.3 Protein-protein inhibitor analyses

For protein–protein complexes where protein B has >30 residues, we employed a surface sampling approach to identify interface regions and assess whether ATOMICA embeddings can distinguish interface patches that are geometrically similar to inhibitor binding sites. We keep PPI complexes that have an A–B binary binding structure, as a result we discard INTEGRASE/LEDGF and TNFA trimers from our analyses. A total of 187 inhibitor structures across six PPIs are evaluated. The final number of inhibitor structures per PPI complex is available in Table S2.

5.3.1 Surface point sampling

We generated molecular surface representations and sampled points uniformly across the surface of each protein B. We use the MSMS tool [106] to compute solvent-accessible surfaces with the parameters: vertex density = 3.0 vertices per Å^2, probe radius = 1.5 Å. For each protein B chain, we extracted atomic coordinates and generated a triangular mesh with vertices, faces, and surface normals. We then sampled 1,000 points per protein B surface using area-weighted triangle sampling. For each sampled triangle, we uniformly sampled a point using barycentric coordinates. This was achieved with generated random values $r_1, r_2 \sim U(0, 1)$, computed barycentric coordinates: $u = 1 - \sqrt{r_1}$, $v = \sqrt{r_1}(1 - r_2)$, $w = \sqrt{r_1} r_2$, and point position: $\mathbf{p} = u\mathbf{v}_0 + v\mathbf{v}_1 + w\mathbf{v}_2$ where $\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2$ are triangle vertices.

5.3.2 Interface patch definition

For each sampled surface point, we defined a local interface patch based on spatial proximity to protein blocks. We computed the Euclidean distance from the surface point to all protein B block centers. Selected blocks were within a radius of 16.0 Å from the surface point, and we discarded points with <8 nearby blocks. For each interface patch center on protein B, we computed the distance to the nearest Cα atom on protein A.

5.3.3 ATOMICA embedding computation

Inhibitor embeddings. For each protein–inhibitor complex (A–I), we apply ATOMICA to the inhibitor bound to the target protein pocket to obtain contextualized inhibitor interface embeddings, $\mathbf{h}^{\text{interface}}_{\text{inhibitor}}$.

Interface patch embeddings. For each protein–protein (A–B) complex, we sample local surface interface patches on the protein B. Each patch is defined as the set of blocks within a 16 Å radius of a surface point on B. We apply ATOMICA to each sampled patch to obtain an interface embedding, $\mathbf{h}^{\text{interface}}_{\text{protein B}}$, that captures the local structural context. These patch embeddings serve as retrieval candidates.

5.3.4 Retrieval analysis

We formulate retrieval as follows: given an inhibitor embedding (query), we rank sampled interface patches on protein B (candidates) by embedding similarity and evaluate whether high-similarity patches localize to the native A–B binding site.

To limit information leakage from the target interface, embeddings from protein A are not used in forming retrieval candidates or computing similarity scores. Specifically, (i) the inhibitor embedding is constructed solely from ligand blocks at the interface, and (ii) candidate embeddings are computed exclusively from patches on protein B. Thus, retrieval compares a ligand-only query

to protein-only interface patches, avoiding trivial localization through direct encoding of the target interface geometry.

We compute the following distance metrics: embedding distance $\text{dist}_{\text{ATOMICA}}(\mathbf{h}^{\text{interface}}_{\text{inhibitor}}, \mathbf{h}^{\text{interface}}_{\text{protein B}})$, given by the cosine distance between inhibitor embedding and interface patch embedding, and spatial distance $\text{dist}_{\mathbb{R}^3}(\mathbf{x}^{\text{interface}}_{\text{inhibitor}}, \mathbf{x}^{\text{interface}}_{\text{protein B}})$, given by the Euclidean distance from the interface patch center to the nearest Cα atom on protein A.

Retrieval procedure. For each inhibitor, we ranked all interface patches by $\text{dist}_{\text{ATOMICA}}$ (ascending). The top-k = 10 (lowest $\text{dist}_{\text{ATOMICA}}$) retrieved patches were selected. A geometric threshold of patches within 12.0 Å of the interface protein A was considered "close". A threshold of 12.0 Å was selected as it represents the 25th percentile of distances of sampled patches from the PPI interface. We computed *Precision@10*: the fraction of top-10 patches within the geometric threshold and *Enrichment@10*: Precision@10 divided by the baseline rate. The baseline rate is given by the proportion of all patches on protein B that are within 12.0 Å from the A–B interface.

