SE3Lig: SE(3)-equivariant CNNs for the reconstruction of cofactors and ligands in protein structures

Siddharth Bhadra-Lobo*
Center for Computational and Integrative Biology
Rutgers University, Camden, NJ 08103
sid.bl@rutgers.edu

Anushriya Subedy*
Center for Computational and Integrative Biology
Rutgers University, Camden, NJ 08103
as3190@rutgers.edu

Sagar D. Khare
Department of Chemistry & Chemical Biology and Institute for Quantitative Biomedicine
Rutgers University, Piscataway, NJ 08854
khare@chem.rutgers.edu

Guillaume Lamoureux
Department of Chemistry and Center for Computational and Integrative Biology
Rutgers University, Camden, NJ 08103
guillaume.lamoureux@rutgers.edu

Abstract

Protein structure prediction algorithms such as AlphaFold2 and ESMFold have dramatically increased the availability of high-quality models of protein structures. Because these algorithms predict only the structure of the protein itself, there is a growing need for methods that can rapidly screen protein structures for ligands. Previous work on similar tasks has shown promise but is lacking scope in the classes of atoms predicted and can benefit from the recent architectural developments in convolutional neural networks (CNNs). In this work, we introduce SE3Lig, a model for semantic in-painting of small molecules in protein structures. Specifically, we report SE(3)-equivariant CNNs trained to predict the atomic densities of common classes of cofactors (hemes, flavins, etc.) and the water molecules and inorganic ions in their vicinity. While the models are trained on high-resolution crystal structures of enzymes, they perform well on structures predicted by AlphaFold2, which suggests that the algorithm correctly represents cofactor-binding cavities.

1 Introduction

The biological activity of a protein is often dictated by its interaction with other proteins and with a number of ligands such as coenzymes, inorganic ions, and water molecules. With the emergence of protein structure prediction models such as AlphaFold2 [1] and ESMFold [2], it has become essential

*equal contributors

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

to develop new tools to rapidly detect whether a protein structure is likely to bind a certain class of ligands, and precisely where those ligands would fit.

Generating ligands from protein cavities is usually performed in the context of structure-based drug design, with the explicit goal of generating novel ligands compatible with a known protein target. While the literature on the topic is vast, a number of "neuralized" solutions have recently been proposed (see [3] for a recent review). Masuda, Ragoza, and Koes [4, 5] have trained a CNN-based variational autoencoder (VAE) that generates 3D ligand densities conditioned on 3D receptor densities. The ligand densities generated are then analyzed to construct ligand molecules, which are assessed for chemical validity and novelty. Luo et al. [6] have proposed a model that encodes the receptor structure using a rotationally and translationally-invariant graph neural network, and that classifies each location on a grid according to the type of ligand atom it can support. Ligand molecules are generated using an autoregressive procedure, by treating each new ligand atom generated as part of the receptor structure and updating the ligand probabilities from which to sample the next ligand atom.

A number of purely graph-based models have been developed as well. Drotar et al. [7] have trained a graph-based VAE architecture that encodes the binding pocked and decodes a graph for the ligand molecule, which is then filtered for validity (drug-likedness and synthesizability) and re-docked into the pocket. Liu et al. [8] have used an SE(3)-equivariant graph-based approach, with an autoregressive procedure that generates the ligand atoms directly into the pocket and does not require re-docking.

By contrast, the present work focuses on quickly assessing protein structures for their ability to bind commonly found cofactors. While drug discovery approaches focus on predicting a generic description of ligand features across a wide chemical variety, we focus on predicting the specific localization of a small class of compounds. It is much more similar in scope to methods like AlphaFill [9], that "fills in" common cofactors, nucleotides, and metal ions using structural homology, and like Metal3D [10], that uses a probability density generated from a 3D-CNN trained on zinc ion binding sites.

Compared to a graph-based method, the main advantage of a grid-based method is that it can predict any number of cofactors or ligands. (Graph-based methods are typically built on the assumption that a single molecule is to be generated.) Although grid-based molecular representations are much sparser than graph-based representations and create a higher computational overhead, they can readily account for any non-protein atoms in the volume.

