Supplementary Information

Deep learning model predicts water interaction sites on the surface of proteins using limited-resolution data

Jan Zaucha[a,*], Charlotte A. Softley[b,c,*], Michael Sattler[b,c], Dmitrij Frishman[a], Grzegorz M. Popowicz[b,c]

[a]*Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Maximus-von-Imhof-Forum 3, 85354 Freising, Germany*
[b]*Biomolecular NMR and Center for Integrated Protein Science Munich at Department Chemie, Technical University of Munich, Lichtenbergstraße 4, 85747, Garching, Germany.*
[c]*Institute of Structural Biology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany.*
Corresponding authors: d.frishman@wzw.tum.de, grzegorz.popowicz@helmholtz-muenchen.de
** These authors contributed equally*

Contents

Materials & Methods

1.1 Training dataset

In order to develop a binary regression model, training data providing the positive (empirically supported water binding sites) and negative (sites with no evidence of binding water) class examples are required.

The positive class data (water molecules from PDB structures) were extracted from a non-redundant representative set (sequence similarity <30%) of 9,067 PDB structures not containing nucleic acids, resolved to at least 1.8 Å resolution using X-ray diffraction (using the download tool provided by PDB[1]; data downloaded on 13.04.19). Water molecules not interacting with at least two atoms (proximity of water's oxygen to target atoms <4.5 Å) of the protein were excluded. Due to computational resource constraints, the positive dataset was limited to 2,800,000 individual water molecules.

Generating the negative class data poses a challenge since the number of positions not occupied by water molecules is virtually infinite. Selecting any positions not occupied by a water molecule would yield a mostly trivial dataset - positions within the core of the protein structure that are physically inaccessible to water molecules or positions too far away from any atoms of the structure. Therefore, the negative class samples were chosen using a heuristic approach, by randomly selecting positions near the protein's estimated solvation envelope as follows. Firstly, EDTSurf [2] was used to generate a triangular mesh matching the Coulombic radius of water (1.4 Å) around the protein (run with option "–f 20" to use the finest grid available)[3]. Based on empirical evidence, we found that the majority of water molecules in PDB structures can be found within 2.4-5.8 Å of the mesh points (this makes sense given that the length of the hydrogen bond is 1.5-4 Å[4]). The mesh was scaled outwards and inwards to a depth of 1.8 Å to cover cavities within the protein, using increments matching the mean grid size (average distance between two adjacent grid points, typically around 0.2 Å) to form multiple mesh layers around the protein (Fig. SI2). Negative class positions were then randomly sampled from the obtained layers, according to a probability distribution selected to estimate the distribution observed empirically (Gaussian distribution centered at layer corresponding to roughly 2.4 Å from the protein's surface, scaled by a factor of $\sqrt{1.5}$ times the number of grid layers). The selected points were verified to check that they were not closer than 1.4 Å to a water molecule encoded within the PDB file. The number of negative class samples generated was selected to match the number of positive samples, yielding a balanced training set.

A random subset of samples, amounting to roughly 15% of the training set, was defined as the "holdout" test set and excluded from the training and cross validation steps, for use once the final network architecture and model parameters had been defined. A list of PDB files used for the test set is available in Supplementary File 1.

1.2 Input features and data preprocessing

In this work, we have made an arbitrary, but intuitive, decision to encode the queried potential water binding site on the surface of the protein in terms of a linear array of vectors representing the full atomic neighborhood of the site (providing information on the distance to each atom as well as the corresponding atom type and temperature factor). Each positive and negative class sample was annotated with vectors pointing towards all proximal (<=7.5 Å) atoms of the protein. The target atom types were recorded according to their chemical properties, distinguishing between: aliphatic and alpha carbons, aromatic carbons, carbonyl carbons of the main chain, hydroxyl oxygens, hydroxyl aromatic, carboxyl oxygens, carbonyl oxygens of the side chain, carbonyl oxygens of the main chain, aromatic nitrogens of tryptophan, aromatic nitrogens of histidine, amide nitrogens of the side chain, amide nitrogens of the main chain, primary amine nitrogens, secondary amine nitrogens, amide carbons of the side chain, amide carbons of the main chain, carboxyl carbons, guanidinium carbons, sulfurs of methionine, sulfurs of cysteine. These twenty atom types are represented in the model using the "one hot encoding". Since temperature factors (B-values) have been shown to affect the probability of resolving water molecules within the crystal[5], their normalized values were added as additional inputs into the model.

