Abstract

Computational methods that operate on three-dimensional molecular structure have the potential to solve important questions in biology and chemistry. In particular, deep neural networks have gained significant attention, but their widespread adoption in the biomolecular domain has been limited by a lack of either systematic performance benchmarks or a unified toolkit for interacting with molecular data. To address this, we present ATOM3D, a collection of both novel and existing benchmark datasets spanning several key classes of biomolecules. We implement several classes of three-dimensional molecular learning methods for each of these tasks and show that they consistently improve performance relative to methods based on one- and two-dimensional representations. The specific choice of architecture proves to be critical for performance, with three-dimensional convolutional networks excelling at tasks involving complex geometries, graph networks performing well on systems requiring detailed positional information, and the more recently developed equivariant networks showing significant promise. Our results indicate that many molecular problems stand to gain from three-dimensional molecular learning, and that there is potential for improvement on many tasks which remain underexplored. To lower the barrier to entry and facilitate further developments in the field, we also provide a comprehensive suite of tools for dataset processing, model training, and evaluation in our open-source atom3d Python package. All datasets are available for download from www.atom3d.ai.
Table 1: Representation choice for molecules. Adding in 3D information consistently improves performance. The depicted 1D representations are the amino acid sequence and SMILES [Weininger, 1988] for proteins and small molecules, respectively.

<table>
<thead>
<tr>
<th>Structure Level</th>
<th>Dimension</th>
<th>Representation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>primary</td>
<td>1D</td>
<td>linear sequence</td>
<td>KVKALPDA</td>
</tr>
<tr>
<td>secondary</td>
<td>2D</td>
<td>chemical bond graph</td>
<td>2D representation</td>
</tr>
<tr>
<td>tertiary</td>
<td>3D</td>
<td>atomistic geometry</td>
<td>3D representation</td>
</tr>
</tbody>
</table>

1 Introduction

A molecule’s three-dimensional (3D) shape is critical to understanding its physical mechanisms of action, and can be used to answer a number of questions relating to drug discovery, molecular design, and fundamental biology. While we can represent molecules using lower-dimensional representations such as linear sequences (1D) or chemical bond graphs (2D), considering the 3D positions of the component atoms—the atomistic geometry—allows for better modeling of 3D shape (Table 1). While previous benchmarking efforts have examined diverse molecular tasks, such as MoleculeNet [Wu et al., 2018] or TAPE [Rao et al., 2019], they focus on these lower-dimensional representations. In this work, we demonstrate the benefit yielded by learning on 3D atomistic geometry and promote the development of 3D molecular learning by providing a collection of datasets leveraging this representation.

Furthermore, the atom is emerging as a “machine learning datatype” in its own right, deserving focused study much like images in computer vision or text in natural language processing. All molecules, including proteins, small molecule compounds, and nucleic acids, can be represented as atoms in 3D space. These atoms can only belong to a fixed class of element types (e.g. carbon, nitrogen, oxygen), and are all governed by the same underlying laws of physics that impose rotational, translational, and permutational symmetries. These systems also contain higher-level patterns that are poorly characterized, creating a ripe opportunity for learning them from data: though certain basic components are well understood (e.g. amino acids, nucleotides, functional groups), many others can not easily be defined. These patterns are in turn composed in a hierarchy that itself is only partially elucidated.

While deep learning methods such as graph neural networks (GNNs) and convolutional neural networks (CNNs) seem especially well suited to atomistic geometry, to date there has been no systematic evaluation of such methods on molecular tasks. Additionally, despite the growing number of 3D structures available in databases such as the Protein Data Bank (PDB) [Berman et al., 2000], they require significant processing before they are useful for machine learning tasks. Inspired by the success of accessible databases such as ImageNet [Deng et al., 2009] and SQuAD [Rajpurkar et al., 2016] in sparking progress in their respective fields, we create and curate benchmark datasets for atomistic tasks, process them into a simple and standardized format, systematically benchmark 3D molecular learning methods, and present a set of best practices for other machine learning researchers interested in entering the field of 3D molecular learning (see Section D). We reveal a number of insights related to 3D molecular learning, including the consistent improvements yielded by using atomic geometry, the relative strengths and weaknesses of different methods, and the presence of several underexplored tasks that provide great opportunities for 3D molecular learning. These are all integrated into the atom3d Python package to lower the barrier to entry and facilitate reproducible research in 3D molecular learning for machine learning practitioners and structural biologists alike.
2 Related Work