5.3.5 Statistical analyses

Spearman Correlation. For each PPI family, we aggregated all inhibitor-patch pairs and computed the Spearman rank correlation coefficient between embedding distances $\text{dist}_{\text{ATOMICA}}$ and geometric distances $\text{dist}_{\mathbb{R}^3}$. Statistical significance was assessed using a two-tailed test ($p < 0.05$), with FDR correction (Benjamini-Hochberg, $q < 0.05$) applied across families.

6 Analysis of dark proteome with ATOMICA-Ligand

Overview. We demonstrate versatility in ATOMICA and finetune the model for annotating ions and ligands to binding sites. The finetuned version of the model is applied to putative binding sites in the dark proteome.

Training ATOMICA-Ligand. The objective is to predict the probability of a specific ion or ligand binding to a given protein interface pocket. We frame this as a binary prediction task and finetune a separate model for each ion and small molecule. A predictive head is a L_{ligand}-layer MLP. For each ion and small molecule, we use RayTune with Optuna and ASHA to finetune ATOMICA-Ligand from ATOMICA and find the optimal hyperparameters among $L_{\text{ligand}} \in [3, 4, 5]$, learning rate $\in [10^{-6}, 10^{-3}]$, non-linearity $\in$ [`relu, gelu, elu`], hidden dimension of MLP $\in [16, 32, 64]$, gradient clipping $\in$ [`None`, 1], and the number of nearest neighbors to define edges to in the graph $k \in [4, 6, 8]$. To address class imbalances in our dataset, we apply a weighted sampling strategy during training, where each protein pocket receives a sampling weight inversely proportional to the total count of its label class. For each ion and small molecule, we finetune ATOMICA-Ligand for 50 epochs on 1 NVIDIA H100 Tensor Core GPU. Three replicate models are trained for each ion and small molecule. For binary classification of binding sites, we set thresholds that maximize the F1 score, constraining these values to fall within the range of 0.05 to 0.95.

We use a one-vs-rest binary formulation for ATOMICA-Ligand because it yields ligand-specific decision boundaries.

Dataset curation. Given an ion or small molecule, we separate all graphs in the pretraining set containing this ion bound to a protein. We cluster protein binders with a 30% protein sequence similarity cutoff, coverage of 80%, sensitivity of 8, and cluster mode 1 using MMseqs2 [90]. The clusters are then divided into training, validation, and test sets in an 8:1:1 ratio. We set up this split for the following metal ions: Ca, Co, Cu, Fe, K, Mg, Mn, Na, Zn, and the following small molecules with these PDB chemical codes: ADP (adenosine diphosphate), ATP (adenosine triphosphate), CIT (citric acid), CLA (chlorophyll A), FAD (flavin adenine dinucleotide), GDP (guanosine diphosphate), GTP (guanosine triphosphate), HEC (heme C), HEM (heme B), NAD (nicotinamide adenine dinucleotide), NAP (NADP+, nicotinamide adenine dinucleotide phosphate, oxidized form), NDP (NADPH, nicotinamide adenine dinucleotide phosphate, reduced form).

Dark proteome annotation. The dark proteome is comprised of proteins that are dissimilar in sequence and structure from all currently annotated proteins. We use the clusters of the dark proteome from FoldSeek cluster on the AlphaFold Protein Structure Database [45]. We limit our analysis to the 33,482 clusters with an average pLDDT > 90. For each cluster, we take the representative protein and run PeSTo [84] on the protein structure to predict ion and small molecule binding sites. We keep residues with PeSTo binding score > 0.8 as the putative binding site, with a minimum of 5 residues required. In total, we extract 2,851 ion binding proteins and 969 small molecule binding proteins from the 33,482 representative proteins. Given these binding interfaces, we run ATOMICA-Ligand for all finetuned ion and small molecules to annotate chemical identities to the binding sites. We evaluated the quality of ATOMICA-Ligand predicted protein-ligand complexes by folding them with AlphaFold3 and evaluating their ipTM scores. For comparison, we established a reference baseline using randomly sampled proteins from the dark proteome with predicted ion and small molecule binding capabilities. These reference proteins were selected and paired with ligands to match both the number and identity of annotated ligands in our predicted complexes. For sequence-based annotation we run the Google Colab notebook with ProtNLM [88].

7 Additional quality metrics and statistical analyses

Visualization of embeddings. We embed 2,105,459 molecular interaction complexes from training, validation, and testing datasets, and project these embeddings into two dimensions using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [107] with 30 neighbors and a minimum distance of 0.001. All interaction complexes were visualized with PyMOL [108].

Visualization of atom and block identity. From molecular interaction complexes in training, validation, and testing datasets, we take the element-wise average atom and block embeddings for each identity. For the atom and block embeddings, we projected the embeddings into two dimensions using Principal Component Analysis (PCA).

