xTrimoDock: Cross-Modal Transformer for Multi-Chain Protein Docking

Anonymous authors
Paper under double-blind review

Abstract

The structure of a protein–protein complex plays a critical role in understanding the dynamics of binding, delineating biological mechanisms, and developing intervention strategies. Rigid protein-protein docking, assuming no conformational change within proteins, predicts the 3D structure of protein complexes from unbound chains. According to the number of chains, rigid docking is divided into binary complex setting that contains only two chains, and more ubiquitous multi-chain complex setting. Most existing docking methods are tailored for binary complexes, and are computationally expensive or not guaranteed to find accurate complex structures. In this paper, we propose a novel model xTrimoDock for the docking of multi-chain complexes, which can simultaneously employ information from both sequence modality and structure modality of involved protein chains. Specifically, xTrimoDock leverages a cross-modal transformer to integrate representations from protein sequences and structures, and conducts a multi-step prediction of rotations and translations to accomplish the multi-chain docking. Extensive experiments reflect the promising results of the proposed model in the harder multi-chain complex setting.

1 Introduction

Protein-protein interactions (PPIs) are essential to the basic functioning of cells and larger biological systems in all living organisms. Due to their importance, elucidating such interactions up to atomic detail is necessary for understanding large multicomponent complexes like ribosomes and discovering protein-based drugs, e.g., antibodies, nanobodies, and peptides. While the experimental golden standard for determining the structure of protein complexes, such as X-ray crystallography and cryo-EM, is extremely time-consuming.

Computational protein docking (Venkatraman et al. (2009); Biesiada et al. (2011); Weng et al. (2019); Sunny & Jayaraj (2021)) provides an alternative route to predict the three-dimensional structures of protein complexes from unbound chains. According to the number of chains, we denote protein docking given two chains of ligand and receptor as binary complex setting, and given multiple chains as multi-chain complex setting. In such cases, there is a rigid body assumption (Ganea et al. (2022)) in many biological environments that no deformations occur within any protein during docking. Therefore, all we need are some appropriate SE(3) transformations shown in Figure 1, i.e., rotation and translation in 3D space, that bring one protein to contact with another one.

Classical docking software (Chen et al. (2003); De Vries et al. (2010); Torchala et al. (2013); Kozakov et al. (2017); Yan et al. (2020)) follow a computationally expensive framework that i) randomly samples a huge number of candidate structures, ii) employs a scoring function (Basu & Wallner (2016); Eismann et al. (2021)) to rank them and iii) refines the top structures based on an energy model. Recently, EquiDock (Ganea et al. (2022)) is the first to apply deep learning for direct prediction of protein complex structures, and achieves a great speed-up but sometimes implausible structures. These methods are suitable for binary complex setting, not the multi-chain complex setting where chains of ligand and receptor are indistinguishable as in reality. As an extension of AlphaFold2 (Jumper et al. (2021)), AlphaFold-Multimer (Evans et al. (2021)) can infer multi-chain complex structure from its amino acid sequence but does not make use of known structures, missing essential information of rigid docking.

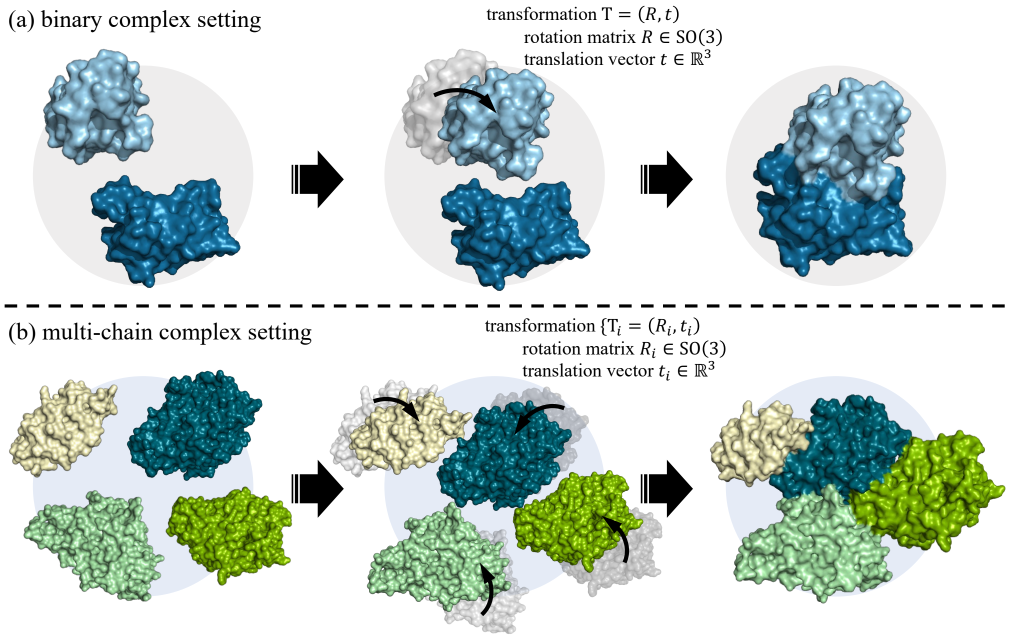

(a) binary complex setting
transformation $\mathrm{T} = (R, t)$
rotation matrix $R \in \mathrm{SO}(3)$
translation vector $t \in \mathbb{R}^3$
(b) multi-chain complex setting
transformation $\{\mathrm{T}_i = (R_i, t_i)$
rotation matrix $R_i \in \mathrm{SO}(3)$
translation vector $t_i \in \mathbb{R}^3$

Figure 1: Surface views of **(a) binary complex setting** and **(b) multi-chain complex setting**. In (a), the number of chains in protein complexes is two. Keeping one chain in its bound location, a SE(3) transformation $T = (\boldsymbol{R}, \boldsymbol{t})$ is predicted to place another chain at the right location and orientation. As for (b), each chain requires a SE(3) transformation $T_i = (\boldsymbol{R}_i, \boldsymbol{t}_i)$ to achieve protein docking.