While three dimensional molecular data have long been pursued as an attractive source of information in molecular learning and chemoinformatics [Swamidass et al., 2005, Azencott et al., 2007], their utility has become increasingly clear in the last couple years. Powered by increases in data availability and methodological advances, 3D molecular learning methods have demonstrated significant impact on tasks such as protein structure prediction [Senior et al., 2020], equilibrium state sampling [Noé et al., 2019], and drug design [Zhavoronkov et al., 2019]. At the same time, broader assessments of tasks involving such molecular data have focused on either 1D or 2D representations [Wu et al., 2018, Rao et al., 2019]. Through ATOM3D, we aim to provide a first benchmark for learning on 3D molecular data. There are a few major classes of algorithms that exist for data in this form.

Graph neural networks (GNNs) have grown to be a major area of study, providing a natural way of learning from data with complex spatial structure. Many GNN implementations have been motivated by applications to atomic systems, including molecular fingerprinting [Duvenaud et al., 2015], property prediction [Schütt et al., 2017, Gilmer et al., 2017, Liu et al., 2019], protein interface prediction [Fout et al., 2017], and protein design [Ingraham et al., 2019]. Instead of encoding points in Euclidean space, GNNs encode their pairwise connectivity, capturing a structured representation of atomic data.

Three-dimensional CNNs (3DCNNs) have also become popular as a way to capture these complex 3D geometries. They have been applied to a number of biomolecular applications such as protein model quality assessment [Pages et al., 2019, Derevyanko et al., 2018], protein sequence design [Anand et al., 2020], protein interface prediction [Townshend et al., 2019], and structure-based drug discovery [Wallauch et al., 2015, Iorg and Altman, 2017, Ragoza et al., 2017, Jiménez et al., 2018]. These 3DCNNs can encode translational and permutational symmetries, but incur significant computational expense and cannot capture rotational symmetries without data augmentation.

In an attempt to address many of the problems of representing atomic geometries, equivariant neural networks (ENNs) have emerged as a new class of methods for learning from molecular systems. These networks are built such that geometric transformations of their inputs lead to well-defined transformations of their outputs. This setup leads to the neurons of the network learning rules that resemble physical interactions. Tensor field networks [Thomas et al., 2018] and Cormorant [Kondor, 2018, Anderson et al., 2019] have applied these principles to atomic systems and begun to demonstrate promise on extended systems [Eismann et al., 2020, Weiler et al., 2018].

3 Datasets for 3D Molecular Learning

We select 3D molecular learning tasks from structural biophysics and medicinal chemistry that span a variety of molecule types. Multiple of these datasets are novel, while others are extracted from existing sources (Table 2). Below, we give a short description of each dataset’s impact and source, as well as the metrics used to evaluate them and the splits. The splits were selected to minimize data leakage concerns and ensure generalizability and reproducibility. Theses datasets are all provided in a standardized format that requires no specialized libraries. Alongside these datasets, we present corresponding best practices (Appendix D) and further dataset-specific details (Appendix E). Taken together, we hope these efforts will lower the barrier to entry for machine learning researchers interested in developing methods for 3D molecular learning and encourage rapid progress in the field.

3.1 Small Molecule Properties (SMP)

Impact – Predicting physico-chemical properties of small molecules is a common task in medicinal chemistry and materials design. Quantum chemical calculations can save expensive experiments but are themselves costly and cannot cover the huge chemical space spanned by candidate molecules.

Source – The QM9 dataset [Ruddigkeit et al., 2012, Ramakrishnan et al., 2014b] contains the results of quantum-chemical calculations for 134,000 stable small organic molecules made up of maximally nine atoms of C, O, N, and F. For each molecule, it contains the geometry of a molecule’s conformation in its ground state as well as calculated energetic, electronic, and thermodynamic properties.

Targets – We predict the molecular properties from the ground-state structure.

Split – We split molecules randomly.
Table 2: Tasks included in the ATOM3D datasets, along with schematic representations of their inputs. P indicates protein, SM indicates small molecule, R indicates RNA. Lines indicate interaction and the smaller square within proteins indicates an individual amino acid. New datasets are in bold.

