We thank the reviewers for their detailed and thoughtful feedback. We respond to each reviewer individually below.

Reviewer 1

-Our model's computational requirements scale linearly with the size of the dataset. We have recently trained the model on much larger datasets using distributed data-level parallelism on a GPU cluster. Given a sufficient number of GPUs, it can be trained quickly on an arbitrarily large dataset.
-As suggested, we will revise the drawings in Fig. 2E to clarify that our representation is truly 3D, not a 2D projection.
-Regarding Table 3 labels: DB4 is a subset of DB5, and DB4 is the training set in Table 2 as well (hence certain numbers are the same). We will clarify this point, perhaps by dropping references to DB4 and simply referring to train/validation/test subsets of DB5.
-We only use CAUROC for the test set. This is computed per training run, and we show the mean and standard deviation of CAUROC across training seeds. So it is a mean (across training seeds) of the median (across complexes) of AUROC. We will clarify this in the Table 2 and Fig. 3 captions.
-Regarding the plot not reaching 0.9 in Figure 3: there are two factors at play here. First, we are plotting mean of CAUROC across 5 training seeds without showing the individual values. When we select the best one by validation loss, we can consistently select a higher CAUROC than the mean value. Second, our final reported results come from a dataset size of 163840, which increases performance over this plot (though we do not currently have experiments for all grid sizes at that dataset size). We will clarify this by plotting all the individual training seed CAUROCs as opposed to the standard deviation, and we will generate this plot for dataset size 163840.

Reviewer 2

-Machine learning applied to molecular structures (in drug design, materials science, and structural biology) is a rapidly growing field of interest to machine learning practitioners—in terms of techniques, datasets, and properties of atomic data. ImageNet has had a major impact in machine learning as an effective dataset against which to pretrain for vision tasks. Likewise, our work shows how one can achieve high performance on molecular learning tasks using large sets of related data and provides a dataset for doing so.
-The use of learned features for protein interface prediction is novel. [23] feeds high-level, hand-defined features into their graph convolution architecture.
-Regarding learned flexibility: We show that unlike previous methods, our method performs well on cases involving conformational change even when the training dataset includes no conformational change at all. The training datasets for previous methods included cases that exhibit substantial conformation change. We will clarify this point.
-Regarding formatting/citations: We will make the suggested changes.
-We use a 3D CNN because graph convolutions do not directly model complex three-dimensional geometric arrangements, and proteins are complex three-dimensional structures.
-We will clarify that Fig. 3B shows that our model—unlike competing methods—is capable of leveraging additional data to increase its performance and achieve state-of-the-art results. We will avoid the term "scalability" in this context.

Reviewer 3

-Regarding dataset availability: We have already uploaded a preliminary version of the dataset on Harvard Dataverse (the subset we trained on) and will upload the full dataset.
-Regarding pruning only against the test set and smoothing across voxels: These are good points and should only improve performance.
-Regarding analyzing why other methods fail to improve performance using the larger dataset: We agree this would be informative. For example, we can determine which high-level features don't translate well to DIPS.
-Regarding information on voxel size and number: Some of this is stated in the Methods section (voxels are 1Å in size and we use 35x35x35 voxels total per surfacelet), but we will add to and highlight this information.
-Our model could indeed be used to predict regions within the same protein that interact with one another, and we agree that this would result in significantly more training data. The key question is how to exclude parts of one surfacelet from the other so signal does not leak about each interacting component. We will point this out.
-We will better explain the experimental setup of Fig. 3A. Grid size refers to the entire window, and voxel size is kept constant. If the voxel size were allowed to vary, then the learned filters would have to encode information at different scales, which might tax the model's capacity.
-Our model scales linearly with the amount of data provided, and we have recently trained it on much larger datasets using distributed data-level parallelism on a GPU cluster. The model should indeed be useful in many molecular prediction tasks, as the reviewer notes.