Contribution We propose a model named xTrimoDock that comprehensively utilizes sequence modality and structure modality of chains to perform multi-chain protein docking. Specifically, an elaborate cross-modal transformer learns representations from protein sequences and structures, and generates chain-level rotation and translation transformations. Then a multi-step prediction mechanism repeatedly feeds the predicted structure recursively into the same cross-modal transformer to gradually refine a highly accurate protein structure. Extensive experiments demonstrate that our xTrimoDock shows promising results on multi-chain protein docking.

2 Related Work

In line with the focus of our work, we briefly review the most related work on pocket druggability prediction, drug-target interface prediction, and protein-protein docking.

Pocket Druggability Prediction The ability of protein pockets to bind drug-like molecules, referred as druggability, is of major interest in the first step of drug discovery. Since protein conformation changes might affect the druggability of pockets, it is necessary to utilize geometric information beyond sequential information. Pioneer work Krasowski et al. (2011); Desaphy et al. (2012); Borrel et al. (2015) predicts druggability based on the predefined descriptors of the rigid pocket structure. Nowadays, an increasing number of methods (Yuan et al. (2020); Zhou et al. (2022)) take into account the conformational changes of proteins.

Drug-Target Interface Prediction Drug-target interactions characterize the binding poses and affinity of compounds to protein targets (Rutkowska et al. (2016); Santos et al. (2017); Zitnik et al. (2019)), playing an essential role in finding effective and safe treatments for new pathogens (Velavan & Meyer (2020)). Deep learning has advanced traditional computational modeling of compounds (Wallach et al. (2015)) by providing increased expressive power in identifying, processing, and extrapolating complex patterns in molecular data (Öztürk et al. (2018); Lee et al. (2019); Eberhardt et al. (2021); McNutt et al. (2021); Bao et al. (2021); Nguyen et al. (2021)). These methods are designed for molecular ligands, and often assume known binding pockets, which are not directly applicable to our rigid docking setting.

Protein-Protein Docking Experimental methods to determine the structure of protein complexes, such as X-ray crystallography and cryo-EM, are mostly financially restrictive and time-consuming when there are tens of thousands of interactions yet to be resolved. Therefore, computational docking methods (Chen et al. (2003); Venkatraman et al. (2009); De Vries et al. (2010); Biesiada et al. (2011); Torchala et al. (2013); Vakser (2014); Schindler et al. (2017); Weng et al. (2019); Yan et al. (2020); Sunny & Jayaraj (2021); Christoffer et al. (2021)) offers an alternative route to predict protein complex structures based on a three-step framework of candidate sampling, ranking (Moal et al. (2013); Basu & Wallner (2016); Launay et al. (2020); Eismann et al. (2021)) and refinement (Verburgt & Kihara (2022)). Recently, deep learning has made a big impact on structural biology (Laine et al. (2021); Dai & Bailey-Kellogg (2021)). AlphaFold2 (Jumper et al. (2021)) and RoseTTAFold (Baek et al. (2021)) have been utilized to improve protein-protein interaction from different aspects (Humphreys et al. (2021); Pei et al. (2022)), e.g., combing physics-based docking methods (Kozakov et al. (2017)) or extending multiple-sequence alignments (Bryant et al. (2022)). Particularly, AlphaFold-Multimer (Evans et al. (2021)) capitalizes on the success of AlphaFold2 to fold and dock two proteins, and EquiDock (Ganea et al. (2022)) is tailored for the effective rigid docking where the performance improvement is limited compared to traditional docking methods.

3 xTrimoDock

In this section, we elaborate the proposed xTrimoDock, a novel framework for multi-chain protein docking. We begin with the overview and subsequently zoom into the details of the cross-modal transformer and multi-step prediction. Lastly, we illustrate the optimization strategy for the model.

3.1 Overview

Sequence Modality Input The available information on sequence modality mainly contains intra-residue and inter-residue information. For intra-residue information, we transform primary sequence features or multiple sequence alignments (MSAs) to residue representations $\{\boldsymbol{z}_i\}$ with $\boldsymbol{z}_i \in \mathbb{R}^d$ and $i \in \{1, \cdots, N_{res}\}$ where d is the dimension of representations and N_{res} is the length of protein sequences. In terms of inter-residue information, relative positional features and primary sequence features form pair representations $\{\boldsymbol{p}_{ij}\}$ with $\boldsymbol{p}_{ij} \in \mathbb{R}^d$ and $i, j \in \{1, \cdots, N_{res}\}$. These representations can be obtained using existing encoders (Jumper et al. (2021); Evans et al. (2021)).

Structure Modality Input To encode structures of protein chains, complex is represented as backbone frames $\{T_i = (\boldsymbol{R}_i \in \mathbb{R}^{3\times3}, \boldsymbol{t}_i \in \mathbb{R}^3)\}$ with $i \in \{1, \cdots, N_{res}\}$ which are SE(3) transformations constructed from the positions of tuples N-Cα-C in residues using the Gram–Schmidt process. We also parameterize atoms of side chains as relative frames from the positions of the atom before the torsion bond, the atom after the torsion bond, and the next atom after that. Once we have determined backbone frames, all atoms of side chains are obtained by the composition of frames (Evans et al. (2021)). The composition of two SE(3) transformations is denoted as

$$(\boldsymbol{R}, \boldsymbol{t}) = (\boldsymbol{R}_1, \boldsymbol{t}_1) \circ (\boldsymbol{R}_2, \boldsymbol{t}_2) = (\boldsymbol{R}_1\boldsymbol{R}_2, \boldsymbol{R}_1\boldsymbol{t}_2 + \boldsymbol{t}_1). \tag{1}$$

Architecture Taking the rigidity of chains into account, we keep the initial backbone frames $\{T_i = (\boldsymbol{R}_i, \boldsymbol{t}_i)\}$ fixed, and achieve multi-chain docking by optimizing chain-level frames $\{T_c = (\boldsymbol{R}_c, \boldsymbol{t}_c)\}$ with $c \in \{1, \cdots, N_{chain}\}$. Thus the complex structure is represented by the composition of frames $\{T_i \circ T_{c_i}\}$ where c_i indicates the chain that residue i belongs to, and the chain-level frames $\{T_c\}$ are initialized to identity transformations $\{(\boldsymbol{I}, 0)\}$.