<table>
<thead>
<tr>
<th>Name (Task Code)</th>
<th>Schematic</th>
<th>Objective</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Molecule Properties (SMP)</td>
<td>[SM]</td>
<td>Properties</td>
<td>QM9 [Ruddigkeit et al., 2012]</td>
</tr>
<tr>
<td>Protein Interface Prediction (PIP)</td>
<td>[P1], [P2]</td>
<td>Amino Acid Interaction</td>
<td>DIPS [Townshend et al., 2019]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DB5 [Vreven et al., 2015]</td>
</tr>
<tr>
<td>Residue Identity (RES)</td>
<td>[P]</td>
<td>Amino Acid Identity</td>
<td>New, created from PDB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[Berman et al., 2000]</td>
</tr>
<tr>
<td>Mutation Stability Prediction (MSP)</td>
<td>[P1], [P2] vs. [P1], [P2]</td>
<td>Effect of Mutation</td>
<td>New, created from SKEMPI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[Jankauskaitė et al., 2019]</td>
</tr>
<tr>
<td>Ligand Binding Affinity (LBA)</td>
<td>[P], [SM]</td>
<td>Binding Strength</td>
<td>PDBBind [Wang et al., 2004]</td>
</tr>
<tr>
<td>Ligand Efficacy Prediction (LEP)</td>
<td>[P], [SM] vs. [P], [SM]</td>
<td>Drug Efficacy</td>
<td>New, created from PDB</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[Berman et al., 2000]</td>
</tr>
<tr>
<td>Protein Structure Ranking (PSR)</td>
<td>[P]</td>
<td>Ranking</td>
<td>CASP-QA [Kryshtafovych et al., 2019]</td>
</tr>
<tr>
<td>RNA Structure Ranking (RSR)</td>
<td>[R]</td>
<td>Ranking</td>
<td>FARFAR2-Puzzles [Watkins and Das, 2019]</td>
</tr>
</tbody>
</table>

3.2 Protein Interface Prediction (PIP)

**Impact** – Proteins interact with each other in many scenarios—for example, antibody proteins recognize diseases by binding to antigens. A critical problem in understanding these interactions is to identify which amino acids of two given proteins will interact upon binding.

**Source** – For training, we use the Database of Interacting Protein Structures (DIPS), a comprehensive dataset of protein complexes mined from the PDB [Townshend et al., 2019]. We predict on the Docking Benchmark 5 [Vreven et al., 2015], a smaller gold standard dataset.

**Targets** – We predict if two amino acids will come into contact when their respective proteins bind.

**Split** – We split protein complexes such that whenever two structures have more than 30% sequence identity, they be in the same split dataset.

3.3 Residue Identity (RES)

**Impact** – Understanding the structural role of individual amino acids is important for engineering new proteins. We can understand this role by predicting the propensity for different amino acids at a given protein site based on the surrounding structural environment [Torng and Altman, 2017].

**Source** – We generate a novel dataset consisting of local atomic environments centered around individual residues extracted from non-redundant structures in the PDB.
** Targets – We formulate this as a classification task where we predict the identity of the amino acid in the center of the environment based on all other atoms.

** Split – We split environments by protein topology class according to the CATH 4.2 classification [Dawson et al., 2017], such that all environments from proteins containing each topology are in the same split.

3.4 Mutation Stability Prediction (MSP)

** Impact – Identifying mutations that stabilize a protein’s interactions is a key task in designing new proteins. Experimental techniques for probing these are labor-intensive [Antikainen and Martin, 2005; Lefèvre et al., 1997], motivating the development of efficient computational methods.

** Source – We derive a novel dataset by collecting single-point mutations from the SKEMPI database [Jankauskaite et al., 2019] and model each mutation into the structure to produce mutated structures.

** Targets – We formulate this as a binary classification task where we predict whether the stability of the complex increases as a result of the mutation.

** Split – We split protein complexes such that whenever two structures have more than 30% sequence identity, they be in the same split dataset.

3.5 Ligand Binding Affinity (LBA)

** Impact – Most therapeutic drugs and many molecules critical for biological signaling take the form of small molecules. Predicting the strength of the protein-small molecule interaction is a challenging but crucial task for drug discovery applications.