2 Methods

2.1 Datasets

We queried the Protein Data Bank (PDB) using the cofactor classification from Ref. [11], keeping only the structures with a resolution below 3.25 Å and excluding all PDB component IDs with fewer than 20 structures (see Table 1). A separate model is trained for each cofactor class: HEM, NAD, FAD, FMN, and SAM. Two additional models are trained on ZN and ATP components (even though they are not cofactors). Models are trained to predict the atomic densities of the components listed in Table 1 along with the densities of 12 common small ligands including water, metal ions, and inorganic ions (see Supplementary Table S1 for the full list). (We expect that the location of a cofactor is easier to infer if small ligands such as water molecules and inorganic ions are inferred as well.)

For each PDB structure, only the first instance of the cofactor was considered and only the atoms having the same chain ID as that cofactor were kept. (All other atoms were discarded.). For each example, the input corresponds to five $51 \times 51 \times 51$ grids of 0.5 Å resolution representing the protein atoms surrounding the cofactor molecule. The channels consist of four "chemical element" channels (C, N, O, S) and one "total" channel (the sum of the previous four). The ground truth corresponds to six $51 \times 51 \times 51$ grids representing the atomic densities of the cofactor molecule (C, N, O, S, Fe, P) and twelve $51 \times 51 \times 51$ grids representing the density of the small ligands (see Supplementary Table S1). Each atom is represented on the grid as a Gaussian density with standard deviation $\sigma = 1.0$ Å.

Each model is trained on 1000 examples randomly selected from the complete PDB dataset (see Table 1) and validated on 100 examples. A second validation set was generated by substituting the

Table 1: Cofactors considered in each model and their occurrence in the dataset.

Model	Cofactor class	PDB IDs	Occurrence*
HEM	Heme	HEM, HEA, HEC	5147
NAD	Nicotinamide adenine dinucleotide	NAD, NAI, NDP	2597
FAD	Flavin adenine dinucleotide	FAD	2044
FMN	Flavin mononucleotide	FMN	1234
SAM	S-adenosylmethionine	SAM, SFG, SAH	1702
ZN	Zinc	ZN	10156
ATP	Adenosine triphosphate	ATP	1046

*Only 1000 examples are used for training.

experimentally-determined protein structure by its corresponding AlphaFold prediction (aligned on the C_α atoms of the crystal structure). The AlphaFold validation set consists of 100 examples for each cofactor, except for ZN and ATP which have 87 examples each (see Supplementary Table S2).

2.2 Architecture

The SE3Lig architecture is presented in Figure 1. For each training example, 18 cofactor/ligand volumetric features are computed from 5 protein features, using an SE(3)-equivariant convolutional neural network (CNN) [12]. The network has four SE(3)-equivariant convolutional layers, each using a fixed kernel size of $9 \times 9 \times 9$. The first layer maps the five input channels onto 16 scalars, 16 vectors and 16 tensors, the second layer maps these onto 16 new scalars, 16 new vectors and 8 new tensors, the third layer operates like the second layer, and the fourth layer maps these to 18 scalars (see Supplementary material for details on architectural decisions). Each convolutional layer is followed by a hyperbolic tangent ("tanh") nonlinearity applied on the element-wise norm of each feature, which preserves SE(3) equivariance. Specifically, the network generates scalar feature maps for each molecular category: the 6 atom channels composing the cofactors, plus 12 channels for each of the 12 ligand types listed in Supplementary Table S1).