Given residue representations $\{\boldsymbol{z}_i\}$, pair representations $\{\boldsymbol{p}_{ij}\}$ and backbone frames of complex $\{T_i \circ T_{c_i}\}$, xTrimoDock predicts 3D coordinates of all heavy atoms based on a cross-modal transformer and multi-step prediction mechanism as depicted in Figure 2. Specifically, cross-modal transformer updates residue representations $\{\boldsymbol{z}_i\}$ from the whole complex $\{T_i \circ T_{c_i}\}$ and pair representations $\{\boldsymbol{p}_{ij}\}$ with structure-aware attention, then a chain-level pooling is performed on updated representations $\{\hat{\boldsymbol{z}}_i\}$ to compute frames $\hat{T}_c$ for each chain c. Finally, multi-step prediction composes

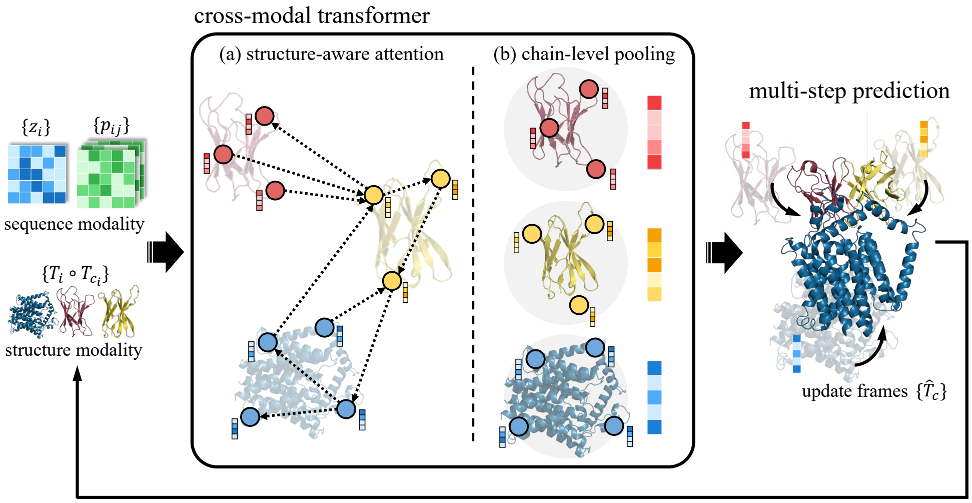


Figure 2: Details on the architecture of **xTrimoDock** which is mainly composed of two modules, cross-modal transformer and multi-step prediction. Taking the information from sequence modality and structure modality as input, **cross-modal transformer** first applies **(a) structure-aware attention** to update residue representations, and utilizes **(b) chain-level pooling** to compute update frames for chains. Then **multi-step prediction** constructs the refined complex structure based on these update frames, and refeeds them to the cross-modal transformer. Finally, the obtained structures and residue representations are constrained by well-designed losses to optimize the model.

update frames $\{\hat{T}_c\}$ and current frames $\{T_c\}$, and feeds them into the same cross-modal transformer to recursively refine a protein structure with more precise atomic details. It is worth noting that the overall architecture is equivariant to the global transformation of the initial complex structure.

3.2 Cross-Modal Transformer

Structure-aware Attention Cross-modal transformer first updates residue representations $\{\boldsymbol{z}_i\}$ by applying structure-aware attention on backbone frames $\{T_i \circ T_{c_i}\}$, and is invariant under any global SE(3) transformation on those frames. Formally, query vectors $\boldsymbol{q}_i^h$, key vectors $\boldsymbol{k}_i^h$, and value vectors $\boldsymbol{v}_i^h$ are computed by

$$\boldsymbol{q}_i^h = \boldsymbol{z}_i \boldsymbol{W}_q^h,\ \boldsymbol{k}_i^h = \boldsymbol{z}_i \boldsymbol{W}_k^h,\ \boldsymbol{v}_i^h = \boldsymbol{z}_i \boldsymbol{W}_v^h,\ h \in \{1, \cdots, N_{head}\}, \tag{2}$$

where $\boldsymbol{q}_i^h, \boldsymbol{k}_i^h, \boldsymbol{v}_i^h \in \mathbb{R}^{d_t}$, and N_{head} is the number of attention heads. To inject information of structure modality, another set of self-attention terms is transformed into three-dimensional space,

$$\boldsymbol{q}_i^{hp} = \boldsymbol{z}_i \boldsymbol{W}_q^{hp},\ \boldsymbol{k}_i^{hp} = \boldsymbol{z}_i \boldsymbol{W}_k^{hp},\ \boldsymbol{v}_i^{hp} = \boldsymbol{z}_i \boldsymbol{W}_v^{hp},\ p \in \{1, \cdots, N_{point}\}, \tag{3}$$

where $\boldsymbol{q}_i^{hp}, \boldsymbol{k}_i^{hp}, \boldsymbol{v}_i^{hp} \in \mathbb{R}^3$ can be taken as virtual points, and N_{point} is the number of them.