Capturing important residues with ATOMICASCORE. To identify amino acids involved in specific non-covalent contacts, we analyzed protein-small molecule complexes in the test set using PLIP [61]. We then quantified the contribution of each amino acid to the interaction by calculating an importance score, ATOMICASCORE. For amino acid i, ATOMICASCORE is defined as $a_i = \text{sim}\left(\mathbf{h}_G^{\text{graph}}, \mathbf{h}_{G\setminus i}^{\text{graph}}\right)$, where $\mathbf{h}_G^{\text{graph}}$ is the embedding of the original complex, and $\mathbf{h}_{G\setminus i}^{\text{graph}}$ represents the embedding of the modified complex in which amino acid i has been replaced with a special mask token and its constituent atoms substituted with a single special mask atom. Here, sim denotes the cosine similarity between the two embedding vectors. In total, we analyze 5,691 protein-small molecule complexes with at least 20 amino acid blocks at the interface. Within the 10 amino acids with the lowest ATOMICASCOREs we count the number of amino acids involved in intermolecular bonds. For reference, we compare this to randomly nominating 10 amino acids at the interface, and to ESM-2 (3B) [5] and score amino acids by the log likelihood of the masked mutant compared to the original amino acid [62].

Training ATOMICA on a single pair of interacting modalities (modality-specific training). To demonstrate representations learned by ATOMICA are generalizable across multiple modalities, we train models with identical architecture and hyperparameters on only single pairs of interacting modalities (small molecules, protein-ion, protein-small molecule, protein-DNA, protein-RNA, protein-peptide, protein-protein, nucleic acid-small molecule). Using the same training setup as ATOMICA, these models are trained on the same training data as ATOMICA but filtered for only one pair of interacting modalities. The models are trained for 150 epochs on 4 NVIDIA H100 Tensor Core GPUs in parallel. The model checkpoint with the lowest validation loss is then used for further finetuning on masked block identity prediction on the same training data for 50 epochs with a learning rate of 1e-4. We also finetune ATOMICA for 50 epochs on block identity prediction for each pair of interacting modalities. To compare the quality of embeddings generated by ATOMICA and versions of it trained on single modalities, we evaluate the accuracy of masked block identity prediction on a test set. This test set was not seen by any of the models and has 30% sequence similarity and minimal small molecule fingerprint similarity to any training and validation data.

8 Experimental validation of heme binders

8.1 Selection of heme binders for experimental validation

Of the 59 proteins predicted to bind to heme non-covalently (HEM ligand) and eight proteins predicted to bind to heme covalently (HEC ligand) from the dark proteome, nine were selected for experimental validation. In addition to their ATOMICA-Ligand score, they were selected based on surpassing a minimum AlphaFold3 ipTM score of 0.85 and the absence of a large transmembrane domain, which would make protein synthesis difficult. The UniProt identifiers of these proteins are: A0A7W1B5T5, A0A1T4N4K0, A0A4P5TA35, A0A7W0X6V6, A0A2V6P8N7, A0A136KY61, A0A7Y8LED7, A0A7V7N0X5, and V5BF69.

8.2 Synthesis of proteins

A0A7W1B5T5 and A0A1T4N4K0 were synthesized with automated flow peptide synthesis. In addition, all nine proteins were synthesized and purified as recombinant proteins by GenScript (Piscataway, NJ, USA).

Reagents and supplies. Disposable fritted polypropylene reaction vessels were purchased from Torviq (cat. no. SF-0500-LL for 6 mL reactors). H-RAM-ChemMatrix resin (0.17 mmol/g, batch no. 18007556) was purchased from ChemMatrix. Fmoc-amino acids were purchased from the Novabiochem product line of Millipore Sigma and were used as received. OmniSolv® grade *N,N*-Dimethylformamide (DMF, biosynthesis grade) was purchased from Millipore Sigma (product DX1732-1) and was equipped using AldraAmine trapping agents (for 1000–4000 mL DMF, catalog number Z511706). *N,N*-Diisopropylethylamine (DIPEA; ReagentPlus $\geq$ 99%), piperidine (ACS reagent, $\geq$ 99.0%), trifluoroacetic acid (HPLC grade, $\geq$ 99.0%), triisopropylsilane (TIPS, $\geq$ 98.0%), acetonitrile (ACN, HPLC grade), 1,2-ethanedithiol (EDT, 98+%), trimethylsilyl chloride (TMS-Cl, 98.0% GC), triphenylphosphine (PPh3, ReagentPlus 99%) and formic acid (FA, $\geq$ 95.0%) were all purchased from Sigma-Aldrich and used as is. O-(7-azabenzotriazol-1-yl)-*N,N,N',N'*-tetramethyluronium hexafluorophosphate (HATU, $\geq$ 97.0%) was purchased from P3 Biosystems. Dichloromethane (DCM, HPLC grade $>$ 99.9%) was purchased from Fisher. Hemin (BioXtra, $\geq$ 96.0%, cat. no. 51280-1G) was purchased from Sigma-Aldrich. Water was deionized in-house using a Milli-Q water purification system from Millipore.