2.3 Training

The final 18 predicted scalar features are used to calculate the cross-entropy between the predicted and ground-truth atomic densities. The loss function is defined as the cross-entropy between 19 scalar features and 19 scalar target channels. The 18 predicted and ground-truth channels are concatenated with 1 additional channel set to a value $\epsilon = 10^{-3}$, which represents the "background" probability of having no cofactor or ligand at a given location. The loss is written as:

$$\mathcal{L} = \mathcal{L}_{\text{cof}} + \mathcal{L}_{\text{lig}} \tag{1}$$

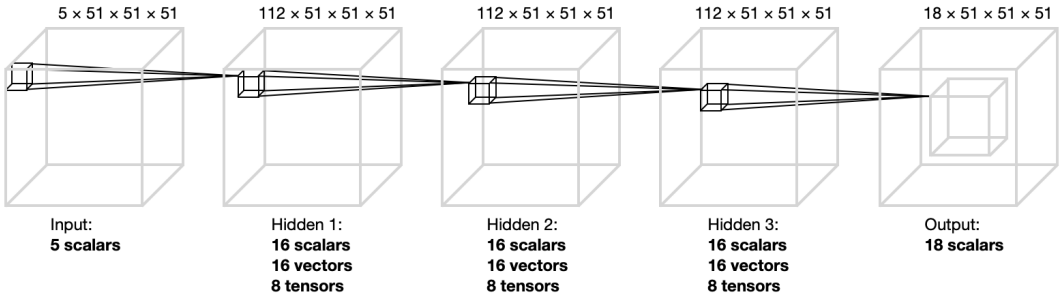

Figure 1: Overall architecture of the SE3Lig model. The model receives as input the 3D atomic structure of the protein projected on a $51 \times 51 \times 51$ grid of 0.5 Å resolution, broken down into 5 channels: one for each chemical element of the protein (C, N, O, S) and a fifth channel corresponding to the sum of the previous four. The input is transformed by an SE(3)-equivariant CNN [12], leading to a $18 \times 51 \times 51 \times 51$ array describing the densities of 18 classes of atoms. Each convolution is followed by a "tanh" nonlinearity applied on the norm of each feature (scalar, vector, or tensor).

with $\mathcal{L}_{\mathrm{cof}}$ containing the contributions from the 6 "cofactor" output channels and $\mathcal{L}_{\mathrm{lig}}$ the contributions from the 12 "ligand" channels.

Loss values alone were not very informative during training, especially when making predictions on 18 sparse or empty channels. For each cofactor class, we also monitored the root mean square distance (RMSD, in Å) between the ground-truth position of an *indicator atom* (Fe for HEM, P for NAD, FAD, FMN and ATP, S for SAM, and Zn for ZN) and the coordinate of the grid point for which that atom has the highest predicted density. For cofactor classes with more than one indicator atom per molecule (NAD, FAD, and ATP), we used the average distance between the predicted "$\arg\max$" coordinate and the ground-truth coordinates of the multiple indicator atoms. Since phosphorus atoms in diphosphate and triphosphate moieties are typically 3 Å apart, the RMSD can be at best ~ 1.5 Å for diphosphate-containing NAD and FAD classes and at best ~ 2 Å for triphosphate-containing ATP class.

RMSD values are used to monitor training and, during evaluation, to determine the usefulness of the output densities at predicting the location of the cofactor. The lower the RMSD is, the closer the highest scoring predicted indicator voxel is to the ground-truth atom coordinate. The lower the loss, the better the precision is of the overall predicted density in that channel as off target values are penalized in cross-entropy loss. The RMSD values are limited by the resolution of the grid and (statistically) can be no less than $0.5 \text{ Å}/\sqrt{6} = 0.204$ Å.

3 Results

3.1 Cofactors

Table 2 reports the cross-entropy losses and indicator atom RMSD (in Å) on the PDB and AlphaFold validation sets for each of the 7 cofactor models. Loss $\mathcal{L}_{\mathrm{cof}}$ is broken down into 6 contributions C, N, O, S, Fe, and P. (A breakdown of the 12 contributions to $\mathcal{L}_{\mathrm{lig}}$ is reported in Supplementary Table S3.) The individual channel contributions to these losses are standardized by dividing each loss component by the number of atoms present in the ground-truth structure, so that they describe the per-atom loss in a given channel. The values in gray represent losses from a channel for which no atoms are expected to be predicted. (For instance, there are no sulfur or phosphate atoms in heme.) The models successfully learn that these channels should be "turned off", leading to a uniform constant loss value across all "off" channels once the softmax function is applied to all 19 channels (including the background channel set to ϵ).