When calculating the attention weights between residues, we take into account two factors. i) The closer the residues are, the greater the interaction between them. ii) Pair representations containing inter-residue information exert a vital part. Therefore, attention weights

$$\alpha_{ij}^h = \text{softmax}_j\left(w_L\left(\frac{1}{\sqrt{d_t}}{\boldsymbol{q}_i^h}^\top \boldsymbol{k}_j^h + b_{ij}^h - \frac{\lambda^h w_C}{2}\sum_p \left\| T_i \circ T_{c_i} \circ \boldsymbol{q}_i^{hp} - T_j \circ T_{c_j} \circ \boldsymbol{k}_j^{hp} \right\|^2\right)\right), \tag{4}$$

$$b_{ij}^h = \boldsymbol{p}_{ij} W_p^h, \tag{5}$$

where λ^h is a learnable scalar, and b_{ij}^h is the bias stemming from pair representations $\{\boldsymbol{p}_{ij}\}$. Factors $w_C = \sqrt{1/3}$ and $w_L = \sqrt{2/9N_{points}}$ are computed such that two sets of attention terms contribute equally and the resulting variance of attention weights is 1. Then, these attention weights weight value terms and pair representations

$$\boldsymbol{o}_i^h = \sum_j \alpha_{ij}^h \boldsymbol{v}_j^h, \boldsymbol{o}_i^{hp} = T_{c_i}^{-1} \circ T_i^{-1} \circ \sum_j \alpha_{ij}^h (T_j \circ T_{c_j} \circ \boldsymbol{v}_j^{hp}), \tilde{\boldsymbol{o}}_i^h = \sum_j \alpha_{ij}^h \boldsymbol{p}_{ij}. \tag{6}$$

The residual representations are updated by these outputs of the attention mechanism

$$\hat{\boldsymbol{z}}_i = \mathrm{Linear}\left(\mathrm{concat}_{h,p}\left(\tilde{\boldsymbol{o}}_i^h, \boldsymbol{o}_i^h, \boldsymbol{o}_i^{hp}, \|\boldsymbol{o}_i^{hp}\|\right)\right). \tag{7}$$

Please note that backbone frames are produced from a global reference such that updated representations are invariant to global transformations of the initial complex structure. Specifically, since the l_2-norm of vectors is invariant under any global transformation T,

$$\begin{aligned} \left\|(T \circ T_i \circ T_{c_i}) \circ \boldsymbol{q}_i^{hp} - (T \circ T_j \circ T_{c_j}) \circ \boldsymbol{k}_j^{hp}\right\|^2 &= \left\|T \circ \left(T_i \circ T_{c_i} \circ \boldsymbol{q}_i^{hp} - T_j \circ T_{c_j} \circ \boldsymbol{k}_j^{hp}\right)\right\|^2 \\ &= \left\|T_i \circ T_{c_i} \circ \boldsymbol{q}_i^{hp} - T_j \circ T_{c_j} \circ \boldsymbol{k}_j^{hp}\right\|^2, \end{aligned} \tag{8}$$

the global transformation is canceled. In the computation of output terms, it also cancels out when mapping back

$$\begin{aligned} (T \circ T_i \circ T_{c_i})^{-1} \circ \sum_j \alpha_{ij}^h ((T \circ T_j \circ T_{c_j}) \circ \boldsymbol{v}_j^{hp}) &= T_{c_i}^{-1} \circ T_i^{-1} \circ T^{-1} \circ T \circ \sum_j \alpha_{ij}^h (T_j \circ T_{c_j} \circ \boldsymbol{v}_j^{hp}) \\ &= T_{c_i}^{-1} \circ T_i^{-1} \circ \sum_j \alpha_{ij}^h (T_j \circ T_{c_j} \circ \boldsymbol{v}_j^{hp}). \end{aligned} \tag{9}$$

Chain-Level Pooling Chain-level pooling module in cross-modal transformer aims at generating chain-level SE(3) transformations, refining locations of protein chains. According to a binary chain mask $\boldsymbol{M}_c \in \mathbb{R}^{N_{res} \times d}$ that indicates which residues belong to chain c, we get chain representations through mean pooling

$$\hat{\boldsymbol{s}}_c = \text{Mean-Pooling}\left(\hat{\boldsymbol{Z}} \odot \boldsymbol{M}_c\right),\ c \in \{1, \cdots, N_{chain}\} \tag{10}$$

where $\hat{\boldsymbol{Z}}$ is matrix form of updated residue presentations $\{\hat{\boldsymbol{z}}_i\}$, and N_{chain} is the number of chains in the protein complex. The update frames for chains are created by predicting a rotation quaternion and a translation vector. The first component of the rotation quaternion is fixed to 1, and the rest defining the Euler axis are learned by a linear layer,

$$\hat{b}_c, \hat{c}_c, \hat{d}_c, \hat{\boldsymbol{t}}_c = \mathrm{Linear}(\hat{\boldsymbol{s}}_c). \tag{11}$$

The non-unit quaternion rotation $(1, \hat{b}_c, \hat{c}_c, \hat{d}_c)$ can be converted to a rotation matrix $\hat{\boldsymbol{R}}_c$ (Evans et al. (2021)). Thus the update frames for chains are $\{\hat{T}_c = (\hat{\boldsymbol{R}}_c, \hat{\boldsymbol{t}}_c)\}$.

3.3 MULTI-STEP PREDICTION

Multi-step prediction mechanism composes backbone frames and update frames of chains to compute immediate structures of the protein complex. Then they are fed to the cross-modal transformer with shared weights to replace initial backbone frames. Multi-step prediction allows the network deeper without significantly increasing the training time or the volume of parameters, and refines the predicted structures multiple times.

Specifically, current frames $\{T_c\}$ and the update frames $\{\hat{T}_c\}$ at chain level are composed as $\{T_c \circ \hat{T}_c\}$, replacing current frames $\{T_c\}$ as input to the cross-modal transformer. These intermediate structures along with the final predicted structure are constrained by the Frame Aligned Point Error (FAPE) loss (Jumper et al. (2021)) to accelerate and stabilize model training. Since the side chains are rigid, the FAPE loss only operates on the backbone frames and Cα atom positions. Given backbone frames $\{T_i \circ \hat{T}_{c_i}\}$ and Cα position $\{\hat{\boldsymbol{x}}_i = \hat{\boldsymbol{r}}_i\}$ of prediction, backbone frames $\{T_i \circ T_{c_i}^{gt}\}$ and Cα position $\{\boldsymbol{x}_i^{gt}\}$ of ground truth, the FAPE loss is defined as

$$\hat{\boldsymbol{x}}_{ij} = \hat{T}_{c_i}^{-1} \circ T_i^{-1} \circ \hat{\boldsymbol{x}}_j,\ \boldsymbol{x}_{ij}^{gt} = T_{c_i}^{gt^{-1}} \circ T_i^{-1} \circ \boldsymbol{x}_j^{gt}, \tag{12}$$

Table 1: Statistics of datasets.

