RWTH AACHEN UNIVERSITY

Computational Design and Cell-Free Expression of Coat Protein Subunits for the Functional Assembly into Potato Virus X-like Particles

Thesis
in Molecular and Applied Biotechnology (B.Sc.)
at RWTH Aachen University

Submitted on: July 3, 2025
by: Kilian Mandon
1. Appraiser: Prof. Dr. Stefan Schillberg
2. Appraiser: Jun. Prof. Dr. Anna Matuszyńska

RWTH Aachen University

Contents

1 Introduction

Virus-like particles (VLPs) are nanostructures assembled from viral protein subunits, but lacking genetic material for replication, and are therefore non-infectious [9]. Due to their properties, VLPs have a wide number of medical applications. The particles can serve as carriers for genes, proteins, or small drugs, as well as employment as scaffolds for proteins. A particular benefit of their use as carriers is their ability for targeted drug delivery [18]. Because of their effectiveness as therapeutics and their high biosafety, several VLPs are already clinically approved, such as the Human papilloma virus vaccine Gardasil® or the Hepatitis B virus vaccine PreHevbrio® [17].

Potato Virus X (PVX) is a *Potexvirus* in the family of the *Alphaflexiviridae* and has proven to be a versatile delivery mechanism for a variety of therapeutic agents. It was successfully modified as a carrier for the tumor necrosis factor-related apoptosis-inducing ligand (TRAIL), a protein drug inducing apoptosis in cancer cells [13]. Further, PVX nanoparticles were stably associated with doxorubicin, a commonly used chemotherapeutic [14] and, more generally, can be functionalized by standard amine chemistry or "click" chemistry [21]. However, PVX is currently unsuited for use as a VLP since its coat protein fails to assemble in absence of its viral genome [19].

This shortcoming could be alleviated by modern developments in Computational Protein Design. The Deep Learning based method ProteinMPNN is capable of generating novel sequences to match a certain backbone structure for monomers, heteromers, and also homomers. The sequences are highly soluble and often assemble correctly into a desired oligomeric state [6]. The backbones used by ProteinMPNN can be generated with RFdiffusion, a diffusion-based algorithm capable of creating probable protein backbones to satisfy a variety of conditions, such as scaffolding a fixed motif, binding to a predetermined target, or satisfying a specific geometry [24]. The already high success rate of these tools can be even further enhanced by filtering designs through metrics based on structure prediction tools [5]. All these developments in the field of protein design were created by David Baker's Lab, whose work in "computational protein design" was awarded with the Nobel Prize in Chemistry in 2024.

In this work, we will describe our efforts to use the aforementioned design tools to realize a modified PVX coat protein that assembles in absence of its viral genome. The designs created through computational methods will be experimentally expressed using the ALiCE® cell-free expression systems, and analyzed for assembly into VLPs. The ALiCE® lysate is derived from *Nicotiana tabacum*, a diagnostic host of PVX [23], rendering the lysate a natural choice for expression.

Part I
Computational Modeling

2 Symmetry Analysis and Model Building

The structure of PVX was determined by [8], up to a resolution of 2.2 Å. The structural data was made available through the PDB, as a file containing 13 consecutive protein subunits, forming one-and-a-half cycles of the helix. Notably, structural determination was not possible for the 29 amino acids long N-terminal domain, due to its flexibility. The N-terminal domain is relevant for the particle's structure and assembly. In experiments by [4], it was found that while a deletion of the 22 N-terminal amino acids still leaves the virus infectious, the morphology of the particles changes from the wild type. The structure-guided computational design process will only be used to design the rigid part of the protein, the N-terminal domain will however be fused to the constructs in the end.

The following chapters require a flexible way to use the symmetry of PVX, such as the ability to generate different configurations of monomers (e.g. a 3×3 neighborhood of monomers), or the ability to dynamically enforce this symmetry during symmetry-guided prediction with AlphaFold (section 5) or symmetry-guided design with RFdiffusion (section 4). Therefore, this section discusses the computation of the symmetry relationship between consecutive monomers, and how it can be applied to generate new configurations of monomers.

Let $\{\vec{\mathbf{r}}_{j,i}^{\text{original}}\}$ denote the backbone atom positions of chain j in the original PDB file, and let $\{\vec{\mathbf{r}}_{j}^{\text{original}}\}$ be their arithmetic mean.

We choose $T_0 = (I_3, \vec{\mathbf{r}}_A^{\text{original}})$ as our new origin, centered on chain A. The backbone atom coordinates in this frame are denoted by $\vec{\mathbf{r}}_{j,i}$, and we have

$$\vec{\mathbf{r}}_{j,i} = T_0^{-1} \circ \vec{\mathbf{r}}_{j,i}^{\text{original}} = \vec{\mathbf{r}}_{j,i}^{\text{original}} - \vec{\mathbf{r}}_A^{\text{original}} \tag{1}$$

The frames of all other chains in these coordinates are computed as the optimal rigid body transform to align the chain with A. That is,

$$T_j = \underset{T \in \mathrm{SE}(3)}{\arg\min} \sum_i \left\| T \circ \vec{\mathbf{r}}_{A,i} - \vec{\mathbf{r}}_{j,i} \right\|^2 \tag{2}$$

Using the Kabsch algorithm [12], T_j can be computed as $T_j = (R_j, \vec{\mathbf{t}}_j)$, where

$$\vec{\mathbf{t}}_j = \vec{\mathbf{r}}_j - R_j \vec{\mathbf{r}}_A = \vec{\mathbf{r}}_j \tag{3}$$

since $\vec{\mathbf{r}}_A = \vec{\mathbf{0}}$, and $R_j \in \mathrm{SO}(3)$ minimizes

$$\sum_i \left\| R_j(\vec{\mathbf{r}}_{A,i} - \vec{\mathbf{r}}_A) - (\vec{\mathbf{r}}_{j,i} - \vec{\mathbf{r}}_j) \right\| \tag{4}$$