Predictions are typically better for the PDB validation set than for the AlphaFold validation set—at least according to the RMSD metric (see Table 2). This is expected since the models are trained on crystal structures and since AlphaFold predictions are not explicitly accounting for the presence of ligands.

Figure 2 shows a typical output from the trained HEM model, overlaid on the ground-truth structure of the cofactor and water molecules. The model can clearly identify all atomic classes of the heme group (C, N, O, Fe) and successfully predicts the location of many water molecules—along with a number of false negatives and false positives. For this example, the predicted "$\arg\max$" Fe indicator atom is 0.64 Å away from the ground-truth coordinate, comparable to the 0.78 Å RMSD for the PDB validation set (see Table 2). (See Supplementary Figure S1 for the distributions of all distances used in calculating the RMSD values of Table 2.)

3.2 Water molecules and inorganic ions

$\mathcal{L}_{\mathrm{lig}}$ values are difficult to compare across models due to the inconsistent empty channel predictions after the softmax operation. In general, HOH density predictions are similar to ground truth, although there are notable false negatives and false positives (see Figure 2E), which significance is not explored in this work. It is likely that some of the false positives represent weakly-bound water molecules.

Since all 7 cofactor models were jointly trained on ligand densities, it is interesting to assess whether that part of the output is transferable from one model to another. Table 3 shows the 7×7 matrix of ligand losses for each model on each PDB validation set. Interestingly, the models evaluated on their own PDB validation set (values on the diagonal, underlined) do not always perform better than other models transferred then evaluated—although they are generally better than the average (bottom row

Table 2: Loss ($\times 10^4$) on the cofactor and ligand density prediction for the PDB and AlphaFold validation sets. Loss terms for which no (nonzero) atomic densities are ever to be predicted are grayed out. (See Supplementary Figure S1 for indicator atom distance distributions.)

Model	Evaluation set	$\mathcal{L}_{\text{cof}}$	C	N	O	S	Fe	P	$\mathcal{L}_{\text{lig}}$	RMSD (Å)
HEM	PDB	0.64	0.04	0.09	0.15	0.10	0.16	0.10	1.32	0.78
	AlphaFold	1.10	0.09	0.14	0.23	0.17	0.29	0.17	2.25	2.46
NAD	PDB	0.71	0.06	0.12	0.09	0.13	0.13	0.17	1.75	4.28*
	AlphaFold	0.94	0.10	0.17	0.12	0.16	0.16	0.22	2.14	4.63*
FAD	PDB	0.88	0.07	0.14	0.11	0.19	0.19	0.20	2.30	3.37*
	AlphaFold	0.80	0.09	0.15	0.11	0.13	0.13	0.20	1.70	3.41*
FMN	PDB	1.23	0.07	0.16	0.12	0.26	0.26	0.36	3.19	1.98
	AlphaFold	1.20	0.08	0.18	0.13	0.23	0.23	0.35	2.84	2.26
SAM	PDB	0.81	0.06	0.10	0.13	0.27	0.12	0.12	1.61	5.07
	AlphaFold	0.84	0.08	0.12	0.16	0.28	0.10	0.10	1.49	5.75
ZN	PDB	2.25	0.38	0.38	0.38	0.38	0.38	0.38	4.50	1.03
	AlphaFold	2.80	0.47	0.47	0.47	0.47	0.47	0.47	5.72	2.06
ATP	PDB	1.53	0.13	0.20	0.13	0.40	0.40	0.27	5.03	4.35*
	AlphaFold	1.25	0.15	0.21	0.14	0.24	0.24	0.26	3.24	4.36*