The model was designed to consist of two input channels: the first layer provides vectors pointing from the (non-)water molecule position to the nearest fifty interacting atoms of the protein (matrix shape: 50x3), while the second layer provides B-values and atom-type information corresponding to each of the fifty vectors (matrix shape: 50x30, one hot encoding).

In order to facilitate model convergence and prediction performance, the landscape of input feature patterns that can appear should be condensed; in particular, rotational transformations of the input feature space introduce redundancies which greatly hinder the ability of the model to learn[6]. One way of mitigating the problem is training-data augmentation including random rotations of the each input feature sample[7] or data-preprocessing to ensure rotational invariance of the inputs. In this work, we opted for the latter solution; the vectors were rotated to a common system of coordinates in Euclidean space by Gram-Schmidt orthonormalization using the vector pointing to the closest atom of the protein ($\vec{v}_1$) and the orthogonal part of the vector at the highest angle against it ($\vec{v}_2$) as the first and second basis vectors for the orientation (the third vector forming the new orthonormal basis $\hat{e}_1, \hat{e}_2, \hat{e}_3$ is their cross product):

$$\hat{e}_1 = \frac{\vec{v}_1}{\left\|\vec{v}_1\right\|}, \quad \hat{e}_2 = \frac{\vec{v}_2 - \hat{e}_1\left(\hat{e}_1 \cdot \vec{v}_2\right)}{\left\|\vec{v}_2 - \hat{e}_1\left(\hat{e}_1 \cdot \vec{v}_2\right)\right\|}, \quad \hat{e}_3 = \hat{e}_1 \times \hat{e}_2$$

Each vector was rotated into the standardized system of coordinates by left-multiplying it by the matrix $M = \left[\hat{e}_3\ \hat{e}_2\ \hat{e}_1\right]$.

In order to allow for effective learning of patterns indicative of strong interactions with water molecules, the vectors were sorted into ordered pairs such that the first vector is the vector with the lowest norm, the second vector is the vector at the highest angle against it, the next vector pair is

again the vector with the lowest norm and the vector at the highest angle against it (drawn from the set of remaining vectors). All data were normalized to fall between [0,1] (B-values were $log10$-transformed beforehand). If the number of interacting atoms with the water molecule was less than fifty, the unoccupied positions were filled with empty values.

1.3 Neural network architecture and hyper-parameter tuning

Due to the vastness of possibilities, the selection of the neural network architecture was inspired by the previous successes in applying similar models to a different application domain[8,9]. The training set was split into six cross-validation folds (excluding the holdout test set) and various network depths (numbers of hidden layers and numbers of residual blocks) and combinations of hyper-parameters were evaluated to establish the choice that yields the highest classification accuracy on the validation set (corresponding to the lowest loss, measured in terms of binary cross-entropy).

The best performing network found has the following architecture: the two input layers (positional vectors and atom-types/B-values) are separately passed through a 1D convolution of thirty filters with linear activation functions (window size was set to four) and subsequently concatenated to mix atom type and B-value-based patterns with the patterns learnt from the array of input vectors describing the ordered positional vectors. The data is then duplicated and passed into two collateral paths, the first one being a series of two residual blocks, each containing a series of two batch normalization layers with rectified linear unit activation functions followed by 1D convolutional layers, and the second being skip connections that join each residual block after passing its respective output data through a 1D convolutional layer. The paths are then joined together in a final residual block followed by a 1D convolution with a single sigmoid activation kernel and, lastly, the dimensionality of the model is collapsed in the final max pooling layer serving as the output for predicting whether or not a position is likely to correspond to a water molecule "hot-spot". Altogether, the model comprises 34,141 trainable parameters (the model architecture is available for visual inspection in Figure SI1, and the model summary is available in Supplementary file 2).