** Source – We use the PDBBind database [Wang et al., 2004; Liu et al., 2015], a curated database containing protein-ligand complexes from the PDB and their corresponding binding strengths.

** Targets – We predict \( pK = -\log(K) \), where \( K \) is the binding affinity in Molar units.

** Split – We split protein-ligand complexes such that whenever two of the proteins have more than 30% sequence identity, they be in the same split dataset.

3.6 Ligand Efficacy Prediction (LEP)

** Impact – Many proteins switch on or off their function by changing shape. Predicting which shape a drug will favor is thus an important task in drug design.

** Source – We develop a novel dataset by curating proteins from several families with both “active” and “inactive” state structures, and model in 527 small molecules with known activating or inactivating function using the program Glide [Friesner et al., 2004].

** Targets – We formulate this as a binary classification task where we predict if a molecule bound to the structures will be an activator of the protein’s function or not.

** Split – We split complex pairs by protein target.

3.7 Protein Structure Ranking (PSR)

** Impact – Proteins are one of the primary workhorses of the cell, and knowing their structure is often critical to understanding (and engineering) their function.

** Source – We use the submissions to the last 18 years of the Critical Assessment of Structure Prediction (CASP) [Kryshtafovych et al., 2019], a blind international competition for predicting protein structure.

** Targets – We formulate this as a regression task, where we predict the global distance test (GDT_TS) from the true structure for each of the predicted structures.

** Split – We split structures temporally by competition year.

3.8 RNA Structure Ranking (RSR)

** Impact – Similar to proteins, RNA plays major functional roles (e.g., gene regulation) and can adopt well-defined 3D shapes. Yet the problem is data-poor, with only a few hundred known structures.

** Source – We use candidate models generated by FARFAR2 [Watkins and Das, 2019] for the first 21 released RNA Puzzle challenges [Cruz et al., 2012], a blind structure prediction competition for RNA.
Targets – We predict the root-mean-squared deviation (RMSD) from the ground truth structure.

Split – We split structures temporally by competition year.

4 Benchmarking Setup

To assess the benefits of 3D molecular learning, we use a combination of existing and novel 3D molecular learning methods, and implement a number of robust baselines. Our 3D molecular learning methods belong to one of each of the major classes of deep learning algorithms that have been applied to atomistic systems: graph networks, three-dimensional convolutional networks, and equivariant networks. Here we describe the main principles of the core networks used in these models. See Appendix F for task-specific details and hyperparameters.

For GNNs, we represent molecular systems as graphs in which each node is an atom. Edges are defined between all atoms separated by less than 4.5 Å, and weighted by the distance between the atoms using an edge weight defined by \( w_{i,j} = \frac{1}{d_{i,j} + \epsilon} \), where \( \epsilon = 10^{-5} \) is a small factor added for numerical stability. Node features are one-hot-encoded by atom type. Our core model uses five layers of graph convolutions as defined by Kipf and Welling [2016], each followed by batch normalization and ReLU activation, and finally two fully-connected layers with dropout. For tasks requiring graph-level outputs, we use global mean pooling to aggregate over nodes. For tasks requiring predictions for single atoms or amino acids, we extract the relevant node embeddings from each graph after all convolutional layers (see Appendix F).

For 3DCNNs, we represent our data as a cube of fixed size (different per task due to the different molecular sizes) in 3D space that is discretized into voxels with resolution of 1 Å to form a grid (for PSR and RSR, we need to decrease the grid resolution to 1.3 Å in order to fit them in the GPU memory). Each voxel is associated with a one-hot-encoded vector which denotes the presence or absence of each atom type. Our core model consists of four 3D-convolutional layers, each followed by ReLU activation, max-pooling (for every other convolution layer), dropout, and two fully-connected layers.

For ENNs, we use SE(3)-equivariant networks that represent each atom of a structure by its position as absolute coordinates in 3D space with one-hot-encoded atom type as features. No rotational augmentation is needed due to the rotational symmetry of the network. The core of all architectures in this work is a network of four layers of covariant neurons that use the Clebsch–Gordan transform as nonlinearity, as described and implemented by Anderson et al. [2019].