*More than one indicator atom

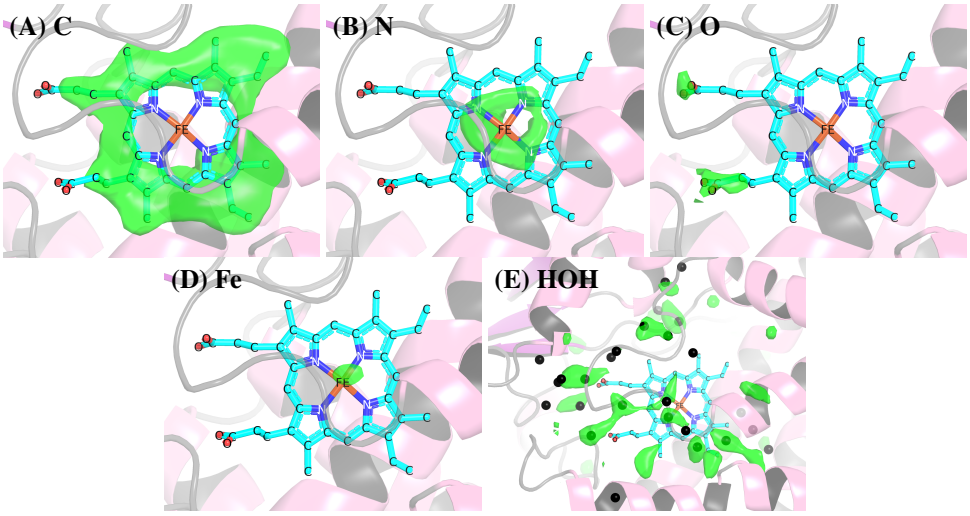

Figure 2: Example HEM prediction on a protein structure from the PDB validation set, shown as probability isosurfaces (in green). Individual channel predictions superimposed to associated ground-truth atomic positions for heme carbon (panel A), heme nitrogen (panel B), heme oxygen (panel C), heme iron (panel D), and water (panel E). The voxel with highest probability for iron is 0.64 Å away from the ground-truth iron coordinate.

of Table 3). The ZN model systematically outperforms other models, with an average ligand loss of 2.79 across all datasets, likely because its training dataset shows the highest number (and diversity) of ligands (see Supplementary Table S1) Conversely, the HEM validation set appears to represent the easiest ligand prediction task, with an average ligand loss of 1.31 across all models.

4 Conclusion

The present work shows that cofactors and ligands missing in experimentally determined structures and AI-predicted structures can be in-painted using a relatively straightforward SE(3)-equivariant CNN architecture dubbed SE3Lig. Since these models are fully convolutional, the learned filters can be directly applied to protein structures of any size, thus allowing for rapid cofactor and ligand

Table 3: Performance of each cofactor model at predicting water molecules and ions from the PBD validation sets of other cofactor models. The value reported is $\mathcal{L}_{\text{lig}} \times 10^4$. "Average" column on the right represents the average performance of a given model over all validation sets. "Average" row at the bottom represents the average performance of all models for a given validation set.

Model	PDB validation set							
	HEM	NAD	FAD	FMN	SAM	ZN	ATP	Average
HEM	1.32	2.12	2.60	4.99	2.63	5.90	5.65	3.60
NAD	1.29	1.75	2.55	3.23	1.56	5.18	5.23	2.98
FAD	1.39	1.98	2.30	3.45	1.75	6.03	5.82	3.25
FMN	1.31	1.94	**2.16**	3.19	1.71	7.08	5.50	3.27
SAM	1.33	1.89	2.67	3.32	1.61	5.85	5.53	3.17
ZN	**1.22**	**1.62**	2.61	**3.04**	**1.48**	**4.50**	5.06	**2.79**
ATP	1.28	1.84	2.63	3.31	1.64	5.60	**5.03**	3.05
Average	**1.31**	1.88	2.50	3.50	1.77	5.73	5.40	

screening and validation. The architecture shows good performance on many classes of cofactors and strongly-bound ligands and it is expected that, using expanded datasets, it can be trained on a greater variety of ligands.

Given its simplicity, the SE3Lig model can likely be used in a scoring function for protein design algorithms, for which designed structures are screened for their propensities at binding certain cofactors or ligands at certain locations. In particular, the model may help develop approaches similar to the ligand-conditioned protein structure prediction algorithm NeuralPLexer [13], for which AI-generated structures would be conditioned on ligands based on the ligand probability densities generated using the SE3Lig. More straightforwardly, those ligand probability densities may also be used to perform cofactor and ligand docking.

