DeepVASP-E: A Flexible Analysis of Electrostatic Isopotentials for Finding and Explaining Mechanisms that Control Binding Specificity

Felix M. Quintana*, Zhaoming Kong, Lifang He, and Brian Y. Chen*†

Dept. Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015, USA

Amino acids that play a role in binding specificity can be identified with many methods, but few techniques identify the biochemical mechanisms by which they act. To address a part of this problem, we present DeepVASP-E, an algorithm that can suggest electrostatic mechanisms that influence specificity. DeepVASP-E uses convolutional neural networks to classify an electrostatic representation of ligand binding sites into specificity categories. It also uses class activation mapping to identify regions of electrostatic potential that are salient for classification. We hypothesize that electrostatic regions that are salient for classification are also likely to play a biochemical role in achieving specificity. Our findings, on two families of proteins with electrostatic influences on specificity, suggest that large salient regions can identify amino acids that have an electrostatic role in binding, and that DeepVASP-E is an effective classifier of ligand binding sites.

Keywords: Specificity Annotation; Interpretable Binding Mechanisms; Volumetric Analysis.

1. Introduction

A small minority of amino acids play central roles in selective binding. Discovering those amino acids, and especially the biochemical mechanisms by which they act, is crucial for understanding how genetic variations influence pathogenicity and for interpreting how preferred binding partners might be changed through protein redesign. Most approaches proposed to date have focused on identifying influential amino acids: Evolutionary techniques[1,2] infer that the conservation of amino acids, or variations that follow major evolutionary divergences, are evidence for a role in function. Cavity based techniques[3–5] infer that proximity to the largest clefts on the solvent accessible surface is evidence for an enriched role in function. Structure comparison algorithms[6–8] infer that having certain atoms or amino acids in specific geometric configurations is evidence for the capacity to catalyze the same chemical reaction. Combinations of these and other concepts have also been considered.[9,10] All these methods can focus human attention on amino acids that may have a functional role, thereby reducing effort wasted on irrelevant amino acids. However, without deducing the biochemical mechanism by which these amino acids contribute to selective binding, the problem of determining the biochemical effect of genetic variation, and the design of validation experiments, must still rely on human expertise. Given that mutations at only a few critical residues have combinatorial effects on function, a computer generated explanation of specificity mechanisms could offer insights at appropriate scale and suggest mechanisms that experts might overlook.

* These authors contributed equally to this work. †Correspondence: chen@cse.lehigh.edu

Towards interpreting the biochemical role of individual residues, this paper proposes a novel general strategy that we call the *Analytic Ensemble* approach, and it examines one aspect of this strategy. We begin with a *training family* of closely related proteins that perform the same biochemical function, with subfamilies that prefer to act on similar but non-identical ligands. These families could be evolutionary in origin, or closer groups of mutants that share binding preferences. Suppose that we design an algorithm that examines only patterns of steric hindrance to identify amino acids that influence specificity. Then any amino acid identified by this narrow approach can be associated with a steric influence on specificity, because the method examines no other mechanism. Imagine a second approach that examines only electrostatic fields to find influential amino acids. Residues found by the second approach must have an electrostatic influence on specificity for the same reasons. Further methods could be developed for hydrogen bonds, hydrophobicity, and so on. While these individual approaches are very narrow, they could collectively analyze a diverse range of structural mechanisms, and their exclusivity has the novel property that it connects their findings to a biochemical mechanism. The same inferential structure is typically not possible with existing approaches because most employ biochemically holistic representations. For example, finding the same residues in the same locations with a structure comparison algorithm could identify amino acids that might be ideally positioned to either form hydrogen bonds, or to sterically hinder a discouraged ligand. In such cases, important residues can be detected, but activity through steric hindrance, for example, cannot be confirmed.

As a part of the Analytic Ensemble, this paper presents DeepVASP-E, an algorithm for identifying electrostatic mechanisms by which amino acids influence specificity. It achieves this purpose by representing the electric field within protein binding sites using a voxel representation of electrostatic isopotentials. Given isopotentials q derived from the binding site of a query protein with unknown binding preferences, DeepVASP-E performs two functions: First, it adapts a three dimensional convolutional neural network (3D-CNN) to classify q into one of the subfamilies based exclusively on the geometric similarity of electrostatic isopotentials. Second, it uses the gradient-weighted class activation mapping, Grad-CAM++[11] to identify regions of q that motivate its classification into that subfamily. We hypothesize that regions identified in this way will identify zones of electrostatic potential that are important for selective binding, thereby proposing a simple electrostatic mechanism by which the query protein achieves specificity.

Deep learning methods have recently made increasing contributions to structural bioinformatics, most notably for the prediction of protein structures.[12–15] DeepVASP-E has some similarities to these methods in its underlying techniques. For example, 3D-CNNs have been applied for the prediction of binding sites[16,17] and the classification of proteins by Enzyme Classification number.[18] Grad-CAM maps have also been applied for the identification of functional amino acids.[19] While it also employs deep learning techniques, DeepVASP-E is using deep learning for a new purpose, to generating biochemical explanations for specificity.