	Chains per Protein				**Residues per Protein**				**Atoms per Protein**			
	Min	*Max*	*Mean*	*Std*	*Min*	*Max*	*Mean*	*Std*	*Min*	*Max*	*Mean*	*Std*
Training Set	2	45	2.87	2.39	67	8290	759.95	689.97	354	62495	5918.64	5375.14
Validation Set	2	28	2.81	2.41	91	6368	735.90	716.65	744	49538	5701.60	5552.25
Test Set	3	3	3.00	0.00	408	650	530.52	92.55	3056	5060	4072.59	713.97

$$d_{ij} = \sqrt{\left\|\hat{\boldsymbol{x}}_{ij} - \boldsymbol{x}_{ij}^{gt}\right\|^2 + \epsilon}, \tag{13}$$

$$\mathcal{L}_{\mathrm{FAPE}} = \frac{1}{Z}\,\mathrm{mean}_{i,j}(\mathrm{minimum}(d_{clamp}, d_{ij})), \tag{14}$$

where ϵ is a small constant to ensure that gradients do not vanish, and the resulting deviations $\{d_{ij}\}$ are penalized by a clamped l_1-loss with a normalization constant Z.

The final residue representations $\{\hat{\boldsymbol{z}}_i\}$ predict per-residue lDDT-Cα scores (Mariani et al. (2013)). The true per-residue lDDT-Cα scores $\{r_i^{gt}\}$ are discretized into bins, and supervise the predicted scores as an auxiliary loss (Evans et al. (2021)). Formally,

$$\boldsymbol{l}_i = \mathrm{softmax}(\mathrm{Linear}(\mathrm{relu}(\mathrm{LayerNorm}(\hat{\boldsymbol{z}}_i)))),\ \boldsymbol{l}_i^{gt} = \text{One-Hot}(r_i^{gt}, \boldsymbol{v}_{bins}), \tag{15}$$

$$\mathcal{L}_{aux} = \mathrm{mean}_i(\boldsymbol{l}_i^{\top} \log \boldsymbol{l}_i^{gt}), \tag{16}$$

where $\mathbf{v}_{bins} = [1, 3, 5, \cdots, 99]^{\top}$ represents the vector of bins. Besides, the atom clashes need to be avoided. Thus, we introduce a clash loss to constrain these structural violations in a way that loss-free structures will pass the stereochemical quality checks in the lDDT metric. The clash loss uses a one-sided flat-bottom-potential to penalize too short distances

$$\mathcal{L}_{clash} = \sum_{i=1}^{N_{nb}} \max\left(d_i^{lit} - \tau - \hat{d}_i, 0\right), \tag{17}$$

where N_{nb} is the number of non-bonded atom pairs, d_i^{lit} is the lower bound of distances between non-bonded atom pairs based on literature Van der Waals radii, and $\hat{d}_i$ is the distance in the prediction. The tolerance τ is set to 1.5 Å. In all, the loss of model optimization is as follows

$$\mathcal{L} = \lambda_1\mathcal{L}_{\mathrm{FAPE}} + \lambda_2\mathcal{L}_{aux} + \lambda_3\mathcal{L}_{clash}. \tag{18}$$

Hyperparameters λ_1, λ_2 and λ_3 balance the importance of different losses.

4 Experiments

In this section, we evaluate the effectiveness of xTrimoDock via extensive experiments. Particularly, xTrimoDock is compared with the state-of-the-art methods and presents competitive performance even under multi-chain evaluation protocol. We further analyze the performance differences for short- and long-chain proteins. Lastly, we conduct ablation studies to investigate the effectiveness of the multi-step prediction.

4.1 Experimental Setup

Dataset We leverage Database of Interacting Protein Structures (DIPS) (Townshend et al. (2019)), which is a protein complex benchmark mined from the Protein Data Bank (Bank (1971)) and tailored for rigid docking. Following the experimental settings of EquiDock (Ganea et al. (2022)), we filter DIPS to only keep proteins with at least 30 residues and at most 10K atoms. Dataset is partitioned in train/validation splits of sizes 12136/369 based on the protein family to separate similar proteins. For a fair comparison, the test set consists of 27 antibody-antigen complexes released after October 2021 that have not been used to train AlphaFold-Multimer (Evans et al. (2021)). Particularly, the antibody-antigen complex is composed of three chains, i.e., the heavy/light chain of the antibody and the one chain from the antigen. The statistics of the dataset are summarized in Table 1.

Table 2: Quantitative results on **complex prediction**.

	RMSD ↓			**TM-score** ↑			**GDT-TS** ↑			**GDT-HA** ↑		
	Med.	*Mean*	*Std*	*Med.*	*Mean*	*Std*	*Med.*	*Mean*	*Std*	*Med.*	*Mean*	*Std*
ZDOCK	21.37	21.35	6.49	0.69	0.66	0.12	0.66	0.65	0.12	0.66	0.65	0.12
HADDOCK	22.26	21.39	5.15	0.69	0.66	0.12	0.66	0.65	0.12	0.66	0.65	0.12
ClusPro	17.01	15.94	7.21	0.69	0.70	0.13	0.68	0.67	0.12	0.68	0.66	0.12
HDOCK	13.20	10.90	10.42	0.72	0.81	0.18	0.70	0.80	0.19	0.70	0.78	0.18
EquiDock	92.67	100.57	67.44	0.66	0.65	0.12	0.66	0.65	0.12	0.66	0.65	0.12
Multimer	13.84	13.31	5.85	0.67	0.67	0.12	0.55	0.57	0.11	0.44	0.45	0.10
xTrimoDock	1.46	1.50	0.48	0.97	0.96	0.02	0.82	0.82	0.07	0.59	0.61	0.10

Baselines To evaluate the effectiveness of our proposed xTrimoDock, we compare it with two categories of representative methods, including four docking software ZDOCK (Chen et al. (2003)), HADDOCK (De Vries et al. (2010)), ClusPro (Kozakov et al. (2017)), HDOCK (Yan et al. (2020)), two deep learning models EquiDock (Ganea et al. (2022)) and AlphaFold-Multimer (Multimer for short, Evans et al. (2021)).