Acknowledgments and Disclosure of Funding

This work was supported by NSF award number 2226816 to S.D.K. and G.L. and by a Busch Biomedical Grant to G.L. The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers University for providing access to the Amarel cluster and associated research computing resources.

References

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

[2] Zeming Lin, Halil Akin1, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[3] Rıza Özçelik, Derek van Tilborg, José Jiménez-Luna, and Francesca Grisoni. Structure-based drug discovery with deep learning. *arXiv:2212.13295*, 2022.

[4] Tomohide Masuda, Matthew Ragoza, and David Ryan Koes. Generating 3D molecular structures conditional on a receptor binding site with deep generative models. *arXiv:2010.14442*, 2020.

[5] Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3D molecules conditional on receptor binding sites with deep generative models. *arXiv:2110.15200*, 2021.

[6] Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3D generative model for structure-based drug design. *arXiv:2203.10446*, 2022.

[7] Pavol Drotár, Arian Rokkum Jamasb, Ben Day, Cătălina Cangea, and Pietro Liò. Structure-aware generation of drug-like molecules. *arXiv:2111.04107*, 2021.

[8] Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. Generating 3D molecules for target protein binding. *arXiv:2204.09410*, 2022.

[9] Maarten L. Hekkelman, Ida de Vries, Robbie P. Joosten, and Anastassis Perrakis. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature Methods*, 20:205–2015, 2022.

[10] Simon L. Dürr, Andrea Levy, and Ursula Rothlisberger. Metal3D: a general deep learning framework for accurate metal ion location prediction in proteins. *Nature Communications*, 14:2713, 2023.

[11] Abhik Mukhopadhyay, Neera Borkakoti, Lukáš Pravda, Jonathan D. Tyzack, Janet M. Thornton, and Sameer Velankar. Finding enzyme cofactors in Protein Data Bank. *Bioinformatics*, 34(23):4046–4053, 2019.

[12] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3D steerable CNNs: Learning rotationally equivariant features in volumetric data. *arXiv:1807.02547*, 2018.

[13] Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F. Miller III, and Animashree Anandkumar. State-specific protein-ligand complex structure prediction with a multi-scale deep generative model. *arXiv:2209.15171*, 2023.

Supplementary material

Table S1: Ligands modeled and their occurrence in the dataset

		Occurrence (for each cofactor class)						
Ligand type	PDB ID	HEM	NAD	FAD	FMN	SAM	ZN	ATP
Calcium	CA	457	75	58	42	72	1640	172
Copper	CU	157	3	22	1	0	153	0
Dinuclear copper	CUA	0	0	0	0	0	7	0
Iron-sulfur cluster	FES	10	14	96	9	7	9	2
Water	HOH	972438	457866	476342	219488	274966	1828814	149529
Potassium	K	162	69	44	9	75	534	68
Magnesium	MG	230	391	202	54	229	1541	828
Manganese	MN	44	72	5	9	15	156	80
Sodium	NA	297	287	98	74	131	649	36
Iron-sulfur cluster	SF4	0	32	129	202	97	28	2
Sulfate ion	SO4	1717	754	874	434	570	3208	242
Zinc	ZN	720	305	113	35	639	10156	69

Table S2: AlphaFold generated validation set example counts

Model	Cofactor class	PDB IDs	Number of examples
HEM	Heme	HEM, HEA, HEC	478
NAD	Nicotinamide adenine dinucleotide	NAD, NAI, NDP	214
FAD	Flavin adenine dinucleotide	FAD	172
FMN	Flavin mononucleotide	FMN	104
SAM	S-adenosylmethionine	SAM, SFG, SAH	137
ZN	Zinc	ZN	87
ATP	Adenosine triphosphate	ATP	87

Table S3: $\mathcal{L}_{\text{lig}}$ loss ($\times 10^4$) and individual loss terms on the ligand density prediction for the PDB and AlphaFold validation sets.