Automated flow peptide synthesis. 125 mg of H-RAM ChemMatrix® resin was weighed in a 6 mL Luer Torviq reactor and swollen using DCM (15 min) and DMF (15 min). The reactor was drained and deposited in the heating reactor of the "Amidator"; an automated-flow peptide synthesiser built in the Pentelute Lab. Sequences for A0A7W1B5T5 and A0A1T4N4K0 were directly taken from UniProt. Synthesis was carried out as previously described [109, 110]. After the synthesis, the resin was drained, washed with DCM $\times$3 and let to dry on the vacuum manifold.

Cleavage protocol. Half of the peptide-anchored resin was transferred to a 10 mL Luer-Lock Torviq syringe equipped with a frit and was cleaved for 4 hours at room temperature using a 5 mL of a Met-Reducing cleavage cocktail containing 77.5% TFA, 5% thioanisole, 5% dimethylsulfide, 5% TIPS, 5% TMS-Cl and 2.5% EDT. After 4 hours, the cleavage mixture was filtered, evaporated under N2 flow, and the resulting solid was dissolved in 20 mL of 50% ACN in water + 0.1% formic acid, flash-frozen using liquid nitrogen, and lyophilized.

Preparative liquid chromatography-mass spectrometry (Prep LC-MS). Preparative HPLC purification was performed on an Agilent Technologies mass-directed purification system 1260 Infinity LC coupled to a 6130 Single Quad MS. A Timberline Instrument TL105 HPLC column heater was used to heat the column to 50 °C. The crude lyophilized peptide powder was solubilized in 6 mL of a denaturing buffer consisting of 6 M Gdm HCl, 100 mM TRIS HCl, pH=7.5. The mixture was filtered using a 0.22 μm nylon syringe filter. An analytical HPLC run was first performed using a low volume of protein to estimate %B and a focused gradient was performed for the purification runs as follows:

Analytical run: <u>Column</u>: Agilent Zorbax 300 Stable Bond C3 PrepHT, 21.1 x 100 mm, 5 μm.

Flow Rate: 20 mL/min. Solvents: A: H_2O + 0.1% TFA, B: ACN + 0.1% TFA. Column Temperature: 50 °C. Gradient: 0-10 min 5% B, 10-40 min 5-95% B, 40-45 min 95% B.

Focused gradient run: To generate the final preparative HPLC method, we estimated the %B of elution during the analytical run and performed a preparative run using a linear gradient of -10% to +5%B in 60 minutes around the %B identified in the analytical run. Column: Agilent Zorbax 300 Stable Bond C3 PrepHT, 21.1 x 100 mm, 5 μm. Flow Rate: 20 mL/min. Solvents: A: H_2O + 0.1% TFA, B: ACN + 0.1% TFA. Column Temperature: 50 °C. Gradient: see description.

Fractions were analyzed by LC-MS, chosen based on apparent purity, flash-frozen using liquid nitrogen and lyophilized.

Protein folding via dialysis. The purified lyophilized fractions were solubilized with a denaturing solution containing 6 M guanidinium chloride, HEPES 10 mM, pH = 7.4 and grouped, using a total volume of less than 6 mL. Solubilized protein was dialysed against a 20 mM TRIS, 200 mM NaCl, pH 7.5 buffer using a 3.5 kDa MWCO dialysis membrane (Slide-A-Lyzer® MINI Dialysis Devices, 2 mL, cat. no. 88403) from Thermo. Buffer changes were performed after 2 and 24 hours for a total of three dialysis cycles over 48 hours.

Liquid chromatography-mass spectrometry (LC-MS). Mass analysis was performed on an Agilent Technologies 1290 Infinity II UHPLC coupled to a 6550 iFunnel Q-TOF LC-MS system. MS spectra were acquired in positive ionization mode with m/z range of 300-2000 Da.:

C4-1-91-10min: Phenomenex Aeris C4 column (2.1 x 150 mm, 3.6 μm, 0.2 mL/min, 40 °C), 1% B for 0–1 min, 1–91% B over 1–7 min, 1% B for 7–9 min. (ESI+, mass range 100–1700 m/z). Column: Phenomenex Aeris WIDEPORE C4 200 Å (2.1 x 150 mm, 3.6 μm, Cat. no. 00F-4486-AN.). Flow Rate: 0.3 mL/min. Solvents: A: H_2O + 0.1% FA, B: ACN + 0.1% FA. Column Temperature: 40 °C. Gradient: 0-2 min 1% B, 2-8 min 1%-91% B, 8-10 min 91%-95% B (MS acquisition from 2 to 8 min).