Following the Kabsch Algorithm, R_j can be computed via the singular value decomposition

$$(\vec{\mathbf{r}}_{A,i} - \vec{\mathbf{r}}_A)^T \cdot (\vec{\mathbf{r}}_{j,i} - \vec{\mathbf{r}}_j) = U \Sigma V^T \tag{5}$$

as

$$R_j = V \cdot \mathrm{diag}(1, 1, d) \cdot U^T \tag{6}$$

where $d = \det(U)\det(V)$ corrects for a potential reflection in the orthogonal matrices U and V.

With all frames T_j expressed in the same coordinate system, we can compute the relative transform

$$T_{j\to j+1} = (R_{j\to j+1}, \vec{\mathbf{t}}_{j\to j+1}) = T_j^{-1} \circ T_{j+1} \tag{7}$$

Given the symmetry of the viral coat structure, these transforms are expected to be equal. The average relative transform $T_R = (R_R, \vec{\mathbf{t}}_R)$ is computed by choosing $\vec{\mathbf{t}}_R$ as the mean over $\{\vec{\mathbf{t}}_{j\to j+1}\}$ and choosing $R_R \in \mathrm{SO}(3)$ as the rotation matrix closest to the average over all $R_{j\to j+1}$, that is $R_R = UV^T$ where $U\Sigma V^T = \frac{1}{n}\sum_j R_{j\to j+1}$ [20] (given the similarity of the $\{R_{j\to j+1}\}$, no reflection can arise by continuity).

The individual rotations $R_{j\to j+1}$ had standard deviation $\Delta R_R = 0.004\,\mathrm{rad}$ in geodesic distance, and the individual translations had standard deviation $\Delta \mathbf{t}_R = 0.04\,\text{Å}$. R_R closely resembles a pure rotation around the z-axis $R_Z(\theta)$, with an angle of $\theta = -0.707\,\mathrm{rad}$. The deviation is $d(R_R, R_Z(\theta)) = 0.005\,\mathrm{rad}$. This value of θ corresponds to a left-handed helix with 8.89 subunits per turn. The computed rise is $\mathbf{t}_z = 3.87\,\text{Å}$ per subunit, resulting in a helical pitch (rise per turn) of 34.4 Å. These values are mostly consistent with the ones stated in [8] (rise 3.96 Å, rotation of 0.707 rad, 8.9 copies per turn, helical pitch 35.2 Å). However, the authors emphasize the slight difference in the helical pitch of 35.2 Å compared to that of similar flexible filamentous plant viruses (PepMV, BaMV, and PapMV), for which the helical pitch ranges from 34.3 Å to 34.6 Å. According to the calculations above, the helical pitch in the PDB entry (which the authors produced through multiple cycles of real space refinement) differs from the original helical parameters fitted to the cryo-EM data and falls into the range of the other plant viruses, thereby potentially diminishing the significance of the reported pitch deviation.

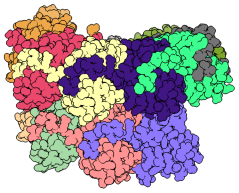
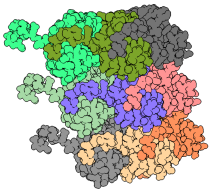
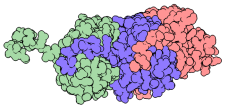
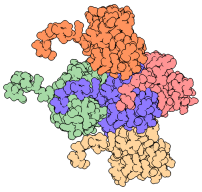
Given the relative transform T_R, model coordinates can be reconstructed based on the coordinates of the monomer A according to

$$\vec{\mathbf{r}}_{j,i}^{\,\text{original}} = T_0 \circ T_R^j \circ \vec{\mathbf{r}}_{A,i} \tag{8}$$

Using Equation 8, four different configurations of monomers are generated and used throughout the following sections (Table 1). A helical configuration consisting of thirteen consecutive monomers, a three-by-three neighborhood of nine monomers, a trimer consisting of three consecutive monomers, and a pentamer consisting of five monomers aranged in a cross-shape.

Despite the small standard deviation of T_R, the deviation of individual atom positions in the helical thirteen-monomer reconstruction compared to the data from the pdb entry reaches up to 0.8 Å. This is due to lever effects caused by small deviations in the rotation. The difference in structure introduces no new clashes, but slightly reduces the contacts by 2 %, as computed with ChimeraX [16].

Table 1: Visualization and chain indices of different monomer configurations, generated based on the average relative transform T_R. The blue chain has index 0, the coordinates for the other chains are computed as $T_R^j \circ \vec{\mathbf{r}}_{A,i}, j \in I$. The generated monomer configurations will be used to create inputs for the algorithms in the following sections.

Type	**Indices**	**Visualization**
Helical	$I = \{0, ..., 12\}$	
3x3	$I = \{0, \pm 1, \pm 8, \pm 9, \pm 10\}$	
Trimer	$I = \{0, \pm 1\}$	
Pentamer	$I = \{0, \pm 1, \pm 9\}$	

3 Sequence Design with ProteinMPNN

ProteinMPNN [6] is a deep learning model for protein sequence design, capable of creating de-novo designs of proteins that fold into a desired shape or bind to specific targets. The algorithm can create sequences for monomers, heterooligomers, and homooligomers.