5 Benchmarking Results

To assess the utility of 3D molecular learning, we evaluate our methods on the ATOM3D datasets and compare performance to state-of-the-art methods using 1D or 2D representations (for a comparison to the overall state-of-the-art, see Table 3). We note that in many cases, 3D molecular learning methods have not been applied to the proposed tasks, and that several of the tasks are novel. In the following sections, we describe the results of our benchmarking and some key insights that can be derived from them. We also aggregate these results along with additional metrics and standard deviations over three replicates in Table 8. For each metric, we bold the best-performing method as well as those within one standard deviation of the best-performing method.

5.1 3D representations consistently improve performance

Our evaluation of 3D methods on the tasks in ATOM3D reveals that incorporating atomistic geometry leads to consistently superior performance compared to 1D and 2D methods. For small molecules, state-of-the-art methods do not use 1D representations, so we focus instead on comparing to representations at the 2D level, i.e. the chemical bond graph. This is the approach taken by the 2D GNN introduced by [Tsubaki et al., 2019] or the N-gram graph method by [Liu et al., 2019], which both obtain similar results (Table 3) on the small-molecule-only dataset SMP. When we add 3D coordinate information as in our ENN implementation, performance improves across all targets in SMP.

For tasks involving biopolymers (proteins and RNA), state-of-the-art methods do not use 2D representations, primarily because most of the chemical bond graph can be re-derived from the 1D
Table 3: Small molecule results. Metric is mean absolute error (MAE).

<table>
<thead>
<tr>
<th>Task</th>
<th>Target</th>
<th>3D</th>
<th>Non-3D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3DCNN</td>
<td>GNN</td>
</tr>
<tr>
<td>µ [Å]</td>
<td>0.754</td>
<td>0.501</td>
<td><strong>0.052</strong></td>
</tr>
<tr>
<td>ε_{gap} [eV]</td>
<td>0.580</td>
<td>0.137</td>
<td><strong>0.095</strong></td>
</tr>
<tr>
<td>U_{q}^0 [eV]</td>
<td>3.862</td>
<td>1.424</td>
<td><strong>0.025</strong></td>
</tr>
</tbody>
</table>

Table 4: Biopolymer results. AUROC is the area under the receiver operating characteristic curve. Asterisks (*) indicate that the exact training data differed (though splitting criteria were the same).

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>3D</th>
<th>Non-3D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3DCNN</td>
<td>GNN</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIP</td>
<td>AUROC</td>
<td><strong>0.844</strong></td>
<td>*0.669</td>
</tr>
<tr>
<td>RES</td>
<td>accuracy</td>
<td>0.451</td>
<td>0.082</td>
</tr>
<tr>
<td>MSP</td>
<td>AUROC</td>
<td>0.574</td>
<td><strong>0.621</strong></td>
</tr>
</tbody>
</table>

representation, i.e. the linear sequence that makes up the biopolymer. We thus compare to representations at the 1D level (Table 4). For MSP and RES, both new datasets, we evaluate against the TAPE model [Rao et al. [2019]], a transformer architecture that operates on protein sequence and is state-of-the-art amongst 1D methods for many tasks. For PIP, we compare to the sequence-only version of BIPSPI [Sanchez-Garcia et al. [2018]], a state-of-the-art boosted decision tree method for protein interaction prediction. We find that 3D methods outperform these 1D methods on all biopolymer-only datasets (PIP, RES, MSP).

For tasks involving both biopolymers and small molecules, we compare to DeepDTA [Ozturk et al. [2018]]. This network uses a 1D representation via a 1DCNN for both the biopolymer and small molecules. For LBA, we additionally compare to DeepAffinity [Karimi et al. [2019]] which uses pairs of a ligand SMILES string and structurally-annotated protein sequences. Using a 3D representation for both the ligand and protein leads to improved performance for the joint protein-small molecule datasets (LBA and LEP, see Table 5).

The biopolymer structure ranking tasks (PSR and RSR) are inherently 3D in nature, as they involve evaluating the correctness of different 3D shapes taken on by the same biopolymer. Thus, critically, a 1D or 2D representation would not be able to differentiate between these different shapes since the linear sequence and chemical bond graph would remain the same. We therefore compare to state-of-the-art 3D methods as shown in Table 5 finding competitive or better results.