In order to minimize over-fitting, L2 regularization was applied (at the regularization rate λ=0.005) to the kernel parameter weights of the convolutional layers. Dropout was not found to improve the prediction accuracy and was not used. Additionally, in order to make the model more robust in the event of adversarial perturbations (or other non-standard inputs), the adversarial regularization wrapper from the TensorFlow Neural Structured Learning Framework was applied on top of the developed model (using the multiplier of the adversarial loss set to 0.2 & step size of 0.05) [10,11]. Lastly, in agreement with previous reports[12], we observed that the 'adam' optimizer did not guarantee a stable convergence of the loss function and it had to be replaced with the classic stochastic gradient descent (with a learning rate of 0.01).

1.4 Training the network

The network described was trained for 100 epochs and achieved an accuracy of 94% on the test set (Fig. SI3A). The area under the receiver operating characteristic curve was over 0.985 (Fig. SI3B). It is important to note that the model performance determined here applies only to the training dataset

used in this study. The accuracy is expected to be highly sensitive to the number of "trivial" non-interacting sites the model was presented with during training. Likewise, filtering out the positive dataset to exclude water molecules for which there is insufficient evidence in the form of electron density clouds could further improve the classification accuracy. We have not performed this step since Nittinger et al. have shown that roughly 90% of water molecules encoded on the surfaces of proteins within PDB files were real[13]. While this fraction is already high, we have further increased it by excluding all water molecules not featuring interactions with at least two atoms of the protein structure (this was generally between 0-15% of water molecules encoded in each PDB file) or further away than 4.5Å of any atom of the protein. The main reason behind not using the dataset provided by Nittinger et al. is that it does not guarantee that the protein structures available are non-redundant – this could potentially lead to information leakage between folds of the cross validation and also the test set. Secondly, their dataset comprises only 2.3 million water molecules which is significantly smaller than the dataset we used.

1.5 Case examples for performance evaluation

In order to evaluate the performance of the model in a realistic setting, we sought examples of structures resolved using both high-resolution X-ray crystallography as compared to: 1) low-resolution X-ray crystallography; 2) NMR; and 3) Cryo-EM. Structures resolved using the latter techniques do not contain water molecules - they serve as the target structures for running predictions. The high-resolution structures, on the other hand, contain water molecule positions which serve as a validation set for the predictions.

To this end, we arbitrarily selected "popular" proteins, for which many structures were available in PDB. We selected three high resolution (<1.8 Å) structures resolved using X-ray crystallography: a carbonic anhydrase (resolved to 1.35 Å; PDB accession 3M1K[14]), cyclophilin A (resolved to 1.63 Å; PDB accession 2CPL[15]) and lysozyme (resolved to 1.4 Å; PDB accession 2Q9D[16]). We used TopSearch[17] to search for the most structurally homologous (in terms of low root mean square deviation - RMSD) counterparts resolved to a lower resolution or not relying on X-ray diffraction. The resultant homologues found included: a crystal structure of a carbonic anhydrase resolved to 3.5 Å (PDB accession 1G6V[18]; RMSD against 3M1K =0.73 Å), an NMR structure of cyclophilin A (PDB accession 1oca[19]; RMSD against 2CPL=0.83 Å) and a cryo-EM structure of a gamma-secretase assembly, which includes a lysozyme (resolved to 4.4 Å; PDB accession 4UIS[20]; RMSD against 2Q9D =1.09 Å). The rigid version of FATCAT[21] was used to superimpose structures on top of each other, so as to localize the binding sites of water molecules from the high-resolution structures within their low-resolution counterparts. Proteins exhibiting structural homology to the above listed structures (according to TopSearch) were excluded from the machine learning training set, so as to ensure independence of these test cases and avert the risk of reporting results obtained due to over-fitting.