The sequence is designed based on a protein backbone as input, that is the position of all backbone atoms of one or multiple chains. The underlying algorithm uses a Message Passing Neural Network (MPNN), a graph-based machine learning model. Each residue in the protein is encoded as a vertex in the graph, and edges are drawn up from each residue to its 48 closest neighbors. Vertex embeddings are initialized as 0 vectors, while the initial edge embeddings are computed based on the distances between the backbone atoms of the residue pair and the difference of their residue indices. After the computation of the initial feature embeddings, ProteinMPNN follows an encoder-decoder architecture, in which the encoder updates the edge and vertex embeddings based on their neighborhood, wheras the decoder uses the embeddings computed by the encoder to predict the amino acid type for each residue. The decoder works in an autoregressive fashion by choosing a random order for decoding the individual residues, then predicting their residue type one-by-one with knowledge of all already predicted residues. Concretely, the algorithm predicts logits $\{\ell_i\}$ for each amino acid and chooses it from a softmax distribution according to

$$P(a_i) = \frac{\exp\left(\frac{\ell_i}{\tau}\right)}{\sum_{j=1}^{20} \exp\left(\frac{\ell_j}{\tau}\right)}$$

Here, $\tau > 0$ denotes a chosen temperature constant in the softmax distribution. For $\tau \to \infty$, the distribution is almost uniform, while for $\tau \to 0$ the amino acid with the highest predicted logit is chosen. The distribution can be biased by adding to the logits before sampling. For homooligomers, the logits of identical residues in different monomers are averaged and only one amino acid is sampled from the distribution for all of them.

In this work, all sequences used in computational and experimental evaluation are generated using ProteinMPNN. The input structure is either chosen as the backbone structure of the wildtype, thereby generating alternative sequences for the structure, or a generated artificial backbone as described in section 4. Of particular note is the choice of the input structure: The helical virus particle consists of approximately 1300 monomers [8], and truncation to a smaller number will lead to an incorrect neighborhood during featurization for newly exposed residues.

However, a modification to the original ProteinMPNN algorithm can circumvent this by allowing sequence prediction for a theoretical infinite extension of a symmetric homooligomer. In ProteinMPNN, feature initialization is solely dependent on the relative neighborhood of each residue, meaning that initialization is identical for all corresponding residues in a symmetric homooligomer. Further, the message-passing algorithm in the network conserves this equivariance. Therefore, a theoretical infinite extension of the homooligomer can be simulated by remapping of interchain edges to the corresponding residue in the same chain (Figure 1), thereby reducing the input to a single monomer.

When testing this new algorithm for different helical viruses, the Graph Reduction procedure showed no significant improvement compared to prediction based on

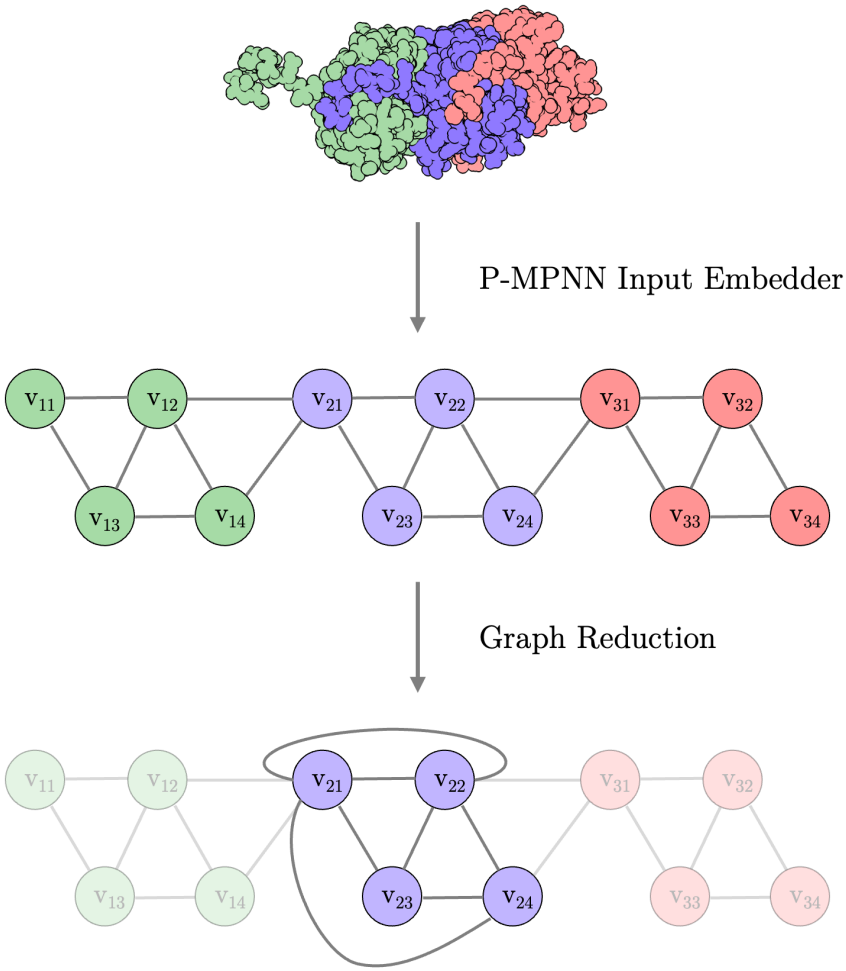


Figure 1: Graph Reduction procedure for symmetric homooligomers. After the default graph initialization from ProteinMPNN, one of the monomers is chosen as the reference monomer. Edges going out from it to other monomers are remapped to the corresponding residue in itself. Afterward, vertices and edges of the non-reference monomers are discarded.