Our attention to electrostatic fields in DeepVASP-E is inspired by findings with VASP-E, a volumetric algorithm for comparing and analyzing electrostatic isopotentials in ligand binding sites and protein-protein interfaces.[20,21] VASP-E uses Constructive Solid Geometry (CSG),

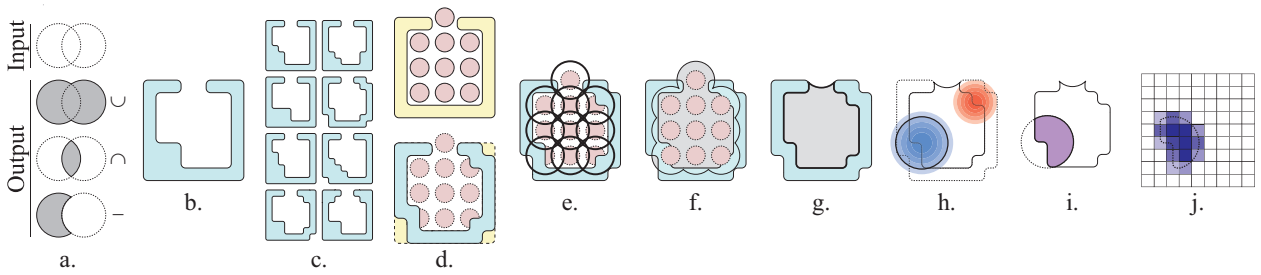

Fig. 1. **Representing the Electrostatic Properties of a Binding Cavity a.** CSG union (∪), intersection (∩), and difference operations (−) performed on input solids (white) and their outputs (grey). **b.** A protein illustrated as a molecular surface (teal). **c.** Conformational samples of the protein. **d.** The pivot structure t' (top, yellow) with ligand atoms (red), and the structural alignment (bottom) of one conformational sample (teal) to t'. **e.** Spheres defining the neighborhood of the ligand atoms (black circles). **f.** The union of the ligand atom spheres (black outline) minus the molecular surface of the conformational sample (teal), is shown in grey. **g.** The binding site in the conformational sample (grey). **h.** The positive (blue gradient) and negative (red gradient) regions of the electrostatic field, shown with the selected electrostatic isopotential (black circle). **i.** The cavity field: the CSG intersection between the electrostatic isopotential and the cavity region (purple). **j.** Translation of the cavity field (dotted outline) into the weighted voxel representation for 3D-CNN training (shaded purple squares). Lorem ipsum dolor sit amet, consectetur adipiscing elit.

---

a technique for computing three dimensional (3D) unions, intersections and differences, to represent and compare electrostatic isopotentials within binding sites as geometric solids (Fig. 1). We found that these comparisons could categorize proteins that prefer different ligands and predict amino acids that influence specificity.[22] This paper builds on this technique by introducing a deep learning approach that can be interpreted to identify the regions of the electrostatic field that are salient to classification.

To perform classification and saliency analysis accurately, it is vital to incorporate conformational variations in the training set. We focus here on cases of limited flexibility, where small sidechain motions and backbone breathing can alter the apparent distribution of charges in the binding cavity and potentially interfere with classification. Proteins with disordered or multiple conformations are outside the scope of this approach. For this work, DeepVASP-E uses conformational samples of the training family from medium-timescale molecular dynamics simulations to train the neural network. Our integration with simulated data has two novel and synergistic advantages: First, it provides a source of highly authentic conformational samples for training that can assist in compensating for conformational change. Second, simulation data are a source of authentic structures that can augment structural datasets so they can satisfy the ravenous need for training data in deep learning systems.

Our results examine the performance of DeepVASP-E on two sequentially nonredundant families of proteins with experimentally established binding preferences. We measured its classification accuracy on electrostatic representations of binding cavities and compared its classification performance to existing techniques. We then validated the accuracy of the saliency maps against experimentally established specificity mechanisms. Together, these results point to new applications in automatically explaining specificity mechanisms in molecular structure.

2. Methods

Overview DeepVASP-E uses a training family T constructed from a family of proteins with well defined subfamilies $\{T_0, T_1, \cdots, T_n\}$ that exhibit distinct binding preferences and a known ligand binding site. To prepare this data, as summarized in Fig. 1, the structure of each protein in T is first simulated using molecular dynamics to produce conformational samples (Fig. 1c). Second, each conformational sample, t, is structurally aligned to a *pivot structure* t', which is a member of T that was chosen because it has a ligand crystallized in the binding site (Fig. 1d). We use the ligand atoms, now aligned to the binding site of t to define the binding site (Fig. 1e-g). Third, we determine the electrostatic field, and given an isopotential threshold, we compute an electrostatic isopotential of t (Fig. 1h). Using CSG, regions of the isopotential that are outside the binding site are removed, producing a region we call the *cavity field* (Fig. 1i). Finally, the cavity field is translated into voxel data, as input for the 3D-CNN (Fig. 1j).