More generally, we find that learning methods that leverage the 3D geometry of molecules hold state-of-the-art on the majority tasks on our benchmark (Table 7).

Table 5: Joint small molecule/biopolymer results. $R_S$ is Spearman correlation, $R_P$ is Pearson correlation, AUROC is area under the receiver operating characteristic curve, and RMSE is root-mean-squared error.

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>3D</th>
<th>Non-3D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>3DCNN</td>
<td>GNN</td>
</tr>
<tr>
<td>LBA</td>
<td>RMSE</td>
<td><strong>1.416</strong></td>
<td>1.570</td>
</tr>
<tr>
<td>glob. $R_P$</td>
<td>0.550</td>
<td><strong>0.637</strong></td>
<td>0.389</td>
</tr>
<tr>
<td>glob. $R_S$</td>
<td>0.553</td>
<td><strong>0.628</strong></td>
<td>0.408</td>
</tr>
<tr>
<td>LEP</td>
<td>AUROC</td>
<td>0.589</td>
<td><strong>0.740</strong></td>
</tr>
</tbody>
</table>
Table 6: Structure ranking results. $R_S$ is Spearman correlation. Mean measures the correlation for structures corresponding to the same biopolymer, whereas global measures the correlation across all biopolymers.

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>3DCNN</th>
<th>GNN</th>
<th>SotA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSR</td>
<td>mean $R_S$</td>
<td>0.431</td>
<td>0.515</td>
<td>0.432 [Pagès et al., 2019]</td>
</tr>
<tr>
<td></td>
<td>glob. $R_S$</td>
<td><strong>0.789</strong></td>
<td>0.755</td>
<td><strong>0.796</strong> [Pagès et al., 2019]</td>
</tr>
<tr>
<td>RSR</td>
<td>mean $R_S$</td>
<td>0.264</td>
<td>0.234</td>
<td>0.173 [Alford et al., 2017]</td>
</tr>
<tr>
<td></td>
<td>glob. $R_S$</td>
<td>0.372</td>
<td><strong>0.512</strong></td>
<td>0.304 [Alford et al., 2017]</td>
</tr>
</tbody>
</table>

5.2 Many 3D molecular learning problems remain underexplored

We hypothesize that the improvement in performance yielded by 3D molecule learning is owed to the fact that a molecule’s behavior is intrinsically linked to its shape. Assuming this shape is accurately represented, it allows for a much more direct relation to the predicted functional properties. Many such problems, however, have not been studied within this framework, leaving significant room for further improvement. One prominent example we explore here is RNA structure ranking, where the state-of-the-art method uses Rosetta [Alford et al., 2017], a hand-designed potential energy function. When we instead apply our 3DCNN and GNN methods that learn directly from the 3D atomistic geometry, we see dramatic increases in performance (Table 6).

We also find room for improvement in domains where 3D molecular learning is already being employed. Protein structure ranking is one such area, and we see that the 3DCNN model is competitive with the state-of-the-art deep learning method by [Pagès et al., 2019] (Table 6), and the GNN model in fact surpasses it in terms of relative assessment (i.e., comparing 3D candidates from the same protein) though not in terms of absolute assessment of correctness (i.e., comparing 3D candidates from different proteins).

Overall, these results demonstrate the potential of 3D molecular learning to address a wide range of problems involving molecular structure, and we anticipate that continued development of such models on less well-studied tasks will aid progress in biomedical research.

5.3 Different tasks require different architectures

While 3D molecular learning methods consistently outperform their non-3D counterparts and provide a systematic way of representing molecular data, our results also provide evidence that architecture selection plays an important role in performance.

For tasks focused on biopolymers and with large amounts of training data (PIP and RES, Figure 1) we observe superior performance from 3DCNNs. We hypothesize this is due to their ability to learn many-body patterns within a single filter, as opposed to standard GNNs which operate on one-body (node) and two-body (edge) features. Such many-body patterns are especially present in biopolymers, as they generally contain complex 3D geometries. This many-body representation hypothesis implies that 3DCNNs have specific advantages in terms of representational power.