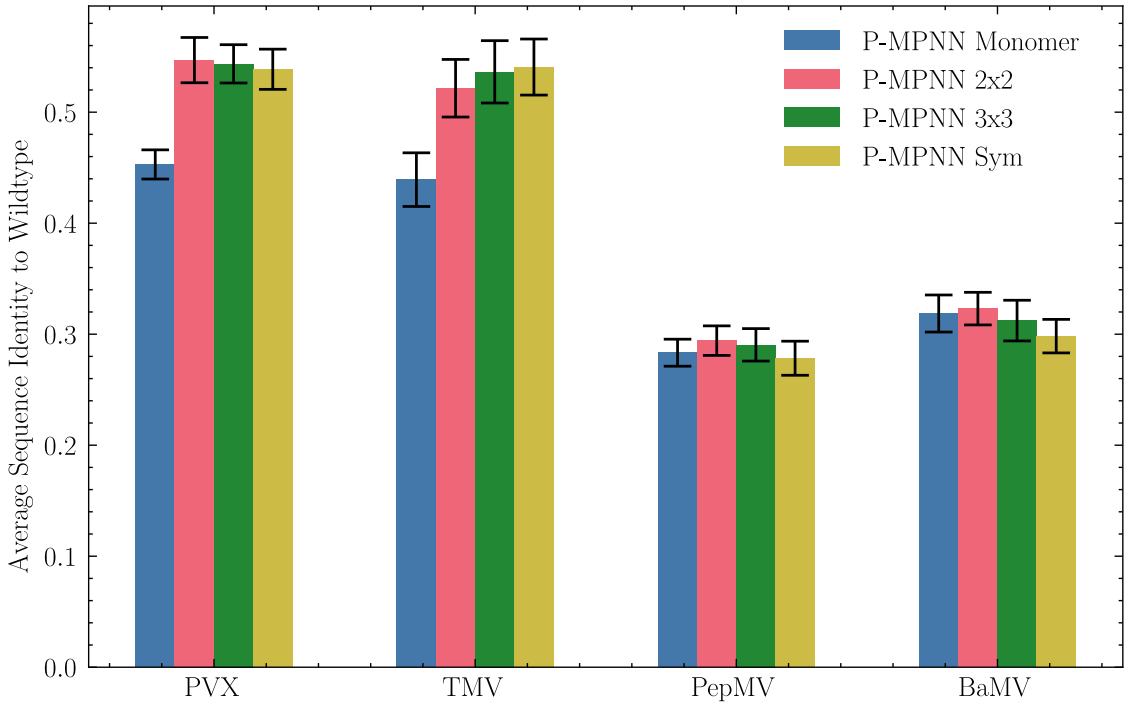


Figure 2: Sequence recovery by ProteinMPNN for different input configurations. The input was chosen as either a single monomer, a 2x2 neighborhood, a 3x3 neighborhood, or a symmetry-preserving graph reduction, modeling a theoretical infinite neighborhood. For each of the four targets Potato Virus X (PVX), Tobacco Mosaic Virus (TMV), Pepino Mosaic Virus (PepMV) and Bamboo Mosaic Virus (BaMV), each model was evaluated 50 times using random decoding orders and a sampling temperature $\tau \to 0$. The errorbars indicate the standard deviation over the repeated evaluation.

a 2x2 neighborhood or a 3x3 neighborhood of monomers (Figure 2). For PVX and Tobacco Mosaic Virus (TMV), the three multimeric inputs (2x2 / 3x3 / infinite neighborhood) performed better than prediction based on a sole monomer, while no such improvement was observed for Pepino Mosaic Virus (PepMV) and Bamboo Mosaic Virus (BaMV) where all methods had similar sequency recovery rates. These results suggest that for the tested proteins, the incorrect neighborhood for small crops doesn't lead to an increased call of wrong amino acids in the aggregated logits. The newly developed infinite symmetry approach performs en par with 2x2 or 3x3 neighborhood prediction, but lowers the amount of required compute to that of a single monomer. However, it is to note that compute cost is generally not a concern when running ProteinMPNN due to its low complexity.

ProteinMPNN was used to generate sequences based on the wildtype backbone structure of PVX using the introduced Graph Reduction technique to model an infinite symmetry and a sampling temperature of $\tau \to 0$, e.g. argmax sampling. Sequences were generated with varying bias b towards the wildtype sequence, that is by increasing the logit of the residue that's present in the wildtype structure by b before sampling the amino acid. For each of the bias values $b \in \{0, 1, 2, 2.5\}$, five sequences. The sequence identity of the generated sequences to the wildtype was about 0.54 (bias 0), 0.73 (bias 1), 0.88 (bias 2) and 0.94 (bias 2.5). The generated sequences were further analyzed as described in the section 5 and section 6 before

selecting some for experimental evaluation.

4 Backbone Design with RFdiffusion

RFdiffusion is a generative machine learning model for protein backbone design. It can be run in different modes to accomplish several tasks such as unconditional monomer generation, protein binder design, scaffolding around a fixed motif, or design of symmetric oligomers (Figure 3), the latter being the most relevant for this work. In practice, RFdiffusion is commonly used together with ProteinMPNN, where RFdiffusion generates synthetic backbone structure and ProteinMPNN tries to realize this backbone with a synthetic amino acid sequence.