Description of protein structures used in benchmark comparisons	Carbonic Anhydrase	Cyclophilin A	Lysozyme
PDB accession	3M1K/1G6V	2CPL/1OCA	2Q9D/4UIS
Experimental method	X-ray/X-ray	X-ray/NMR	X-ray/Cryo-EM
Resolution	1.35Å/3.5Å	1.6Å/-	1.35Å/4.4Å

RMSD	0.73Å	0.83Å	1.09Å
Num. water molecules	204/0	95/0	197/0

1.6 Protein crystallization and structure determination

The N-terminal domain of *T. cruzi* PEX14 was expressed and purified as previously described[22]. Protein was crystallised at 40 mg/ml in crystallization buffer containing 0.2 M Na acetate, 0.1 M Tris.HCl at pH 8.5 with 30% (w/v) PEG 4000. Crystals were grown at 20 °C and cryoprotected in glycerol. Samples were measured at the id30b beamline of the ESRF synchrotron (refinement statistics are provided in Table SI1). The data was indexed using XDS[23] and scaled using XScale. The initial phases were obtained using molecular replacement carried out using Phaser[24], with *T. brucei* PEX-14 as the search model. Manual rebuilding using electron density maps was carried out in Coot[25]. Phenix Elbow[26] was used for obtaining restraints for small molecules in the crystallization conditions. Further refinements were carried out using Phenix Refine. 5% of the reflections were used for cross-validation analysis. The final model was deposited in the PDB; PDB code 6ZFW. Figures were produced using Pymol[27], Coot, Affinity Designer and Inkscape. For comparative analysis the Coot *Find Waters* and Phenix *Update Waters* options were used with the default recommended settings.

PEX14tc	
Wavelength	0.82656
Resolution range	48.44 - 1.58 (1.636 - 1.58)
Space group	P 1 21 1
Unit cell	35.688 117.382 51.301 90 109.235 90
Total reflections	176261 (18072)
Unique reflections	53448 (5358)
Multiplicity	
Completeness (%)	94.2 (92.9)

Mean I/sigma(I)	9.5 (1.35)
Wilson B-factor	22.60
CC1/2	99.8 (61.2)
Reflections used in refinement	53441 (5356)
Reflections used for R-free	2656 (271)
R-work	0.1712 (0.2767)
R-free	0.2065 (0.3193)
Number of non-hydrogen atoms	3263
macromolecules	2884
ligands	40
solvent	329
Protein residues	345
RMS(bonds)	0.012
RMS(angles)	1.21
Ramachandran favored (%)	99.70
Ramachandran allowed (%)	0.30
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	0.31

Clashscore	9.60
Average B-factor	29.54
macromolecules	28.28
ligands	46.63
solvent	38.59

Table SI1. Data collection and refinement statistics.

References

1 S. K. Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J. M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranović, D. Guzenko, B. P. Hudson, T. Kalro, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Periskova, A. Prlić, C. Randle, A. Rose, P. Rose, R. Sala, M. Sekharan, C. Shao, L. Tan, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, M. Zhuravleva and C. Zardecki, *Nucleic Acids Res.*, 2019, **47**, D464–D474.

2 D. Xu and Y. Zhang, *PLoS One*, 2009, **4**, e8140.

3 A.-J. Li and R. Nussinov, *Proteins Struct. Funct. Genet.*, 1998, **32**, 111–127.

4 G. A. Jeffrey, *An introduction to hydrogen bonding*, Oxford University Press, 1997.

5 M. Gnesi, O. Carugo and IUCr, *J. Appl. Crystallogr.*, 2017, **50**, 96–101.

6 M. Egmont-Petersen, D. de Ridder and H. Handels, *Pattern Recognit.*, 2002, **35**, 2279–2301.

7 M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri and D. R. Koes, *J. Chem. Inf. Model.*, 2017, **57**, 942–957.

8 K. He, X. Zhang, S. Ren and J. Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, pp. 770–778.

9 L. Sundaram, H. Gao, S. R. Padigepati, J. F. McRae, Y. Li, J. A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzloglou, X. Li and K. K. H. Farh, *Nat. Genet.*, 2018, **50**, 1161–1170.

10 I. J. Goodfellow, J. Shlens and C. Szegedy, *arXiv Prepr. arXiv*.

11 T. D. Bui, S. Ravi and V. Ramavajjala, in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*, ACM Press, New York, New York, USA, 2018, pp. 64–71.