Data were processed using Agilent MassHunter BioConfirm software version 10.0. Mass deconvolution was carried out using the 600-2000 m/z mass range and output deconvoluted masses were fixed between 10,000-20,000 Da.

Protein concentration. After three cycles of dialysis (48 hours total), the protein solution was concentrated using pre-humidified 3 kDa MWCO centrifugal filters (Amicon® Ultra – 4). Samples were deposited onto the filters and spun at 4,000 RPM for 99 minutes, or until the combined remaining solution volume was under 500 µL. The resulting solution was taken up into protein lo-bind Eppendorf tubes and assayed for protein concentration (A280) using a Tecan Spark® plate reader. Extinction coefficients were estimated using the ProtParam tool from ExPASy Swiss Institute of Bioinformatics - Bioinformatics Resource Portal using the primary sequence of the protein candidates and specifying C-terminal amide and reduced cysteines to estimate the concentration of the proteins [111].

Analytical high-performance liquid chromatography (HPLC). Analytical HPLC was carried out on an Agilent 1290 series system with UV detection at 214 nm.

C3-5-65-60min: Zorbax 300-SB C3 column (2.1 $\times$ 150 mm, 3.6 μm, 0.2 mL/min, 40 °C),

5% B for 0–5 min, 5–65% B over 5-65 min, 65-95% B for 65-66 min, 95% B for 66-70 min, 95-5% B for 70-75 min. Column: Zorbax 300-StableBond C3 (2.1 × 150 mm, 5 μm, cat. no. 883750-909.) Flow Rate: 1 mL/min. Solvents: A: H_2O + 0.1% FA, B: ACN + 0.1% FA. Column Temperature: 40 °C. Gradient: 0-5 min 5% B, 5-65 min 5%-65% B, 65-66 min 65%-95% B, 66-70 min 95% B, 70-75 min 95-5% B.

UV 214 nm integrations were performed automatically by the built-in Agilent OpenLab CDS, ChemStation Edition Rev.C.01.10[287].

Holoprotein reconstitution. To reconstitute the apo protein, a stock solution of hemin in DMF (concentration variable, 33.3x the concentration of the protein in solution) was added to the protein in buffer in equimolar amounts (1 eq.) with a final concentration of 3% DMF. The solution was vortexed and incubated at room temperature for 5 minutes to allow binding as previously described[112].

UV-Vis spectroscopy heme binding assay. Nanodrop methodology (Synthetic candidates): 2 μL of the reconstituted holoprotein solution was deposited into a nanodrop plate (NanoQuant plate) and assayed for a Soret peak at ∼450 nm using a Tecan Spark® plate reader at wavelengths between 300 and 650 nm (1 nm step, 3.5 nm slit). All samples were subtracted from a blank containing 3% DMF in buffer, and a negative control using Hemin alone and Hemin + bovine serum albumin (BSA) was performed. Presence of a Soret peak (Soret λ_{max}) around 420-430 nm indicated positive heme binding [112, 113]. 384-well plate methodology (Recombinant candidates): Since the concentration of most recombinant candidates was lower than the detection threshold for nanodrop quantification, 120 µL of the reconstituted proteins were added to a 384-well plate from Greiner Bio-One (MICROPLATE, 384 WELL, PS, μCLEAR®, WHITE, NON-BINDING, cat. no. 781903). Absorbances were assayed for a Soret peak at ∼450 nm using a Tecan Tecan Spark® plate reader at wavelengths between 300 and 650 nm (1 nm step, 3.5 nm slit). All samples were subtracted from a blank containing 3% DMF in buffer, and a negative control using Hemin alone and Hemin + BSA was performed. Presence of a Soret peak (Soret λ_{max}) in the range of 420-430 nm indicated positive heme binding [112, 113].

References

1. Google DeepMind AlphaFold Team & Isomorphic Labs Team. Performance and structural coverage of the latest, in-development alphafold model (2023). Accessed: 2023-04-16.

2. Krishna, R. *et al.* Generalized biomolecular modeling and design with rosettafold all-atom. *Science* eadl2528 (2024).

3. Leclercq, M. & Droit, A. Protein language models: Applications and perspectives. *Journal of Proteome Research* (2025).

4. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

5. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

6. Su, J. *et al.* Saprot: Protein language modeling with structure-aware vocabulary. *bioRxiv* 2023–10 (2023).

7. Heinzinger, M. *et al.* Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics* **6**, lqae150 (2024).

8. Shu, L., Tang, J., Guan, X. & Zhang, D. A comprehensive survey of genome language models in bioinformatics. *Briefings in Bioinformatics* **27**, bbaf724 (2026).

9. Penić, R. J., Vlašić, T., Huber, R. G., Wan, Y. & Šikić, M. Rinalmo: General-purpose rna language models can generalize well on structure prediction tasks. *Nature Communications* **16**, 5671 (2025).

10. Chen, J. *et al.* Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300* (2022).

11. Dalla-Torre, H. *et al.* Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* **22**, 287–297 (2025).

12. Duran-Frigola, M. *et al.* Extending the small-molecule similarity principle to all levels of biology with the chemical checker. *Nature Biotechnology* **38**, 1087–1096 (2020).

13. Ross, J. *et al.* Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* **4**, 1256–1264 (2022).

14. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **4**, 279–287 (2022).

15. Tsuchiya, Y., Yamamori, Y. & Tomii, K. Protein–protein interaction prediction methods: from docking-based to ai-based approaches. *Biophysical Reviews* **14**, 1341–1348 (2022).

16. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **17**, 184–192 (2020).

17. Sverrisson, F., Feydy, J., Correia, B. E. & Bronstein, M. M. Fast end-to-end learning on protein surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15272–15281 (2021).

18. Gainza, P. *et al.* De novo design of protein interactions with learned surface fingerprints. *Nature* 1–9 (2023).

19. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using alphafold2. *Nature communications* **13**, 1265 (2022).

20. Das, S. & Chakrabarti, S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Scientific reports* **11**, 1761 (2021).

21. Renaud, N. *et al.* Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. *Nature communications* **12**, 7068 (2021).

22. Meli, R., Morris, G. M. & Biggin, P. C. Scoring functions for protein-ligand binding affinity prediction using structure-based deep learning: a review. *Frontiers in bioinformatics* **2**, 885983 (2022).

23. Jia, Y. *et al.* Deep contrastive learning enables genome-wide virtual screening. *Science* **391**, eads9530 (2026).

24. Zhou, G. *et al.* Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations* (2023).

25. Gao, B. *et al.* Self-supervised pocket pretraining via protein fragment-surroundings alignment. In *The Twelfth International Conference on Learning Representations* (2024).

26. Wei, J., Chen, S., Zong, L., Gao, X. & Li, Y. Protein–rna interaction prediction with deep learning: structure matters. *Briefings in bioinformatics* **23**, bbab540 (2022).

27. Lam, J. H. *et al.* A deep learning framework to predict binding preference of rna constituents on protein surface. *Nature communications* **10**, 4941 (2019).

28. Xia, Y., Xia, C.-Q., Pan, X. & Shen, H.-B. Graphbind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research* **49**, e51–e51 (2021).

29. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology* **33**, 831–838 (2015).

30. Sun, L. *et al.* Predicting dynamic cellular protein–rna interactions by deep learning using in vivo rna structures. *Cell research* **31**, 495–516 (2021).

31. Rube, H. T. *et al.* Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nature Biotechnology* **40**, 1520–1527 (2022).

32. Tiyoula, F. N. *et al.* Artificial intelligence in metalloprotein binding site prediction: A systematic review bridging bioinformatics and biotechnology. *International Journal of Biological Macromolecules* 146666 (2025).

33. Dürr, S. L., Levy, A. & Rothlisberger, U. Metal3d: a general deep learning framework for accurate metal ion location prediction in proteins. *Nature Communications* **14**, 2713 (2023).

34. Mohamadi, A. *et al.* An ensemble 3d deep-learning model to predict protein metal-binding site. *Cell Reports Physical Science* **3** (2022).

35. Li, J. *et al.* Artificial intelligence for rna–ligand interaction prediction: advances and prospects. *Drug Discovery Today* 104366 (2025).

36. Sun, S. & Gao, L. Contrastive pre-training and 3d convolution neural network for rna and small molecule binding affinity prediction. *Bioinformatics* **40**, btae155 (2024).

37. Huang, Z. *et al.* Deeprsma: a cross-fusion-based deep learning method for rna–small molecule binding affinity prediction. *Bioinformatics* **40**, btae678 (2024).

38. Jena, S. *et al.* Noncovalent interactions in proteins and nucleic acids: beyond hydrogen bonding and π-stacking. *Chemical Society Reviews* **51**, 4261–4286 (2022).