Generation by RFdiffusion is performed through a reverse Riemannian diffusion process on the manifold SE(3). Compared to other diffusion-based algorithms like AlphaFold3 (section 5), RFdiffusion doesn't operate on the atom coordinates using standard euclidean diffusion, but diffuses the backbone transforms instead. However, it converts the transforms from and to atom coordinates in each iteration. For unconditional generation, the model starts with randomly initialized backbone coordinates and creates a based solely on a specified number of residues. For the creation of symmetric oligomers, the user specifies a set of transforms $\mathfrak{R} = \{R_k\}_{k=1}^{K} \in \mathrm{SO}(3)$ that define the symmetry. The final protein will satisfy $x^{(k)} = R_k x^{(1)}$, where $x^{(k)}$ denotes the coordinates of the k-th monomer. This is achieved by explicitly setting the coordinates as such after initialization and in each further iteration (1).

Algorithm 1 Generation of symmetric oligomers

def SampleSymmetric($M, \mathfrak{R} = \{R_k\}_{k=1}^{K}$):

\# RFdiffusion generation of oligomer with symmetry $\mathfrak{R}$

1: $x^{(T,1)} = \text{SampleReference}(M)$

2: **for all** $t = T, \ldots, 1$ **do**

\# Symmetrize chains

3: $X^{(t)} = [R_1 x^{(t,1)}, \ldots, R_K x^{(t,1)}]$

4: $\hat{X}^{(0)} = \text{RFdiffusion}(X^{(t)})$

5: $[x^{(t-1,1)}, \ldots, x^{(t-1,K)}] = \text{ReverseStep}(X^{(t)}, \hat{X}^{(0)})$

6: **end for**

7: **return** $\hat{X}^{(0)}$

In the original RFdiffusion code, only point group symmetries are supported, that is symmetries that satisfy $\{x^{(1)}, ..., x^{(K)}\} = \{R_j x^{(1)}, ..., R_j x^{(K)}\}$ for each R_j, up to reordering of the monomers. Technically, the code could work on general euclidean transforms $T_j \in \mathrm{SE}(3)$ (such as the transforms from section 2 specifying the symmetry of PVX) as well. However, there are certain drawbacks in doing so. The positions in early steps in the diffusion process follow a Gaussian distribution. While the rotations by point group symmetries conserve that distribution, general euclidean transforms don't. This means that the positions that are fed into the noise

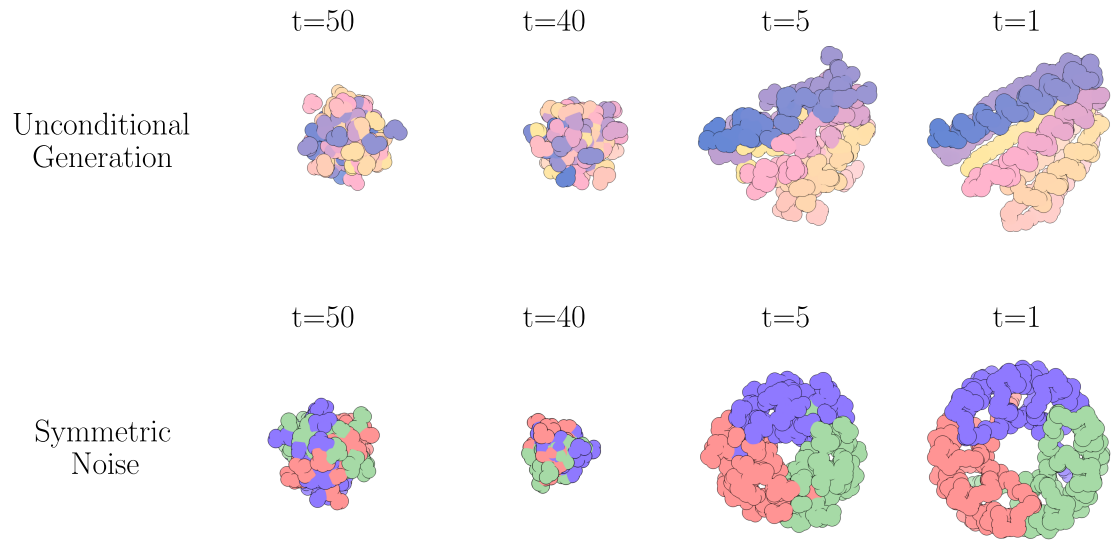


Figure 3: RFdiffusion trajectories for unconditional generation and symmetric noise. For unconditional generation, RFdiffusion samples the initial positions independently from a Gaussian distribution and generates the protein without any constraints. For the generation of oligomers with a specific symmetry, RFdiffusion only samples coordinates for one monomer and initializes the other coordinates by applying the respective symmetry transform to the coordinates of that reference monomer. In each diffusion step, this is repeated to enforce the symmetry.

prediction network in the diffusion process don't follow the distribution the model is trained on. Further, the authors of RFdiffusion observed that for point group symmetries, the noise prediction model conserves the symmetry almost perfectly. Due to this, the explicit symmetrization in each iteration is generally not necessary and barely affects the trajectory, if the initial noise is symmetrized. This arises from the equivariance of the SE(3)-transformer architecture used in RFdiffusion. Using the symmetry transforms for PVX as evaluated in section 2, the atom positions in early stages of the diffusion process don't follow a Gaussian distribution, and same-seeded trajectories using either full symmetry enforcement or only initially symmetrized noise differ strongly (Figure 4). Rather, for initial symmetrization, the atoms quickly collapse to a Gaussian-like distribution, before spreading out again to form the final multimer.