12 N. S. Keskar and R. Socher, *arXiv Prepr. arXiv*.

13 E. Nittinger, N. Schneider, G. Lange and M. Rarey, *J. Chem. Inf. Model.*, 2015, **55**, 771–783.

14 J. Schulze Wischeler, A. Innocenti, D. Vullo, A. Agrawal, S. M. M. Cohen, A. Heine, C. T. T.

Supuran and G. Klebe, *ChemMedChem*, 2010, **5**, 1609–1615.

15 H. Ke, *J. Mol. Biol.*, 1992, **228**, 539–550.

16 Z. Guo, D. Cascio, K. Hideg and W. L. Hubbell, *Protein Sci.*, 2008, **17**, 228–39.

17 M. Wiederstein, M. Gruber, K. Frank, F. Melo and M. J. Sippl, *Structure*, 2014, **22**, 1063–1070.

18 A. Desmyter, K. Decanniere, S. Muyldermans and L. Wyns, *J. Biol. Chem.*, 2001, **276**, 26285–90.

19 M. Ottiger, O. Zerbe, P. Güntert and K. Wüthrich, *J. Mol. Biol.*, 1997, **272**, 64–81.

20 L. Sun, L. Zhao, G. Yang, C. Yan, R. Zhou, X. Zhou, T. Xie, Y. Zhao, S. Wu, X. Li and Y. Shi, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 6003–8.

21 Y. Ye and A. Godzik, *Bioinformatics*, 2003, **19**, ii246–ii255.

22 M. Dawidowski, L. Emmanouilidis, V. C. Kalel, K. Tripsianes, K. Schorpp, K. Hadian, M. Kaiser, P. Mäser, M. Kolonko, S. Tanghe, A. Rodriguez, W. Schliebs, R. Erdmann, M. Sattler and G. M. Popowicz, *Science*, 2017, **355**, 1416–1420.

23 M. Krug, M. S. Weiss, U. Heinemann and U. Mueller, *J. Appl. Crystallogr.*, 2012, **45**, 568–572.

24 A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni and R. J. Read, *J. Appl. Crystallogr.*, 2007, **40**, 658–674.

25 P. Emsley, B. Lohkamp, W. G. Scott and K. Cowtan, *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 2010, **66**, 486–501.

26 P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger and P. H. Zwart, *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 2010, **66**, 213–221.