39. Rodrigues, C. H., Pires, D. E., Blundell, T. L. & Ascher, D. B. Structural landscapes of ppi interfaces. *Briefings in Bioinformatics* **23**, bbac165 (2022).

40. Khazanov, N. A. & Carlson, H. A. Exploring the composition of protein-ligand binding sites on a large scale. *PLoS computational biology* **9**, e1003321 (2013).

41. Krüger, D. M., Neubacher, S. & Grossmann, T. N. Protein–rna interactions: structural characteristics and hotspot amino acids. *Rna* **24**, 1457–1465 (2018).

42. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **72**, 171–179 (2016).

43. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).

44. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences* **112**, 15898–15903 (2015).

45. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).

46. Kulkarni, P. & Uversky, V. N. Intrinsically disordered proteins: the dark horse of the dark proteome. *Proteomics* **18**, 1800061 (2018).

47. Wei, H., Wang, W., Peng, Z. & Yang, J. Q-biolip: A comprehensive resource for quaternary structure-based protein–ligand interactions. *bioRxiv* 2023–06 (2023).

48. Yang, J., Roy, A. & Zhang, Y. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research* **41**, D1096–D1103 (2012).

49. Wollschläger, T., Kemper, N., Hetzel, L., Sommer, J. & Günnemann, S. Expressivity and generalization: Fragment-biases for molecular gnns. *arXiv preprint arXiv:2406.08210* (2024).

50. Hermosilla, P. *et al.* Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *International Conference on Learning Representations* (2021).

51. Wang, L., Liu, H., Liu, Y., Kurtin, J. & Ji, S. Learning protein representations via complete 3d graph networks. *arXiv preprint arXiv:2207.12600* (2022).

52. Kong, X., Huang, W. & Liu, Y. Generalist equivariant transformer towards 3d molecular interaction learning. *arXiv preprint arXiv:2306.01474* (2023).

53. Batzner, S. *et al.* E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications* **13**, 2453 (2022).

54. Musaelian, A. *et al.* Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications* **14**, 579 (2023).

55. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking (2023).

56. Townshend, R. J. *et al.* Geometric deep learning of rna structure. *Science* **373**, 1047–1051 (2021).

57. Luo, S. *et al.* One transformer can understand both 2d & 3d molecular data. In *The Eleventh International Conference on Learning Representations* (2022).

58. Zaidi, S. *et al.* Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations* (2023).

59. Godwin, J. *et al.* Simple GNN regularisation for 3d molecular property prediction and beyond. In *International Conference on Learning Representations* (2022).

60. Jin, W. *et al.* Dsmbind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv* 2023–12 (2023).

61. Adasme, M. F. *et al.* Plip 2021: Expanding the scope of the protein–ligand interaction profiler to dna and rna. *Nucleic acids research* **49**, W530–W534 (2021).

62. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, 29287–29303 (Curran Associates, Inc., 2021).

63. Mallet, V., Oliver, C., Broadbent, J., Hamilton, W. L. & Waldispühl, J. Rnaglib: a python package for rna 2.5 d graphs. *Bioinformatics* **38**, 1458–1459 (2022).

64. Wyss, L., Mallet, V., Karroucha, W., Borgwardt, K. M. & Oliver, C. A comprehensive benchmark for RNA 3D structure-function modeling. *ArXiv* **abs/2503.21681** (2025).

65. Schlichtkrull, M. *et al.* Modeling relational data with graph convolutional networks (2017). 1703.06103.

66. Joshi, C. K. *et al.* gRNAde: Geometric deep learning for 3d RNA inverse design. In *The Thirteenth International Conference on Learning Representations* (2025).

67. Wang, N. *et al.* Multi-purpose rna language modelling with motif-aware pretraining and type-guided fine-tuning. *Nature Machine Intelligence* **6**, 548–557 (2024).

68. Shen, T. *et al.* Accurate rna 3d structure prediction using a language model-based deep learning approach. *Nature Methods* **21**, 2287–2298 (2024).

69. Ontiveros-Palacios, N. *et al.* Rfam 15: Rna families database in 2025. *Nucleic acids research* **53**, D258–D267 (2025).

70. Simonovsky, M. & Meyers, J. Deeplytough: learning structural comparison of protein binding sites. *Journal of chemical information and modeling* **60**, 2356–2366 (2020).

71. Li, S. *et al.* Pocketanchor: Learning structure-based pocket representations for protein-ligand interaction prediction. *Cell Systems* **14**, 692–705 (2023).

72. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics* **10**, 168 (2009).

73. ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale Website (2024). Accessed 2026-02-09.

74. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).

75. Basse, M. J. *et al.* 2p2idb: a structural database dedicated to orthosteric modulation of protein–protein interactions. *Nucleic acids research* **41**, D824–D827 (2012).

76. Krivtsov, A. V. *et al.* A menin-mll inhibitor induces specific chromatin changes and eradicates disease in models of mll-rearranged leukemia. *Cancer cell* **36**, 660–673 (2019).

77. He, S. *et al.* High-affinity small-molecule inhibitors of the menin-mixed lineage leukemia (mll) interaction closely mimic a natural protein–protein interaction. *Journal of medicinal chemistry* **57**, 1543–1556 (2014).

78. Zhang, Z., Morstein, J., Ecker, A. K., Guiley, K. Z. & Shokat, K. M. Chemoselective covalent modification of k-ras (g12r) with a small molecule electrophile. *Journal of the American Chemical Society* **144**, 15916–15921 (2022).

79. Kessler, D. *et al.* Drugging all ras isoforms with one pocket. *Future medicinal chemistry* **12**, 1911–1923 (2020).

80. Cammisa, M., Correra, A., Andreotti, G. & Cubellis, M. V. Identification and analysis of conserved pockets on protein surfaces. *BMC bioinformatics* **14**, 1–9 (2013).

81. Harel, A., Bromberg, Y., Falkowski, P. G. & Bhattacharya, D. Evolutionary history of redox metal-binding domains across the tree of life. *Proceedings of the National Academy of Sciences* **111**, 7042–7047 (2014).

82. Varadi, M. *et al.* Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research* **52**, D368–D375 (2024).

83. Durairaj, J. *et al.* Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646–653 (2023).

84. Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F. & Dal Peraro, M. Pesto: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nature Communications* **14**, 2175 (2023).

85. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* 1–3 (2024).

86. Li, T., Bonkovsky, H. L. & Guo, J.-t. Structural analysis of heme proteins: implications for design and prediction. *BMC structural biology* **11**, 1–13 (2011).

87. Syed, K. & Mashele, S. S. Comparative analysis of p450 signature motifs exxr and cxg in the large and diverse kingdom of fungi: Identification of evolutionarily conserved amino acid patterns characteristic of p450 family. *PLoS One* **9**, e95616 (2014).

88. Gane, A. *et al.* Protnlm: Model-based natural language protein annotation (2022).

89. Bajusz, D., Rácz, A. & Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* **7**, 1–13 (2015).

90. Steinegger, M. & Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026–1028 (2017).

91. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics* **148**, 241722 (2018).

92. Klicpera, J., Becker, F. & Günnemann, S. GemNet: Universal directional graph neural networks for molecules. In *Conference on Neural Information Processing Systems (NeurIPS)* (2021).

93. Liu, Y. *et al.* Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013* (2021).

94. Geiger, M. & Smidt, T. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453* (2022).

95. Godwin, J. *et al.* Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971* (2021).

96. Zaidi, S. *et al.* Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations* (2022).

97. Feng, S., Ni, Y., Lan, Y., Ma, Z.-M. & Ma, W.-Y. Fractional denoising for 3d molecular pre-training. In *International Conference on Machine Learning*, 9938–9961 (PMLR, 2023).

98. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239* (2020).

99. Liaw, R. *et al.* Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118* (2018).

100. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019).

101. Li, L. *et al.* A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems* **2**, 230–246 (2020).

102. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

103. Fey, M. & Lenssen, J. E. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428* (2019).

104. Biewald, L. Experiment tracking with weights and biases (2020). Software available from wandb.com.

105. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. & Dror, R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411* (2020).

106. Sanner, M. F., Olson, A. J. & Spehner, J.-C. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305–320 (1996).

107. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction (2020). 1802.03426.

108. Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8 (2015).

109. Hartrampf, N. *et al.* Synthesis of proteins by automated flow chemistry. *Science* **368**, 980–987 (2020).

110. Simon, M. D. *et al.* Rapid flow-based peptide synthesis. *ChemBioChem* **15**, 713–720 (2014).

111. Gasteiger, E. *et al.* Expasy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research* **31**, 3784–3788 (2003).

112. Fittolani, G., Kutateladze, D. A., Loas, A., Buchwald, S. L. & Pentelute, B. L. Automated flow synthesis of artificial heme enzymes for enantiodivergent biocatalysis. *Journal of the American Chemical Society* (2025).

113. Franzen, S. & Boxer, S. G. On the origin of heme absorption band shifts and associated protein structural relaxation in myoglobin following flash photolysis. *Journal of Biological Chemistry* **272**, 9655–9660 (1997).