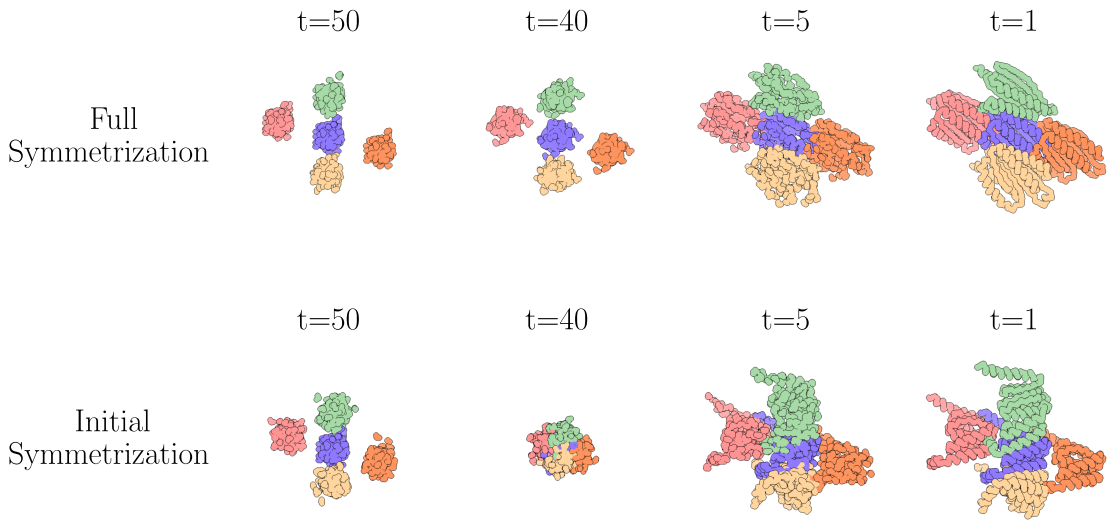


Figure 4: RFdiffusion trajectories for full and initial symmetry enforcement.

RFdiffusion with full symmetry enforcement of the PVX symmetry was used to generate backbone structures for further testing. In total, five different backbone structures were designed, and ProteinMPNN was used as described in section 3 to generate five sequences for each of them. Additionally, backbone structures were generated using partial denoising, where only a limited amount of noise was added to the wild type structure before denoising again. This was done using 5, 10, 15, and 20 noise steps in RFdiffusion. For each of these noise levels, three denoised backbones were generated using RFdiffusion, each realized by three sequences through ProteinMPNN. The de novo generated backbones typically consist of a simple structure of alpha helices and beta sheets (Figure 5). Possibly due to the aforementioned incongruities in running the algorithm with euclidean transforms, backbone structures generated for the PVX symmetry tended to have structural violations, in particular interchain clashes. The sequences were further evaluated using the methods described in section 5 and section 6.

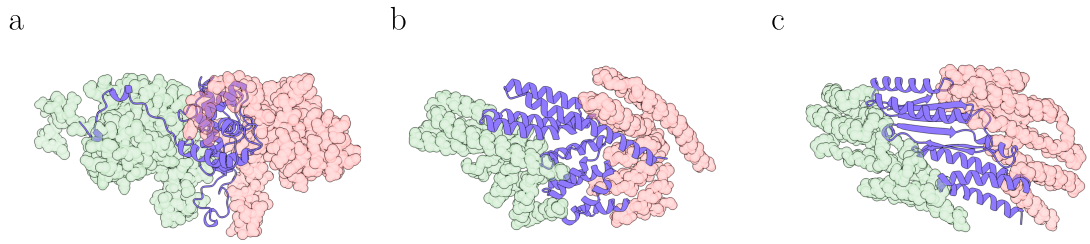


Figure 5: RFdiffusion examples.

5 Evaluation with AlphaFold

Despite outstanding performance of RFdiffusion and ProteinMPNN for de novo protein design, the success rate is often too low for a small-scale experimental evaluation. In experiments by Watson et al. [24], for the task of symmetric oligomer designs, 87 out of 608 designs showed an oligomerization state consistent with the design models. In light of this, methods for in silico assessment of protein designs were established to improve the chances for successful designs.

For the task of binder design, Bennett et al. managed to increase the success rate of binder design nearly 10-fold using metrics based on AlphaFold 2 [3]. In their design assays, a low Cα RMSD and a low predicted aligned error (pAE) between inter-chain residue pairs was predictive of binder success. Unfortunately, AlphaFold 2 fails to predict the multimeric structure of the wild type. Even using the "AF2 initial guess" method [3] of providing the expected backbone structure to the model through the recycling embedder was unsuccessful in recovering the prediction.

The recently developed model AlphaFold 3 [2] performs better on the prediction of the wild type, but still only makes a prediction with Cα RMSD of 5 Å or less in 8 % of the evaluations (Figure 7). Since AlphaFold 3 is indeterministic, repeated runs result in different outcomes. However, the architecture for structure prediction in AlphaFold 3 is largely different from AlphaFold 2, replacing the structure module with a diffusion algorithm. As seen in section 4, diffusion algorithms allow for changes to the denoising process to guide the prediction, such as a symmetry constraint.

While both RFdiffusion and AlphaFold 3 use diffusion, their exact implementations vary, requiring additional considerations when transferring the symmetrization process used in RFdiffusion to AlphaFold 3. In particular, the diffusion trajectories in AlphaFold 3 are not scaled to unit variance and the model changes its position and orientation throughout the process (Figure 6). This motion of the model can be tracked by using a reference frame $T_{\text{ref}} = (R_{\text{ref}}, \vec{\mathbf{t}}_{\text{ref}})$ and enforcing the symmetry in that frame as

$$\vec{\mathbf{x}}^{(j)} = T_{\text{ref}} \circ T_j \circ T_{\text{ref}}^{-1} \circ \vec{\mathbf{x}}^{(1)} \tag{9}$$

Motion of the model happens in three stages of the diffusion sampler in AlphaFold 3: First, the function CentreRandomAugmentation recenters the prediction before applying a random rotation and translation to the model. Second, Gaussian noise is added to the model in each iteration, potentially shifting it. Third, the prediction by the denoiser can be shifted, resulting in a translation of the model when applying the denoising step. These motions do not occur in RFdiffusion. Since the algorithm is SE(3) invariant, it does not require augmentation. No noise is added, and the prediction is aligned to the current model before updating it.