27 L. W. Delano, *http://www.pymol.org*.

Supplementary Figures

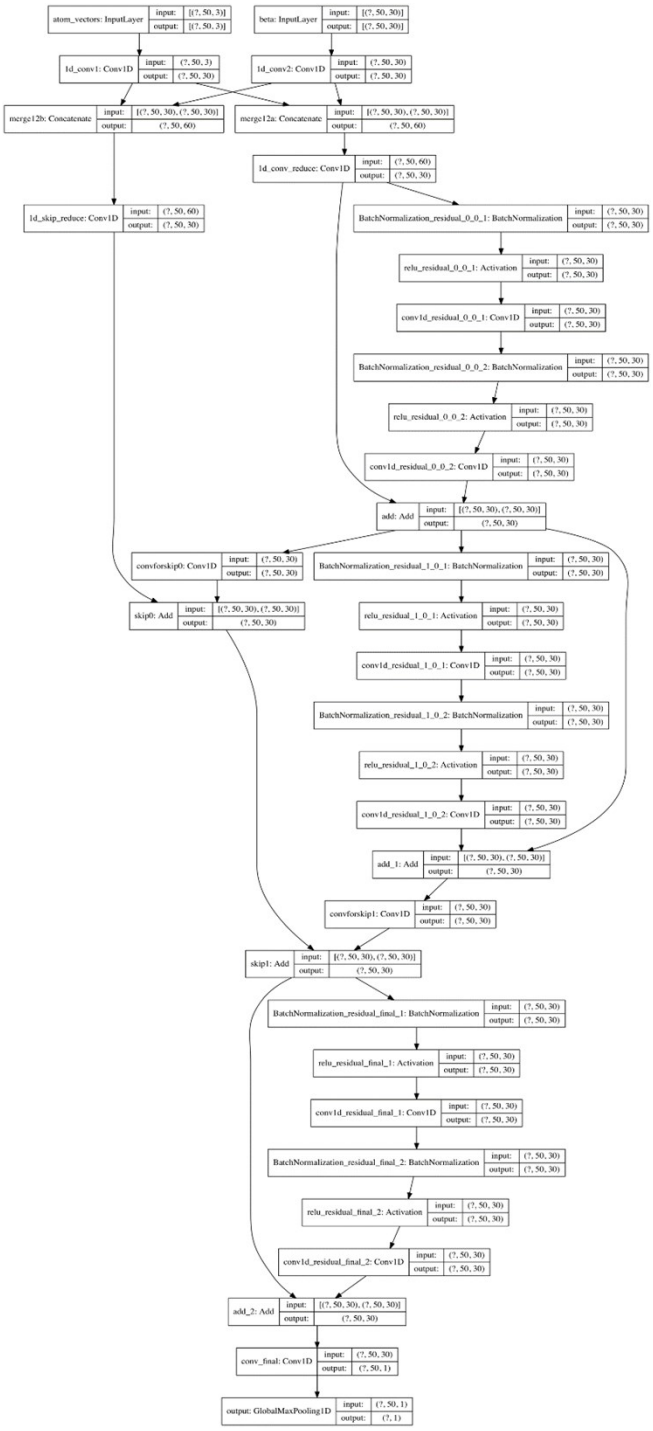

Figure SI1: Deep residual neural network model architecture.

Layers are represented by rectangles, while connectors are represented by arrows. The dimensionality of the input and output of each layer is specified on the right hand side of each rectangle.

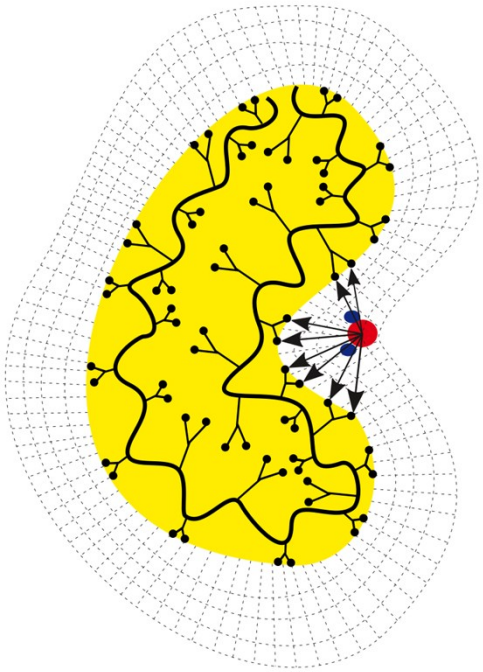

Figure SI2: Schematic diagram of mesh layers covering the area around a protein surface. Water molecules present within the grid are encoded in terms of an array of vectors pointing towards all proximal (<=7.5Å) atoms of the protein.