When classifying the structure of a novel query protein q into one of the n subfamilies, we treat it as a single conformation of the novel protein. Thus, to prepare q for classification, we begin with the structural alignment of q to t' and follow the data preparation steps above until voxel data is generated. Using a 3D-CNN model that has been trained, we produce classifications of q into one of the n subfamilies (Fig. 2).

To predict the voxels that are electrostatically significant for selective binding, we adapt the gradient-weighted class activation map method Grad-CAM++ to identify the regions most associated with classification.[11] The resulting voxels define binding cavity regions that are significant for classification. We hypothesize that these differentiating potentials are nearby amino acids that are crucial for binding, which we will verify against experimentally established results in Section 3.

Training Families To test DeepVASP-E, we selected the serine protease and enolase superfamilies as training families (Table 1). Within each training family, we selected three distinct subfamilies with different binding preferences. From the serine proteases, we selected the trypsin, chymotrypsin, and elastase subfamilies, and we selected the enolase, mandelate racemase, and muconate lactonizing enzyme families from the enolase superfamily.

Table 1. PDB codes of training families used in this study.

Serine Protease Superfamily		**Enolase Superfamily**	
Trypsins	1a0j, 1ane, 1aq7, 1bzx, 1fn8, 1h4w, 1trn, 2eek, 2f91	Enolases	1iyx, 1te6, 3otr
Chymotrypsins	1ex3	Mandelate Racemases	1mdr, 2ox4
Elastases	1b0e, 1elt	Muconate Lactonizing Enzyme	2pgw

The serine proteases selectively cleave peptide bonds by recognizing amino acids on both sides of the scissile bond. The P1 residue, which is immediately before the bond, is recognized by the S1 specificity pocket. In Chymotrypsins, P1 is preferred to be large and hydrophobic.[23] Trypsins prefer P1 residues that are positively charged, to complement their negatively charged S1 pocket.[24] Elastases recognize a small hydrophobic P1 residues.[25]

The enolase superfamily share a binding site at the center of a TIM-barrel fold with an N-terminal "capping domain". Members of this superfamily achieve a range of different functions that generally abstract a proton from a carbon adjacent to a carboxylic acid:[26,27] Enolases catalyze the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate,[28] mandelate race-

mases convert (R)-mandelate to and from (S)-mandelate,[29] and muconate lactonizing enzymes catalyze the reciprocal cycloisomerization of cis,cis-muconate and muconolactone.[26]

The structures in our training families were selected from the Protein Data Bank[30] (PDB) on 6.21.2011. Using the Enzyme Commission classifications (EC number), we found 676 serine protease and 66 enolase structures among the families selected for our data set. From this group, proteins with mutations, disordered regions, or enolases in closed or partially closed (and thus inactive) conformations were removed. From the remaining set, a set of sequentially nonredundant representatives were selected such that no representative had greater than 90% sequence identity with any other representative. This filtration resulted in an average sequence identity of 54.7% and 24.7% among the serine proteases and the enolases, respectively. Technical problems with molecular dynamics simulation prevented 8gch, 1aks, and 2zad from being included in this set. From each of the remaining structures we removed waters, ions, hydrogens, and other non-protein atoms.

Conformational Sampling To produce conformational samples of each protein, we used GROMACS 4.5.4.[31] In preparation, each structure was centered in a cubic waterbox with 10 Å to the nearest point on the box. Inside, solvent was populated using an equilibrated 3-site SPC/E solvent model.[32] Charge balanced sodium and potassium ions were added at a low concentration ($< 0.1\%$ salinity).

Next, we performed Isothermal-Isobaric (NPT) equilibration of this system in four 250 picosecond timesteps, using a steepest descent algorithm. Starting with a position restraint force of 1000 kJ/(mol*nm), each step reduced the restraint by 250 kJ/(mol*nm). System energies were generated at the start of this equilibration, with initial temperature set at 300K and initial pressure at 1 bar. The Nosé-Hoover thermostat[33] was used for temperature coupling. P-LINCS[34] was used to update bonds. Electrostatic interaction energies were calculated by particle mesh Ewald summation (PME).[31] The Parrinello-Rahman algorithm was used for pressure coupling.[35] Temperature and pressure scaling were performed isotropically. The atomic positions and velocities of the final equilibration step were used to start the primary simulation, with all position restraints removed.

The primary simulation was sustained for 100 nanoseconds in 1 femtosecond timesteps. P-LINCS and PME were chosen for their parallel efficiency, and OpenMPI was used for interprocess and network communication. Simulations were run on multiple nodes with 16 cores each, with PME distribution automatically selected by GROMACS. The trajectory file of each completed simulation was then converted into individual timesteps in the PDB file format, with waterbox atoms removed. From these timesteps, we selected 600 conformational samples at uniform intervals, and used them to train DeepVASP-E.

Structural Alignment and Binding Site Representation After conformational sampling of every protein in both training families, each sample was aligned to the pivot structure using ska.[36] Among the serine proteases, the pivot was bovine chymotrypsin (pdb: 8gch), and for enolases, the pivot was pseudomonas putida mandelate racemase (pdb: 1mdr). Pivot proteins were selected because they are co-crystallized with a ligand, which is used to localize the binding site in all conformational samples (Fig. 1e). Due to technical errors in MD simulation, 8gch was used only for this localization step.