Metrics To measure the quality of predictions, we report universally accepted metrics Root Mean Square Deviation (RMSD), TM-score (Template Modeling score), GDT-TS (Global Distance Test-Total Score), and GDT-HA (Global Distance Test-High Accuracy). Given the ground truth positions $\{\boldsymbol{x}_i^{gt}\}$ and predicted positions $\{\hat{\boldsymbol{x}}_i\}$, the Kabsch algorithm computes a SE(3) transformation T_{align} to superimpose them, and RMSD is $1/N_{atom}\sum_i \|\hat{\boldsymbol{x}}_i - T_{align} \circ \boldsymbol{x}_i^{gt}\|$. And TM-score is defined by

$$\text{TM-score} = \max\left[\frac{1}{L}\sum_{i}^{L}\frac{1}{1+\left(\frac{d_i}{d_0(L)}\right)^2}\right],$$

where L is the length of the amino acid sequence, d_i is the distance between the i-th pair of residues in the prediction and ground truth, and $d_0(L) = 1.24\sqrt[3]{L-15} - 1.8$ is a distance scale that normalizes distances. For a given value of cutoff, GDT-score represents the maximum proportion of atoms that can makes RMSD less than the cutoff. GDT-scores are usually calculated w.r.t. different cutoffs, thus GDT-TS is the mean of 1, 2, 4, 8Å, and GDT-HA corresponds to 0.5, 1, 2, 4Å. We compute these metrics using the tool DeepAlign (Wang et al. (2013)).

Implementations ZDOCK, HADDOCK, ClusPro and HDOCK provide user-friendly local packages suitable for automatic experiments or webservers for manual submissions. For Equidock[1] and Multimer[2], we use their pretrained models released on GitHub for inference. Baselines except Mutlimer are designed for the binary complex setting, so the heavy chain and light chain of antibody are merged during the evaluation.

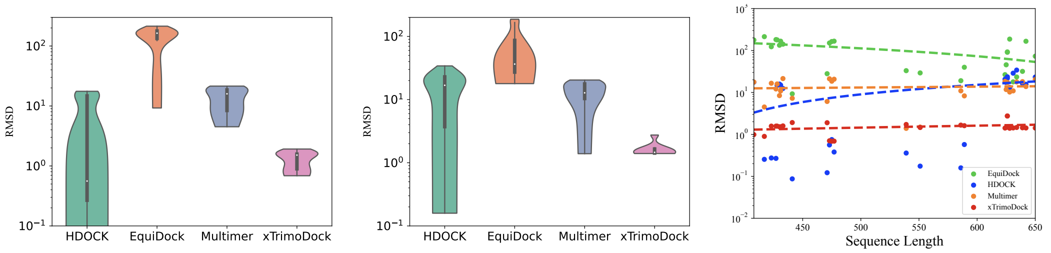
In terms of xTrimoDock, we utilize the single representations and pair representations from Multimer as input, and train it with crop size 128. For hyperparameters in the cross-modal transformer, we set the number of heads $N_{head} = 12$, the number of virtual point $N_{point} = 8$, and embedding dimension $d_t = 16$. We use Adam optimizer with learning rate 10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the coefficients in loss are $\lambda_1 = 0.5$, $\lambda_2 = 0.01$ and $\lambda_3 = 0.5$.

4.2 Results and Analysis

Complex Prediction Based on the results shown in Table 2, xTrimoDock generally performs competitive predictions under the multi-chain evaluation. This demonstrates that the usage of cross-

[1] https://github.com/octavian-ganea/equidock_public

[2] https://github.com/aqlaboratory/openfold



(a) Short-chain proteins

(b) Long-chain proteins

(c) Relationship between sequence length and RMSD

Figure 3: Violin plots of RMSD for (a) **short-chain proteins** and (b) **long-chain proteins**. The box inside violin indicates 25-75 percentiles, and the median is shown by a white scatter. We also depict (c) **RMSD w.r.t. sequence length**, where scatters refer to RMSD of a specific protein and dotted lines are regression lines of scatters.

Table 3: **Ablation studies** on the multi-step prediction mechanism of xTrimoDock. (w/o MS: without multi-step prediction)

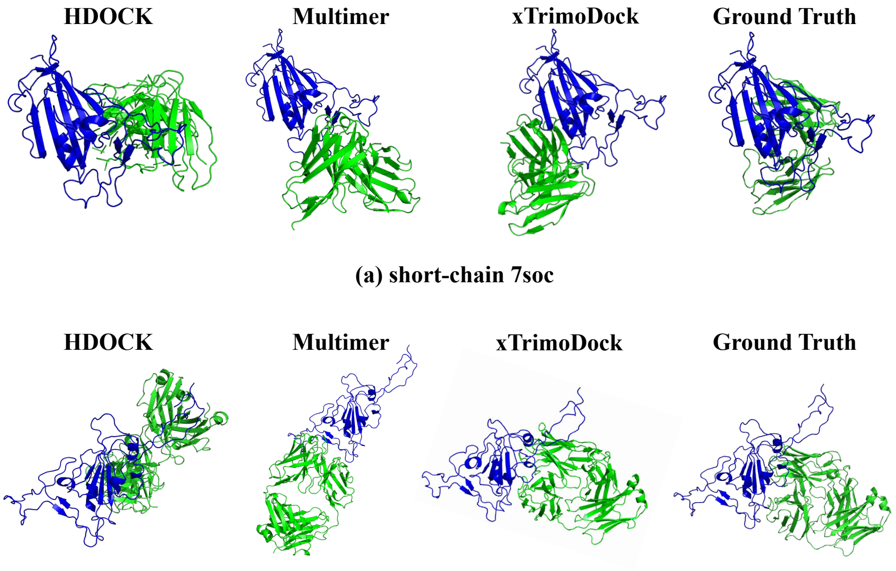
	RMSD ↓			**TM-score** ↑			**GDT-TS** ↑			**GDT-HA** ↑		
	Med.	*Mean*	*Std*	*Med.*	*Mean*	*Std*	*Med.*	*Mean*	*Std*	*Med.*	*Mean*	*Std*
w/o MS	3.77	3.84	0.13	0.91	0.91	0.01	0.76	0.75	0.05	0.59	0.57	0.05
xTrimoDock	1.46	1.50	0.48	0.97	0.96	0.02	0.82	0.82	0.07	0.59	0.61	0.10