To account for this, the motion by the function CentreRandomAugmentation can be applied to the reference frame T_{ref} as well, and shifts to the model can be considered by setting the translation $\vec{\mathbf{t}}_{\text{ref}}$ to the center of the reference monomer in each iteration. Further, the prediction by the denoiser can be shifted to match the center of the current model before applying the symmetry. The details of the implementation are outlined in Supplementary Algorithm S2. In this work, the symmetry constraint was applied to the initial noise and the denoised prediction. Symmetrization of the current model at the start of each iteration, as done in RFdiffusion, is likely to be similarly effective. The denoised prediction could also lead to a slight

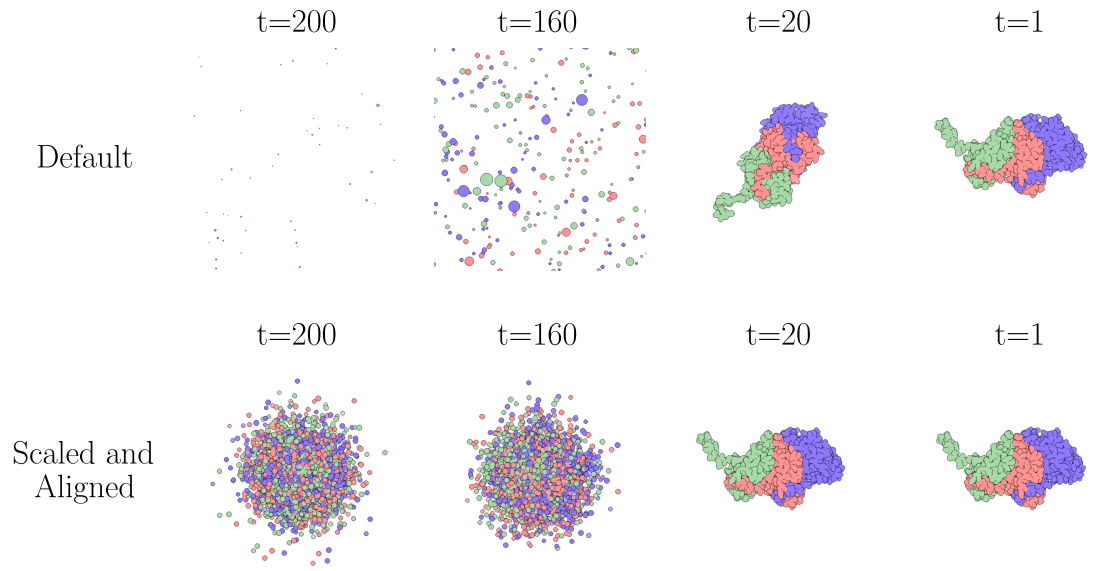


Figure 6: Diffusion Trajectories of AlphaFold3 with symmetry enforcement.

rotation of the model over time. This can be accounted for by rotating the reference frame towards the best alignment of the current model with the expected backbone coordinates, but has little effect on the accuracy. Notably, the atom coordinates in early steps of the diffusion process have a standard deviation that is significantly larger than the translation in the symmetry transforms of PVX. Due to this, symmetrization in AlphaFold 3 does not strongly affect the point distributions, as it did for RFdiffusion (section 4).

The described process of symmetrization during diffusion can in fact often recover the prediction. For a standard AlphaFold pass, including a Multiple Sequence Alignment (MSA) of the query sequence, the symmetry-guided prediction scores an RMSD less than 5 Å in 50 % of the runs, while the original model only reached 5 Å in 8 % of the evaluations (Figure 7). For designs created with RFdiffusion, sequences often have low similarity with sequences from known databases, so there is little to no MSA data available. In the case of MSA-free prediction, symmetrization is unfortunately unable to improve the prediction. This might pose a problem for the RFdiffusion designs. However, the sequences created based on the wild type structure of PVX have sufficient similarity to build meaningful MSAs.

Using AlphaFold with symmetrization and the described tracking of the symmetry reference frame to the expected orientation, all sequences generated in section 3 and section 4 were evaluated based on the Cα-RMSD between their designed backbone and the AlphaFold prediction. The designs from partial diffusion in section 4 all resulted in high RMSDs larger than 15 Å and were not further analyzed. Two of the RFdiffusion designs and three of the pure ProteinMPNN designs were chosen for investigation with GROMACS, the choice being based on their low RMSD score (Table 2). While the pure ProteinMPNN designs show little amount of structural violations, the RFdiffusion designs have a substantial number of interchain and intrachain clashes, as computed with ChimeraX. Aside of its role in filtering the sequence designs, the AlphaFold prediction was also used as the initial structure for the GROMACS simulations in section 6.