However, as the size of the datasets decrease (Table 9), we see more even performance when comparing 3DCNNs and GNNs. In particular, performance is more mixed on the intermediate-sized PSR and RSR datasets, and GNNs nearly fully supplant 3DCNNs on the small-sized MSP, LBA, and LEP datasets. This would be in line with the
While this work demonstrates the potential of 3D structures and provides an initial set of benchmarks, we provide several benchmark datasets and compare the performance of different types of 3D methods. We also show that selection of an appropriate architecture is critical for optimal performance on a given task; depending on the structure of the underlying data, a 3DCNN, GNN, or ENN may be most appropriate, especially in light of our many-body representation hypothesis. Equivariant networks in particular are continuing to improve in efficiency and stability, and we expect these to become more viable due to their close modeling of physical laws.

Finally, in addition to the datasets described here, there are many other open areas in biomedical research and molecular science that are ripe for 3D molecular learning, especially as structural data becomes readily available. Such tasks include virtual screening and pose prediction of small molecule drugs, or the incorporation of conformational ensembles instead of static structures, which would more faithfully represent the entire set of states a molecule could adopt. Like many computational methods in structural biology, this focus on static structures is currently a limitation of ATOM3D; however, through our easily extensible framework we anticipate the addition of new datasets and tasks from across the research community.

Through this work, we hope to lower the entry barrier for machine learning practitioners, encourage the development of algorithms focused on 3D atomistic data, and promote a emerging paradigm within the fields of structural biology and medicinal chemistry.
References


Risi Kondor. N-body networks: a covariant hierarchical neural network architecture for learning 
atomic potentials, 2018.

Christian Kramer and Peter Gedeck. Leave-cluster-out cross-validation is appropriate for scoring func-
2010.

G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack. Improved prediction of protein side-chain 

Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical 
assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, 

Andrew Leaver-Fay, Michael Tyka, Steven Lewis, Oliver Lange, James Thompson, Ron Jacak, 
Kristian Kaufman, P. Renfrew, Colin Smith, Will Sheffler, Ian Davis, Seth Cooper, Adrien Treuille, 
Daniel Mandell, Florian Richter, Yi-H. Ban, Sarel Fleishman, Jacob Corn, David Kim, and 
Philip Bradley. Rosetta3: an object-oriented software suite for the simulation and design of 

Fabrice Lefèvre, Marie-Hélène Rémy, and Jean-Michel Masson. Alanine-stretch scanning mutagene-
sis: a simple and efficient method to probe protein structure and function. *Nucleic acids research*, 

Yan Li, Zhihai Liu, Jie Li, Li Han, Jie Liu, Zhixiong Zhao, and Renxiao Wang. Comparative 

Yang Li and Jianyi Yang. Structural and sequence similarity makes a significant impact on machine-
learning-based scoring functions for protein–ligand interactions. *Journal of Chemical Information 

Shengchao Liu, Mehmet F Demirel, and Yingyu Liang. N-gram graph: Simple unsupervised repre-
sentation for graphs, with applications to molecules. In H. Wallach, H. Larochelle, A. Beygelzimer, 

Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. 
PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31 

Christophe Magnan and Pierre Baldi. Sspro/accpro 5: Almost perfect prediction of protein secondary 
structure and relative solvent accessibility using profiles, machine learning, and structural similarity. 

Simon Mitternacht. FreeSASA: An open source C library for solvent accessible surface area 

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium 
states of many-body systems with deep learning. *Science*, 365(6457), 2019. ISSN 0036-8075. doi: 
10.1126/science.aaw1147. URL [https://science.sciencemag.org/content/365/6457/eaaw1147](https://science.sciencemag.org/content/365/6457/eaaw1147).

Hakime Öztürk, Elif Ozkirimli, and Arzucan Özugür. DeepDTA: Deep Drug-Target Binding Affinity 

Guillaume Pagès, Benoit Charmettant, and Sergei Grudinin. Protein model quality assessment using 

Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein-
2017.


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see Section [B]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see Appendix [B]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)... 
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please see our Github at https://github.com/drorlab/atom3d.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see our website (https://www.atom3d.ai/) or Github (https://github.com/drorlab/atom3d) for details on splits and default hyperparameters. For details on model selection, see Appendix [F]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report standard deviations over 3 replicates in Table [8]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix [F]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
      All datasets are available from our website at https://www.atom3d.ai/.
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes] We obtained explicit consent from all datasets which we use directly in this work.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] No datasets include personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]