After alignment, we apply a technique from VASP, described earlier,[37] to produce a solid representation of the binding site in the conformational sample. Paraphrasing here, we begin by producing spheres centered on the atoms of the ligand, with radius 5.0 Å (Fig. 1e). The CSG union of the spheres (Fig. 1f) defines the neighborhood of the ligand, U. We also compute a molecular surface S and envelope surface E of the conformational sample (not shown in Fig. 1f for clarity) using the classic rolling probe technique.[38] Here, S and E are produced with probes with radius 1.4 Å and 5.0 Å respectively. E represents the region inside the protein, including the cavity. Using CSG, we compute $(U - S) \cap E$ to produce the cavity (Fig. 1g).

Computing Cavity Fields Beginning with each conformational sample, we first model all hydrogen atoms using the "reduce" component of MolProbity.[39] We then use DelPhi[40] to solve the Poisson-Boltzmann equation, producing the electrostatic potential field nearby (Fig. 1h). Finally, using isopotential thresholds -1.0 kt/e and 1.0 kt/e, we use VASP-E to compute electrostatic isopotentials from this field, representing each as a geometric solid. These thresholds were selected because in past experiments we considered a range of thresholds these values produced the clearest outputs.[20–22] Higher absolute values create smaller isopotentials with less detail, while lower absolute values can be too large. Finally, we compute the CSG intersection between each isopotential and the cavity region defined above to produce a positive and a negative cavity field. Only positive cavity field is shown in Fig. 1i for clarity. Henceforth we perform separate computations on positive and negative cavity fields, enabling positive and negative charge to separately generate explanations for influencing binding specificity.

Voxelized Binding Site Representations Each cavity field is then translated into a voxel representation for 3D-CNN classification (Fig. 1j). First, we produce a bounding box around all positive or all negative cavity fields from all proteins in the training family. We then divide the bounding box into voxel cubes that are 0.5 Å on a side, padding it slightly to ensure an integer number of cubes in all dimensions. Finally, we use CSG intersections with cubes to estimate the volume in cubic angstroms of the cavity field inside each voxel. A tensor of voxel intersection volumes is then passed into the 3D-CNN for training or classification.

2.1. *Convolutional Neural Network*

Our 3D-CNN architecture (Fig. 2) accepts voxel data as input. The architecture of DeepVASP-E is inspired by LeNet-5, a classic CNN architecture for recognizing handwritten and printed characters,[41] and VoxNet, a 3D-CNN method for recognizing 3D point clouds.[42] The chief design constraint for the CNN component of DeepVASP-E is the three dimensional resolution of the cavity field, which ranges between 30 to 42 cubes in each dimension, leading to a large number of neurons per layer that must be trained relative to typical 2D image analysis methods. Unfortunately, we have observed that reducing the number of neurons by using a coarser resolution can interfere with classification accuracy.[37] For this reason, we support the full resolution of the input cavity fields and a shallower topology similar to these classic methods. The approach concludes with a fully connected layer with a softmax activation function to three categories corresponding to the three subfamilies of our training familes.

Class Specific Saliency Mapping Gradient-weighted^{++} class activation mapping (Grad-CAM++)[11] was used to generate saliency maps for all of the proteins and respective classes, which is a popular method for explaining CNN predictions. It uses the gradient information

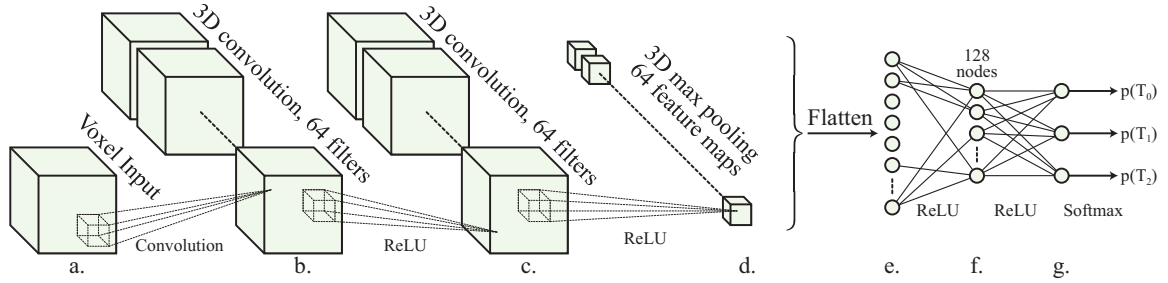

Fig. 2. **a.** Real valued voxel inputs with each of the three dimensions ranging from 30 - 42 cubes. **b, c.** 3D convolutional layers with 64 filters. $5 \times 5 \times 5$ kernels with stride 1 were used (dotted lines), with padding to maintain resolution. A ReLU activation function is used in both layers. **d.** 3D max pooling layer with pool size $2 \times 2 \times 2$ and stride (2,2,2), producing outputs of size $(x/2, y/2, z/2)$ with ReLU activation function. **e.** Flattening layer with ReLU activation function. **f** Fully connected layer reducing layer e to 128 nodes, ReLU activation layer. **g.** output layer with 3 nodes corresponding to the number of classes with softmax activation function, to produce probabilities of classification.