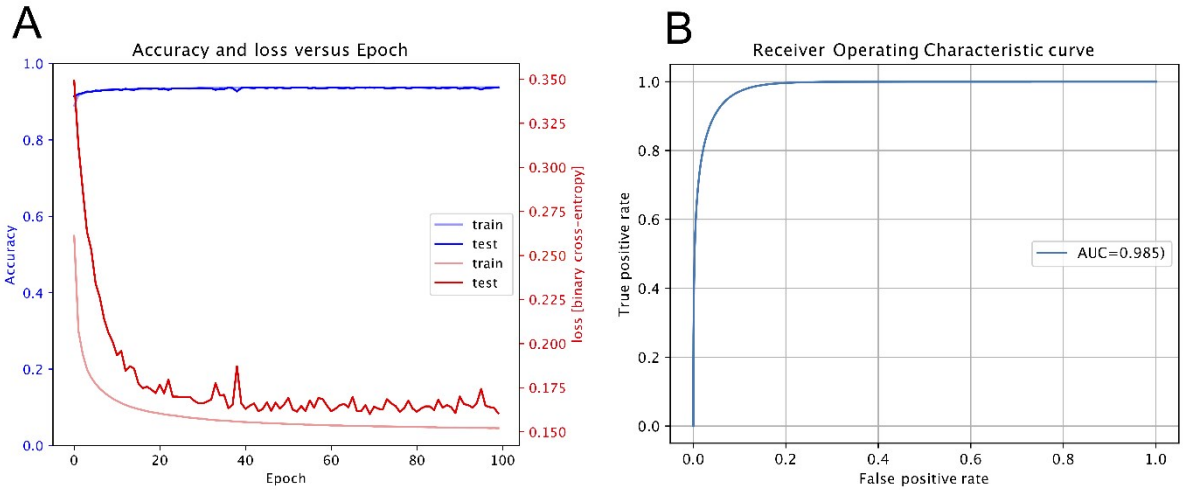

Figure SI3: Performance of the final model.

A) Left axis: model accuracy across training epochs (light blue and dark blue line for train and test set, respectively; Right axis: binary cross-entropy loss across training epochs (light red and dark red line for train and test set, respectively B) Receiver Operating Characteristic curve i.e. false-positive versus true positive rate; area under the curve= 0.985.

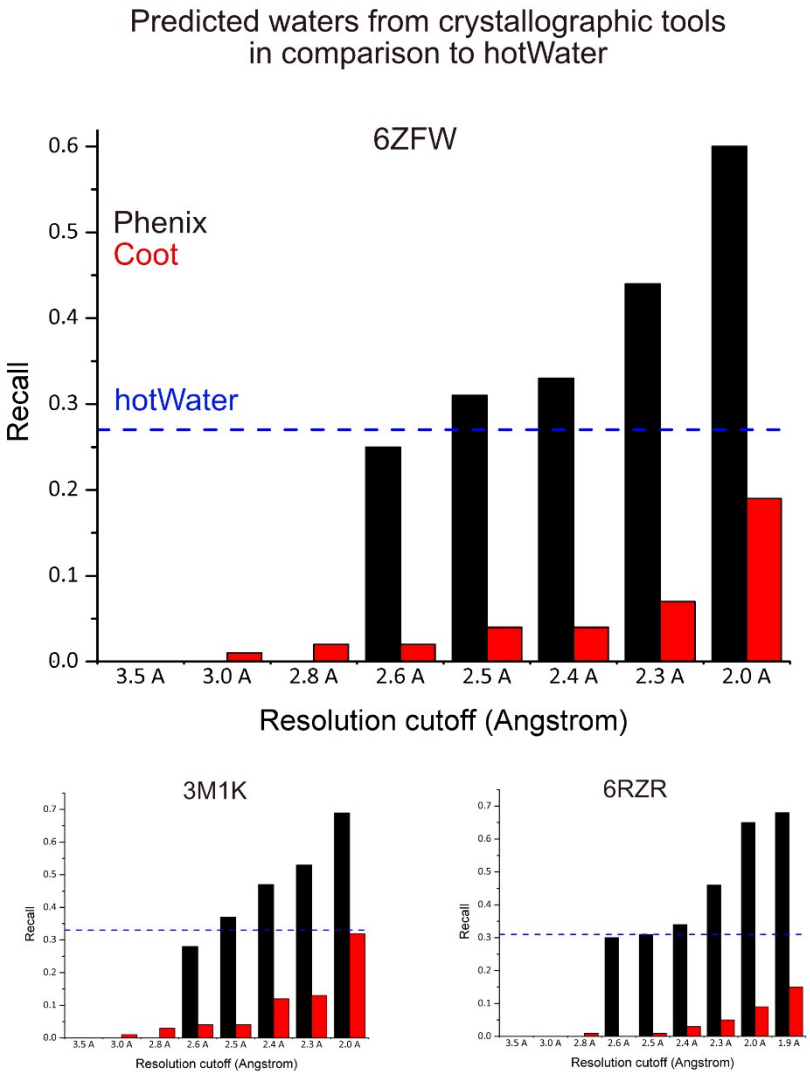

Figure SI4: Water prediction for X-ray crystallography

Case studies of the three of the six test proteins, for which raw crystallography data was provided, depicting the recall of Phenix Update Waters and the Coot Find Water functions at a variety of resolutions, in comparison to that of the HotWater algorithm.