modal information and the iterative refinement of structures can characterize proteins in a more precise manner. We note that most baselines are evaluated in the binary setting where the heavy chain and light chain of antibody are merged, giving them an advantage in the performance. Moreover, we observe that docking software still provide reliable predictions at the expense of large computational costs, and deep learning methods sometimes have leeway in performance. For example, the mean and standard deviation of RMSD evaluated from EquiDock are large, indicating that some inappropriate SE(3) transformations are learned, and the ligand and receptor are even far apart.

Effects of Sequence Length We divided proteins into short- and long- chain proteins based on the mean value 530.52 of the number of residues per protein in the test set, and chose representative baselines HDock, EquiDock and Multimer along with xTrimoDock to analyze performance differences. Violin plots are used to display distributions of RMSD on short- and long-chain proteins respectively, and the relationship between RMSD and sequence length is also drawn in Figure 3. It is found that xTrimoDock performs well in structure prediction of short- and long-chain proteins, and an interesting phenomenon is that deep learning methods is not as much sensitive to the sequence length as docking software.

Visualization We randomly select a short-chain protein with $PDB\_id = 7soc$ and a long-chain protein with $PDB\_id = 7sn2$, visualizing their ground truth structures and predictions of competitive methods in the Figure 4. We intuitively see that xTrimoDock usually exhibits high accuracy to identify the docking pockets and docked poses.

Ablation Study for Multi-Step Prediction Recall that multi-step prediction mechanism in xTrimoDock refines the predicted complex structure via recursively feeding the immediate structures into cross-modal transformer. We alter the multi-step prediction by forward propagation only once to validate the effectiveness of this mechanism. Results of the ablation study are reported in Table 3. We can observe that xTrimoDock is consistently better than the variant. Such observations imply that multi-step prediction contributes markedly to accurate predictions.



(a) short-chain 7soc

(b) long-chain 7sn2

Figure 4: Visualization of (a) **short-chain** and (b) **long-chain protein complexes** respectively.

Table 4: **Total inference time** of different methods. (unit: hour)

	ZDOCK	**HADDOCK**	**ClusPro**	**HDOCK**	**EquiDock**	**Multimer**	**xTrimoDock**
Inference time	28.83	12.15	34.65	8.1	0.24	0.68	0.41

Computational Efficiency We show total inference time in Table 4. The results are in line with our intuition that three-step framework of docking software spends lots of time sampling, ranking and refining candidates. Fortunately, deep learning methods achieve 10-150x speed-up. This is especially important for the drug design, which need to be extremely fast to scan the vast biological and chemical spaces for both desired and unexpected effects. For instance, the human proteome contains up to 100,000 protein types, and a novel drug might negatively inhibit essential proteins. Therefore, the current hope is to scan for these interactions in a computational manner before bringing a few promising candidates to in vitro and in vivo testing.

5 Conclusion

In this paper, we have presented a promising approach for underexplored multi-chain rigid docking that organically utilizes information from structure modality and sequence modality. Additionally, our method smartly adopts multi-step prediction to refine chain-level SE(3) transformations by recursively feeding them into the same modules. Various experiments show competitive results on the harder multi-chain docking.

Limitations and Broader Impact A limitation of xTrimoDock is that the protein flexibility is not taken into account. Flexibility is of overwhelming importance for protein function, and the changes in protein structure during interactions with binding partners can be dramatic. Therefore, we look forward to extending xTrimoDock for flexible docking, and apply it to more tasks in drug design. As another interesting direction for future work, we might attempt to replace multi-step prediction with a diffusion model to see if we can get some new insights. Last, we hope that our work can inspire the community to pay more attention on deep learning in biological scenarios.

REFERENCES

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Protein Data Bank. Protein data bank. *Nature New Biol*, 233:223, 1971.

Jingxiao Bao, Xiao He, and John ZH Zhang. Deepbsp—a machine learning method for accurate prediction of protein–ligand docking structures. *Journal of Chemical Information and Modeling*, 61(5):2231–2240, 2021.

Sankar Basu and Björn Wallner. Dockq: a quality measure for protein-protein docking models. *PloS one*, 11(8):e0161879, 2016.

Jacek Biesiada, Aleksey Porollo, Prakash Velayutham, Michal Kouril, and Jaroslaw Meller. Survey of public domain software for docking simulations and virtual screening. *Human genomics*, 5(5): 1–9, 2011.

Alexandre Borrel, Leslie Regad, Henri Xhaard, Michel Petitjean, and Anne-Claude Camproux. Pockdrug: A model for predicting pocket druggability that overcomes pocket estimation uncertainties. *Journal of chemical information and modeling*, 55(4):882–895, 2015.

Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using alphafold2. *Nature communications*, 13(1):1–11, 2022.

Rong Chen, Li Li, and Zhiping Weng. Zdock: an initial-stage protein-docking algorithm. *Proteins: Structure, Function, and Bioinformatics*, 52(1):80–87, 2003.

Charles Christoffer, Siyang Chen, Vijay Bharadwaj, Tunde Aderinwale, Vidhur Kumar, Matin Hormati, and Daisuke Kihara. Lzerd webserver for pairwise and multiple protein–protein docking. *Nucleic Acids Research*, 49(W1):W359–W365, 2021.

Bowen Dai and Chris Bailey-Kellogg. Protein interaction interface region prediction by geometric deep learning. *Bioinformatics*, 37(17):2580–2588, 2021.