flowing into the last convolutional layer of the CNN to extract the importance of each neuron for the model output. Given a voxel's spatial location (x, y, z) for a particular class c, we use Grad-CAM++ to generate a class-specific saliency map L^c as:

$$L^c_{x,y,z} = \sum_k w^c_k A^k_{x,y,z} \tag{1}$$

where $A^k_{x,y,z}$ is the kth feature map in the last convolutional layer of 3D-CNN, and w^c_k is the corresponding weight defined as follows:

$$w^c_k = \sum_x \sum_y \sum_z [\frac{\frac{\partial^2 Y^c}{(\partial A^k_{x,y,z})^2}}{2\frac{\partial^2 Y^c}{(\partial A^k_{x,y,z})^2} + \sum_p \sum_q \sum_r A_{p,q,r} \frac{\partial^3 Y^c}{(\partial A^k_{x,y,z})^3}}] ReLu(\frac{\partial Y^c}{\partial A^k_{x,y,z}}) \tag{2}$$

where $Y^c = \exp(S^c)$ is the class score, and S^c is the penultimate layer score of class c (Fig. 2g). $\frac{\partial Y^c}{\partial A^k_{x,y,z}}$, $\frac{\partial^2 Y^c}{(\partial A^k_{x,y,z})^2}$, and $\frac{\partial^3 Y^c}{(\partial A^k_{x,y,z})^3}$ are the first-, second-, and third-order gradients w.r.t. $A^k_{x,y,z}$.

2.2. *Experimental Design*

Each protein in the training family contributes 600 conformational samples to the data set. To train the 3D-CNN model, we first leave all samples of one protein, the *evaluation set*, out of the dataset. Second, from the remaining samples, we randomly select 20% to create a *test set* to measure model classification performance. Next, the remaining snapshots are divided randomly into a *validation* set and *training* set at a 1:4 ratio. Weights on each node of the 3D-CNN are assigned and validated with the training and validation sets. This process repeats in each epoch until accuracy on the validation set stabilizes. We perform this process five times with a new, distinct, randomly selected test set. Finally, the weights of the model with the highest accuracy of all five folds are used to predict the subfamily category of the samples in the evaluation set. This evaluation is repeated, iteratively leaving out each protein in the training family.

In this design, the evaluation set is separated to ensure that data used for model refinement never leaks into the performance evaluation. The separation and random selection of the test set supports model evaluation in the presence of multiple categories.

Table 2. Comparison of classification results (avg ± std).

Training Family	Metric	PCA	DeepVasp-E
Serine Proteases Positive Isopotential	Accuracy	97.00 ± 5.43	$\mathbf{98.58 \pm 1.88}$
	F1	98.42 ± 2.90	$\mathbf{99.33 \pm 0.89}$
Serine Proteases Negative Isopotential	Accuracy	$\mathbf{100.0 \pm 0.00}$	98.34 ± 0.61
	F1	$\mathbf{100.0 \pm 0.00}$	98.42 ± 0.51
Enolase Superfamily Positive Isopotential	Accuracy	98.67 ± 1.97	$\mathbf{99.53 \pm 0.06}$
	F1	99.17 ± 1.17	$\mathbf{99.47 \pm 0.10}$
Enolase Superfamily Negative Isopotential	Accuracy	97.83 ± 4.83	$\mathbf{99.86 \pm 0.015}$
	F1	97.83 ± 5.31	$\mathbf{100.00 \pm 0.00}$

2.3. *Comparison with Existing Methods*

While no other methods currently predict biochemical mechanisms that affect specificity, we compared the classification accuracy of DeepVASP-E to that of classic principal component analysis (PCA).[43] For PCA, we learn a low-dimensional feature embedding of input data, and the embedding dimension is selected from $\{5, 10, 15, \cdots, 100\}$ using the same cross-validation strategy described in Section 2.2, then a logistic regression model is applied as a classifier.

Implementation Details and Availability The deep learning backend is tensorflow-gpu (2.4.1) with Python 3.8. All experiments was performed on a workstation with 16 cores and 32GB main memory, using an Nvidia RTX 3090 GPU with 24GB of VRAM memory. 5-fold training for a single evaluation protein required approximately ten minutes.

3. Results

Specificity Classification We evaluated the performance of DeepVASP-E for predicting the specificity category of each member of both training families. The average accuracy and F1-score of the compared methods are presented in Table 2.

Overall, DeepVASP-E clearly outperformed PCA on three out of four datasets and performed slightly worse on the fourth dataset. In addition, DeepVASP-E exhibited substantially less variability in performance: Across the five folds of our validation experiment, fluctuations in accuracy and F1-score for DeepVASP-E had standard deviations less than those of PCA. These findings point to lower consistency in the classification performance of PCA relative to DeepVASP-E. Compared to PCA, another benefit of the proposed DeepVASP-E model is that it maintains the adjacency structure of the voxel data, rather than vectorizing it, to support explainability. We exploit this advantage using Grad-CAM++ below.