Model	Validation set	$\mathcal{L}_{\text{lig}}$	CA	CU	CUA	FES	HOH	K	MG	MN	NA	SF4	SO4	ZN
HEM	PDB	1.32	0.10	0.10	0.10	0.10	0.20	0.10	0.11	0.11	0.10	0.10	0.10	0.11
	AlphaFold	2.25	0.17	0.17	0.17	0.17	0.33	0.17	0.18	0.18	0.17	0.17	0.17	0.18
NAD	PDB	1.75	0.13	0.13	0.13	0.13	0.24	0.14	0.15	0.14	0.13	0.14	0.13	0.15
	AlphaFold	2.14	0.16	0.16	0.16	0.16	0.26	0.17	0.18	0.18	0.16	0.18	0.16	0.20
FAD	PDB	2.30	0.19	0.19	0.19	0.19	0.22	0.19	0.19	0.19	0.19	0.19	0.20	0.19
	AlphaFold	1.70	0.13	0.13	0.13	0.13	0.24	0.13	0.13	0.13	0.13	0.13	0.14	0.13
FMN	PDB	3.19	0.26	0.26	0.26	0.27	0.28	0.26	0.27	0.26	0.26	0.28	0.26	0.26
	AlphaFold	2.84	0.23	0.23	0.23	0.23	0.28	0.23	0.24	0.23	0.23	0.25	0.23	0.24
SAM	PDB	1.61	0.12	0.12	0.12	0.12	0.26	0.12	0.12	0.13	0.12	0.13	0.12	0.13
	AlphaFold	1.49	0.10	0.10	0.11	0.10	0.28	0.11	0.12	0.11	0.10	0.12	0.10	0.12
ZN	PDB	4.50	0.38	0.38	0.38	0.38	0.38	0.39	0.40	0.38	0.38	0.39	0.39	0.46
	AlphaFold	5.72	0.47	0.47	0.47	0.47	0.37	0.48	0.48	0.47	0.47	0.48	0.47	0.63
ATP	PDB	5.03	0.40	0.40	0.40	0.40	0.41	0.42	0.54	0.43	0.40	0.41	0.40	0.42
	AlphaFold	3.24	0.24	0.24	0.24	0.24	0.35	0.25	0.40	0.27	0.24	0.25	0.24	0.26