Sjoerd J De Vries, Marc Van Dijk, and Alexandre MJJ Bonvin. The haddock web server for data-driven biomolecular docking. *Nature protocols*, 5(5):883–897, 2010.

Jérémy Desaphy, Karima Azdimousa, Esther Kellenberger, and Didier Rognan. Comparison and druggability prediction of protein–ligand binding sites from pharmacophore-annotated cavity shapes, 2012.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021.

Stephan Eismann, Raphael JL Townshend, Nathaniel Thomas, Milind Jagota, Bowen Jing, and Ron O Dror. Hierarchical, rotation-equivariant neural networks to select structural models of protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 89(5):493–501, 2021.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew W Senior, Timothy Green, Augustin Žídek, Russell Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *BioRxiv*, 2021.

Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi S. Jaakkola, and Andreas Krause. Independent se(3)-equivariant models for end-to-end rigid protein docking. In *ICLR*. OpenReview.net, 2022.

Ian R Humphreys, Jimin Pei, Minkyung Baek, Aditya Krishnakumar, Ivan Anishchenko, Sergey Ovchinnikov, Jing Zhang, Travis J Ness, Sudeep Banjade, Saket R Bagde, et al. Computed structures of core eukaryotic protein complexes. *Science*, 374(6573):eabm4805, 2021.

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

Dima Kozakov, David R Hall, Bing Xia, Kathryn A Porter, Dzmitry Padhorny, Christine Yueh, Dmitri Beglov, and Sandor Vajda. The cluspro web server for protein–protein docking. *Nature protocols*, 12(2):255–278, 2017.

Agata Krasowski, Daniel Muthas, Aurijit Sarkar, Stefan Schmitt, and Ruth Brenk. Drugpred: a structure-based approach to predict protein druggability developed using an extensive nonredundant data set. *Journal of chemical information and modeling*, 51(11):2829–2842, 2011.

Elodie Laine, Stephan Eismann, Arne Elofsson, and Sergei Grudinin. Protein sequence-to-structure learning: Is this the end (-to-end revolution)? *Proteins: Structure, Function, and Bioinformatics*, 89(12):1770–1786, 2021.

Guillaume Launay, Masahito Ohue, Julia Prieto Santero, Yuri Matsuzaki, Cécile Hilpert, Nobuyuki Uchikoga, Takanori Hayashi, and Juliette Martin. Evaluation of consrank-like scoring functions for rescoring ensembles of protein–protein docking poses. *Frontiers in molecular biosciences*, 7: 559005, 2020.

Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6): e1007129, 2019.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):1–20, 2021.

Iain H Moal, Mieczyslaw Torchala, Paul A Bates, and Juan Fernández-Recio. The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC bioinformatics*, 14(1): 1–15, 2013.

Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

Jimin Pei, Jing Zhang, and Qian Cong. Human mitochondrial protein complexes revealed by large-scale coevolution analysis and deep learning-based structure modeling. *Bioinformatics*, 38(18): 4301–4311, 2022.

Anna Rutkowska, Douglas W Thomson, Johanna Vappiani, Thilo Werner, Katrin M Mueller, Lars Dittus, Jana Krause, Marcel Muelbaier, Giovanna Bergamini, and Marcus Bantscheff. A modular probe strategy for drug localization, target identification and target occupancy measurement on single cell level. *ACS chemical biology*, 11(9):2541–2550, 2016.

Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, 16(1):19–34, 2017.

Christina EM Schindler, Isaure Chauvot de Beauchêne, Sjoerd J de Vries, and Martin Zacharias. Protein-protein and peptide-protein docking and refinement using attract in capri. *Proteins: Structure, Function, and Bioinformatics*, 85(3):391–398, 2017.

Sharon Sunny and PB Jayaraj. Fpdock: Protein–protein docking using flower pollination algorithm. *Computational Biology and Chemistry*, 93:107518, 2021.

Mieczyslaw Torchala, Iain H Moal, Raphael AG Chaleil, Juan Fernandez-Recio, and Paul A Bates. Swarmdock: a server for flexible protein–protein docking. *Bioinformatics*, 29(6):807–809, 2013.

Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Ilya A Vakser. Protein-protein docking: From interaction to interactome. *Biophysical journal*, 107 (8):1785–1793, 2014.

Thirumalaisamy P Velavan and Christian G Meyer. The covid-19 epidemic. *Tropical medicine & international health*, 25(3):278, 2020.

Vishwesh Venkatraman, Yifeng D Yang, Lee Sael, and Daisuke Kihara. Protein-protein docking using region-based 3d zernike descriptors. *BMC bioinformatics*, 10(1):1–21, 2009.

Jacob Verburgt and Daisuke Kihara. Benchmarking of structure refinement methods for protein complex models. *Proteins: Structure, Function, and Bioinformatics*, 90(1):83–95, 2022.

Izhar Wallach, Michael Dzamba, and Abraham Heifets. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*, 2015.

Sheng Wang, Jianzhu Ma, Jian Peng, and Jinbo Xu. Protein structure alignment beyond spatial proximity. *Scientific reports*, 3(1):1–7, 2013.

Gaoqi Weng, Ercheng Wang, Zhe Wang, Hui Liu, Feng Zhu, Dan Li, and Tingjun Hou. Hawkdock: a web server to predict and analyze the protein–protein complex based on computational docking and mm/gbsa. *Nucleic acids research*, 47(W1):W322–W330, 2019.

Yumeng Yan, Huanyu Tao, Jiahua He, and Sheng-You Huang. The hdock server for integrated protein–protein docking. *Nature protocols*, 15(5):1829–1852, 2020.

Jui-Hung Yuan, Sungho Bosco Han, Stefan Richter, Rebecca C Wade, and Daria B Kokh. Druggability assessment in trapp using machine learning approaches. *Journal of Chemical Information and Modeling*, 60(3):1685–1699, 2020.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. 2022.

Marinka Zitnik, Francis Nguyen, Bo Wang, Jure Leskovec, Anna Goldenberg, and Michael M Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, 2019.

A Appendix

You may include other additional sections here.