Mechanism Prediction We hypothesize that the most salient voxels identified by DeepVASP-E will be regions of electrostatic isopotential that are mechanistically involved in the specificity of the protein. Thus, we also evaluated the accuracy of the most salient voxels as predictions of functionally significant charged regions, and verified them against experimentally established findings in the literature, cited throughout.

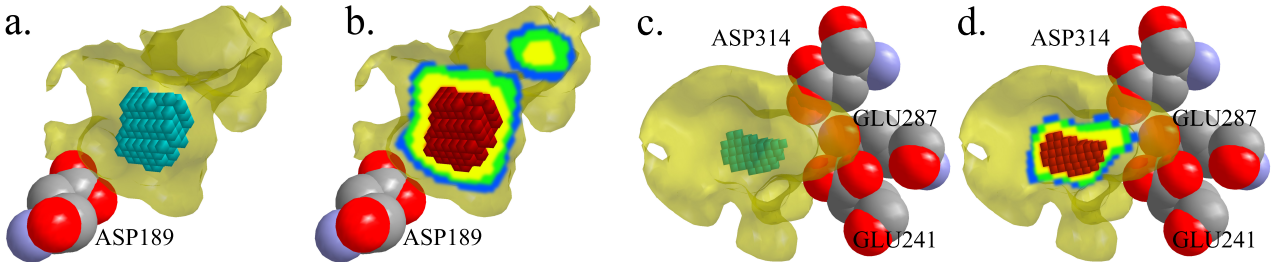

Fig. 3. **Negatively charged salient regions of the trypsin and enolase binding cavities.** The binding cavities of atlantic salmon trypsin (pdb: 1a0j, panels **a**,**b**) and enterococcus hirae enolase (pdb: 1iyx, panels **c**, **d**), are shown in transparent yellow. The most salient 150 voxels identified by DeepVASP-E are shown as teal cubes in **a** and **c**. In **b** and **d**, the gradient of red, yellow, green and blue cubes illustrates four groups of 150 cubes with decreasing salience. These electrostatic regions are created by the nearby amino acid D189, in trypsin, and E241, E287 and D314, in enolase.

Overall, we observed that the most salient voxels appear nearby charged amino acids that are known to affect specificity. The clearest example of this observation is the case of aspartate 189 in atlantic salmon trypsin. This amino acid has been experimentally established to play a pivotal role in the selection of positively charged substrates through electrostatic complementarity.[24] Looking at the geometry of the negative isopotential, the most salient region for classification is a region deep in the cavity nearby aspartate 189 (Fig. 3a). It is clear that the most salient voxels in the negative isopotential are identifying a region that enables specificity in trypsin. As we look at additional voxels with diminishing salience, we can see that they emanate away from this influential region (Fig. 3b). It is clear that this region plays a crucial role in distinguishing the subfamilies of the serine proteases, and that saliency mapping is able to detect electrostatic influences on specificity.

Small regions of positive charge appear to distinguish the other subfamilies of the serine proteases. While these isopotentials are not large, they appear to support the separation of elastase from chymotrypsin. Small salient regions were often observed around the nitrogen atoms of the V216, G193 and N192 that are near the binding site. Valine 216 is known to have a steric role in elastases for excluding larger hydrophobic substrates, but not an electrostatic one.[25] These findings illustrate that small regions of electrostatic isopotential are not random, and that they may distinguish between proteins in different subfamilies without themselves having an electrostatic role in specificity.

Among the enolase subfamily of the enolase superfamily, negatively charged isopotentials produced many salient voxels nearby E287, D241, and D314, which are known for stabilizing the magnesium ions necessary for dehydrating the enolase ligand[44] (Fig. 3c, d). Similar salient voxels were observed near corresponding amino acids of the other members of the enolase subfamily. Among the mandelate racemaces, negatively charged isopotentials produced salient voxels nearby E247 and E317, which have a role as a general acid catalyst.[45]

Positively charged isopotentials in the enolase subfamily were associated with regions of salient voxels nearby K339 and K390 in pdb 1IYX. The same amino acids in the other members of the enolase subfamily, with slightly different indices, also produced regions of salient voxels. Altogether, K339 and K390 are believed to electrostatically stabilize the carboxylate moiety of

the enolase substrate.[46] Among the mandelate racemases, salient voxels were associated with H297, K166, and K164 in pdb 1MDR. H297 and K166 are associated with proton exchange as a result of their net charge, and K164 is believed to electrostatically stabilize the carboxylate oxygen of the substrate.[45]

4. Conclusions

We have presented DeepVASP-E, a deep learning algorithm for detecting salient features for the classification of electrostatic isopotentials within ligand binding cavities. DeepVASP-E is the first algorithm to contribute to an Analytic Ensemble strategy, by which it is an intentionally narrow predictor of only electrostatic influences on specificity. When combined with other mechanism-specific predictors, we hypothesize that a more comprehensive picture of the multiple mechanisms governing molecular recognition can emerge.