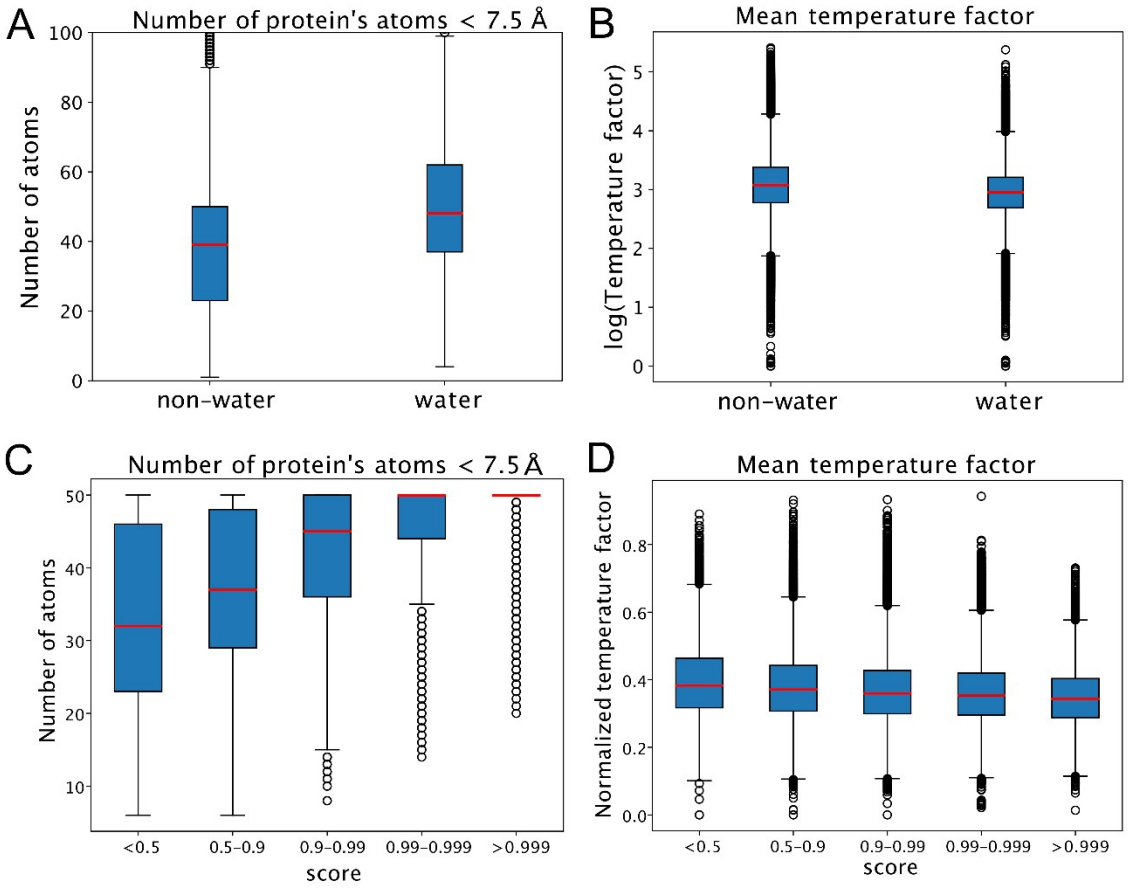

Figure SI5: Comparison of positive and negative input data and outputs

Top: Features describing characteristics of positive (water molecules within PDB files) and negative (heuristically chosen non-water positions; see SI 1.1) class samples; Bottom: Features describing the characteristics of water molecules according to the assigned hot-spot prediction scores: A, C) Number of interacting protein's atoms (<7.5 Å from the oxygen atom of the water); B,D) mean temperature factor of all protein atoms in the neighbourhood of the position (<7.5 Å from the water molecule).