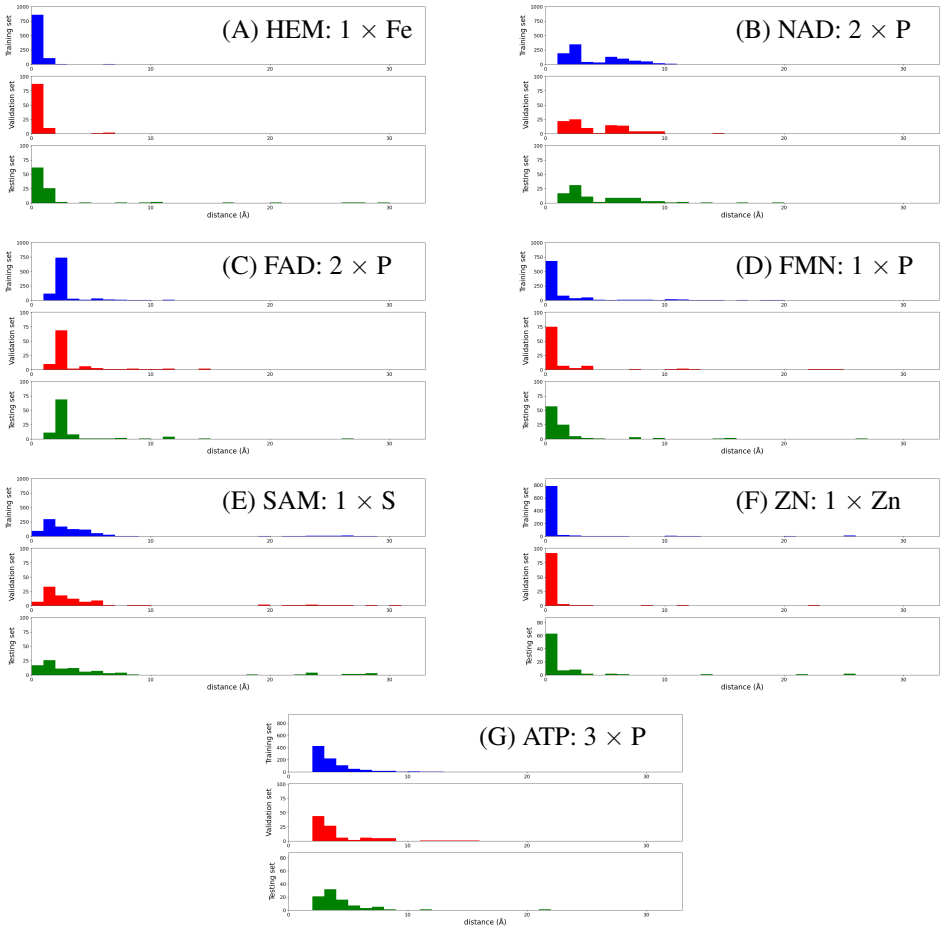

Figure S1: Distributions of indicator atom distances for each of the 7 SE3Lig cofactor models, for the training set (blue), PDB validation set (red), and AlphaFold validation set (green). For cofactors with a single indicator atom (HEM, FMN, SAM, and ZN), the distance is measured between the position of the "$\arg\max$" voxel and the ground-truth position of the atom. For cofactors with more than one indicator atoms (NAD, FAD, and ATP), the average distance of the "$\arg\max$" voxel to all ground-truth atoms is reported.

Ablation studies

The HEM model was used in an ablation study to determine the adequate depth and width of the architecture. Table S4 shows the various experiments conducted using different numbers of layers and features.

The effect layer depth has on the loss and indicator atom RMSD suggesting a minimum of 3 layers is needed to develop focused density features in our in-painting approach. The addition of subsequent layers improves the predicted density precision and indicator atom RMSD accuracy, however we decided 4 layers were sufficient.

Higher-rank features (i.e. vectors and tensors) can both improve the loss and indicator atom RMSD, however they contribute more to the precision of the predicted density than to the accuracy in the placement of atom density "hotspots". Similarly, the replacement of higher-rank features with lower-rank features degrades the precision of the model and is reflected in the relative increases in loss values.

The Fe atom placement accuracy is comparable across the 3 replacement experiments, suggesting that scalar, vector, and tensor features are sufficient to describe single atom channel positions, however higher order features better describe more complicated cofactor structures and ligands (e.g. 1 Fe atom vs. 4 N atoms, not reflected in indicator atom RMSD).

The removal of the 12 ligand channels does not improve loss on the cofactor atoms nor the RMSD and suggests that including ligand prediction targets helps the model with cofactor atom placement, especially HOH due to its overall abundance in the dataset (see Table S1).

Table S4: Effect of various network ablations on the performance of the HEM model on the PDB validation set. (All models trained for 200 epochs, batch size of 1. Last row shows a model trained using only $\mathcal{L}_{\text{cof}}$.)

Layers	Scalars	Vectors	Tensors	$\mathcal{L}_{\text{cof}} \times 10^4$	$\mathcal{L}_{\text{lig}} \times 10^4$	RMSD_{Fe} (Å)
6	16	16	8	0.31	0.43	0.90
4	16	16	8	0.64	1.32	0.78
3	16	16	8	1.32	2.95	2.48
2	16	16	8	3.48	7.81	14.9
4	16	16	8	0.64	1.32	0.78*
4	8	8	4	0.64	1.28	1.29
4	4	4	2	0.67	1.24	1.57
4	16	16	8	0.64	1.32	0.78*
4	24	16	0	0.65	1.34	0.90
4	40	0	0	0.71	1.45	0.74
4	16	16	8	0.64	1.32	0.78*
4	16	16	8	0.69	—	0.88

*Row duplicated from above.