In our results, large salient regions identified with DeepVASP-E occupied regions of electrostatic potential that are significant for achieving binding specificity. Charged amino acids adjacent to these regions are often associated with an electrostatic role in binding specificity, according to established experimental results. We also observed that small salient regions may identify regions that assist in classification but do not contribute an electrostatic role in specificity. These findings support the use of salient regions to identify electrostatic mechanisms that influence specificity, especially if smaller salient regions can be filtered out.

Altogether, these capabilities point to applications in anticipating mutations that alter binding preferences or novel mutations that maintain similar binding preferences. These include forecasting mutations that may arise in viral evolution, leading to vaccine resistance, or in protein redesign, for altering binding specificity.

Acknowledgements The authors are grateful to Dr. Edward Kim for his generous advice on interpretable machine learning methods and to Mr. Desai Xie for his early work on the project. This work was funded in part by NIH Grant R01GM123131 to Brian Y. Chen.

References

1. O. Lichtarge, H. R. Bourne and F. E. Cohen, An evolutionary trace method defines binding surfaces common to protein families, *Journal of molecular biology* **257**, 342 (1996).
2. A. Armon, D. Graur and N. Ben-Tal, Consurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information, *Journal of molecular biology* **307**, 447 (2001).
3. M. Nayal and B. Honig, On the nature of cavities on protein surfaces: application to the identification of drug-binding sites, *Proteins: Structure, Function, and Bioinformatics* **63**, 892 (2006).
4. J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz and J. Liang, Castp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues, *Nucleic acids research* **34**, W116 (2006).
5. R. A. Laskowski, N. M. Luscombe, M. B. Swindells and J. M. Thornton, Protein clefts in molecular recognition and function., *Protein Science* **5**, p. 2438 (1996).
6. C. Guda, S. Lu, E. D. Scheeff, P. E. Bourne and I. N. Shindyalov, Ce-mc: a multiple protein structure alignment server, *Nucleic acids research* **32**, W100 (2004).
7. B. Y. Chen, V. Y. Fofanov, D. M. Kristensen, M. Kimmel, O. Lichtarge and L. E. Kavraki, Algorithms for structural comparison and statistical analysis of 3d protein motifs, in *Biocomputing 2005*, (World Scientific, 2005).

8. H. J. Wolfson and I. Rigoutsos, Geometric hashing: An overview, *IEEE computational science and engineering* **4**, 10 (1997).
9. B. Huang and M. Schroeder, Ligsite csc: predicting ligand binding sites using the connolly surface and degree of conservation, *BMC structural biology* **6**, 1 (2006).
10. B. Y. Chen, D. H. Bryant, V. Y. Fofanov, D. M. Kristensen, A. E. Cruess, M. Kimmel, O. Lichtarge and L. E. Kavraki, Cavity-aware motifs reduce false positives in protein function prediction, in *Computational Systems Bioinformatics*, 2006.
11. A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018.
12. W. Gao, S. P. Mahajan, J. Sulam and J. J. Gray, Deep learning in protein structural modeling and design, *Patterns* **1**, p. 100142 (2020).
13. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Highly accurate protein structure prediction with alphafold, *Nature* **596**, p. 583–589 (2021).
14. P. Hoseini, L. Zhao and A. Shehu, Generative deep learning for macromolecular structure and dynamics, *Current Opinion in Structural Biology* **67**, 170 (2021).
15. R. Pearce and Y. Zhang, Deep learning techniques have significantly impacted protein structure prediction and protein design, *Current Opinion in Structural Biology* **68**, 194 (2021).
16. J. Jiménez, S. Doerr, G. Martínez-Rosell, A. S. Rose and G. De Fabritiis, Deepsite: protein-binding site predictor using 3d-convolutional neural networks, *Bioinformatics* **33**, 3036 (2017).
17. M. Skalic, A. Varela-Rial, J. Jiménez, G. Martínez-Rosell and G. De Fabritiis, Ligvoxel: inpainting binding pockets using 3d-convolutional neural networks, *Bioinformatics* **35**, 243 (2019).
18. A. Amidi, S. Amidi, D. Vlachakis, V. Megalooikonomou, N. Paragios and E. I. Zacharaki, Enzynet: enzyme classification using 3d convolutional neural networks on spatial representation, *PeerJ* **6**, p. e4750 (2018).
19. V. Gligorijević, P. D. Renfrew, T. Kosciolek, J. K. Leman, D. Berenberg, T. Vatanen, C. Chandler, B. C. Taylor, I. M. Fisk, H. Vlamakis *et al.*, Structure-based protein function prediction using graph convolutional networks, *Nature communications* **12**, 1 (2021).
20. B. E. Nolan, E. Levenson and B. Y. Chen, Influential mutations in the smad4 trimer complex can be detected from disruptions of electrostatic complementarity, *J. Comput. Biol.* **24**, 68 (2017).
21. Y. Zhou, X.-P. Li, B. Y. Chen and N. E. Tumer, Ricin uses arginine 235 as an anchor residue to bind to p-proteins of the ribosomal stalk, *Scientific reports* **7**, 1 (2017).
22. B. Y. Chen, Vasp-e: Specificity annotation with a volumetric analysis of electrostatic isopotentials, *PLoS computational biology* **10**, p. e1003792 (2014).
23. K. Morihara and H. Tsuzuki, Comparison of the specificities of various serine proteinases from microorganisms, *Archives of biochemistry and biophysics* **129**, 620 (1969).
24. L. Gráf, A. Jancso, L. Szilágyi, G. Hegyi, K. Pintér, G. Náray-Szabó, J. Hepp, K. Medzihradszky and W. J. Rutter, Electrostatic complementarity within the substrate-binding pocket of trypsin, *Proceedings of the National Academy of Sciences* **85**, 4961 (1988).
25. G. I. Berglund, A. O. Smalås, H. Outzen and N. P. Willassen, Purification and characterization of pancreatic elastase from north atlantic salmon (salmo salar)., *Mol. Marine Biol. Biotechnol.* **7**, 105 (1998).
26. P. C. Babbitt, M. S. Hasson, J. E. Wedekind, D. R. Palmer, W. C. Barrett, G. H. Reed, I. Rayment, D. Ringe, G. L. Kenyon and J. A. Gerlt, The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the α-protons of carboxylic acids, *Biochemistry* **35**, 16489 (1996).
27. J. A. Gerlt, P. C. Babbitt and I. Rayment, Divergent evolution in the enolase superfamily: the interplay of mechanism and specificity, *Archives of biochemistry and biophysics* **433**, 59 (2005).

28. K. Kühnel and B. F. Luisi, Crystal structure of the escherichia coli rna degradosome component enolase, *Journal of molecular biology* **313**, 583 (2001).
29. S. L. Schafer, W. C. Barrett, A. T. Kallarakal, B. Mitra, J. W. Kozarich, J. A. Gerlt, J. G. Clifton, G. A. Petsko and G. L. Kenyon, Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the d270n mutant, *Biochemistry* **35**, 5662 (1996).
30. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, The protein data bank, *Nucleic acids research* **28**, 235 (2000).
31. B. Hess, C. Kutzner, D. Van Der Spoel and E. Lindahl, Gromacs 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation, *Journal of chemical theory and computation* **4**, 435 (2008).
32. M. Iannuzzi, A. Laio and M. Parrinello, Efficient exploration of reactive potential energy surfaces using car-parrinello molecular dynamics, *Physical Review Letters* **90**, p. 238302 (2003).
33. S. Nosé, A unified formulation of the constant temperature molecular dynamics methods, *The Journal of chemical physics* **81**, 511 (1984).
34. B. Hess, P-lincs: A parallel linear constraint solver for molecular simulation, *Journal of chemical theory and computation* **4**, 116 (2008).
35. M. Parrinello and A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, *Journal of Applied physics* **52**, 7182 (1981).
36. A.-S. Yang and B. Honig, An integrated approach to the analysis and modeling of protein sequences and structures. i. protein structural alignment and a quantitative measure for protein structural distance, *Journal of molecular biology* **301**, 665 (2000).
37. B. Y. Chen and B. Honig, Vasp: a volumetric analysis of surface properties yields insights into protein-ligand binding specificity, *PLoS computational biology* **6**, p. e1000881 (2010).
38. A. Shrake and J. A. Rupley, Environment and exposure to solvent of protein atoms. lysozyme and insulin, *Journal of molecular biology* **79**, 351 (1973).
39. V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson, Molprobity: all-atom structure validation for macromolecular crystallography, *Acta Crystallographica Section D: Biological Crystallography* **66**, 12 (2010).
40. W. Rocchia, E. Alexov and B. Honig, Extending the applicability of the nonlinear poisson-boltzmann equation: multiple dielectric constants and multivalent ions, *The Journal of Physical Chemistry B* **105**, 6507 (2001).
41. Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86**, 2278 (1998).
42. D. Maturana and S. Scherer, Voxnet: A 3d convolutional neural network for real-time object recognition, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.
43. S. Wold, K. Esbensen and P. Geladi, Principal component analysis, *Chemometrics and intelligent laboratory systems* **2**, 37 (1987).
44. T. Hosaka, T. Meguro, I. Yamato and Y. Shirakihara, Crystal structure of enterococcus hirae enolase at 2.8 å resolution, *Journal of biochemistry* **133**, 817 (2003).
45. J. A. Landro, J. A. Gerlt, J. W. Kozarich, C. W. Koo, V. J. Shah, G. L. Kenyon, D. J. Neidhart, S. Fujita and G. A. Petsko, The role of lysine 166 in the mechanism of mandelate racemase from pseudomonas putida: Mechanistic and crystallographic evidence for stereospecific alkylation by (r)-. alpha.-phenylglycidate, *Biochemistry* **33**, 635 (1994).
46. J. Qin, G. Chai, J. M. Brewer, L. L. Lovelace and L. Lebioda, Fluoride inhibition of enolase: crystal structure and thermodynamics, *Biochemistry* **45**, 793 (2006).