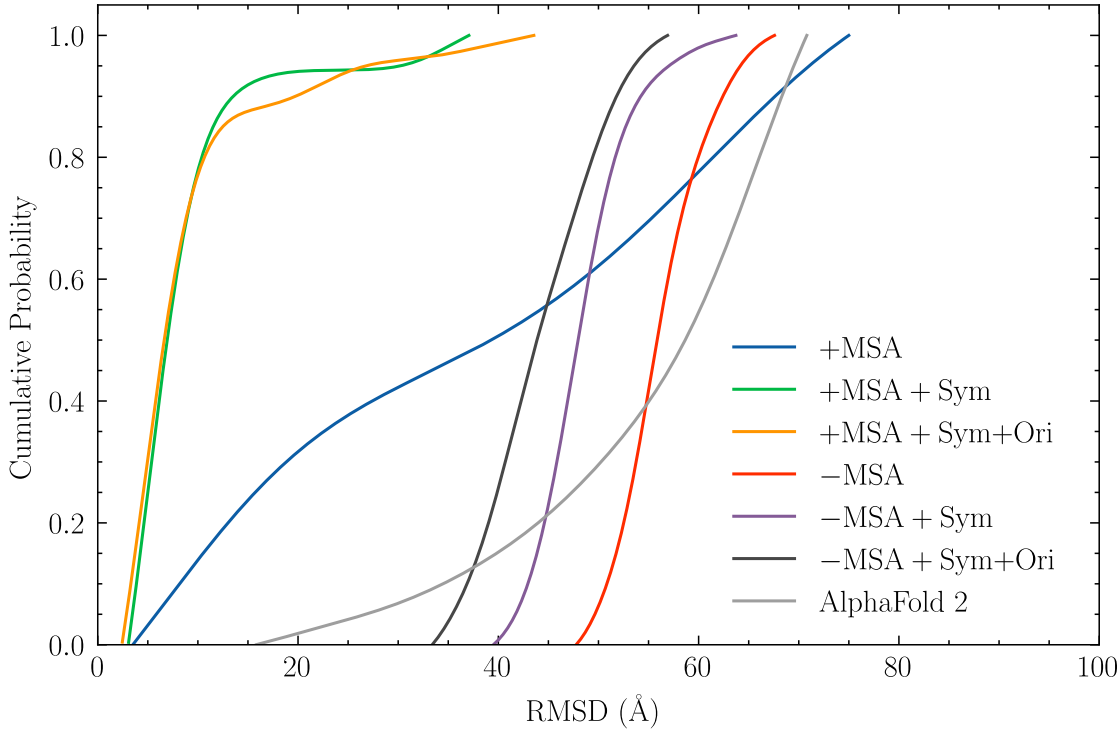


Figure 7: Comparison of Cα RMSD of AlphaFold 3 on PVX using different variants of the algorithm.

Table 2: Comparison of designs from different sources

Design	**Source**	**RMSD**	**Clashes (intra|inter)**	**Sequence Identity**
Design A	WT P-MPNN Bias 2	0.97	(1|4)	90%
Design B	WT P-MPNN Bias 0	0.67	(1|4)	53%
Design C	WT P-MPNN Bias 2.5	0.75	(2|4)	92%
Design D	RFdiffusion	3.74	(216|1012)	8%
Design E	RFdiffusion	2.94	(504|3148)	10%

6 Evaluation with GROMACS

For further in silico evaluation of the designs, Molecular Dynamics (MD) simulations using GROMACS [1] were conducted. Similar physics based evaluations have already been proven to be a suitable metric for assessing the quality of novel binder designs [5]. While physics based metrics tended to be less effective for design evaluation than methods based on AlphaFold [3], the metric might be particularly viable for the designs created in this work, since the methods of symmetry-guided design (section 4) and symmetry-guided prediction (section 5) might introduce a bias for the AI tools.

For viruses, all-atom MD simulations are widely used for several tasks regarding structural analysis and assembly [15]. In particular, Freddolino et al. were able to analyze structural integrity of Satellite Tobacco Mosaic Virus (STMV), showing that the capsid becomes unstable in absence of RNA [7]. This instability was observed as an increase in the RMSD of the viral atoms compared to the initial structure over the course of the simulation.

A similar trend was observed in this work as well. In MD simulations of a slice of 13 monomers at a temperature of 310 K (also conducted at 300 Kwithsimilarresults), RMSD for the wild type excluding RNA increased significantly quicker than for the wild type with its genomic RNA included (Figure 8). Three of the artificial designs showed an RMSD development in-between these two, while the designs based on RFdiffusion quickly rose to RMSDs more than twice as large as that of the wild type without RNA (data not shown), likely due to the observed structural violations in the designs. Notably, when running the simulation of the wild type without RNA based on an initial guess by AlphaFold instead of the PDB structure, the RMSD progresses similar to those of the artificial designs.

In the context of binder design, another metric that proved to be an effective predictor was the Rosetta ddG estimate of the complex's free binding energy [5]. Similar free energy simulations can also be conducted in GROMACS using Umbrella Sampling and the Weighted Histogram Analysis Method (WHAM) [10]. Here, an estimate of the binding energy is calculated by forcing the proteins apart from each other using a moving potential, then running simulations along intermediate steps of the trajectory. The binding energy can be computed from these simulations through estimating the thermodynamical likelihood $P(x)$ of each distance. Concretely, the potential of mean (the free energy along the pulling coordinate) can be computed as

$$F(x) = -k_B T \log(P(x)) + C \tag{10}$$

where k_B is the Boltzmann constant, T is the temperature, C is a constant offset, and $P(x)$ is the probability distribution over the distance x between the monomers. The free binding energy is then the difference of the potential of mean force in the unbound and the bound state.

This calculation of the free binding energy was conducted for the wild type and three of the designs (Figure 9). The free energy estimate for Design A of 165 k/mol was slightly higher than that of the wild type of 160 k/mol. Designs B and C were evaluated to lower binding energies of 120 k/mol and 115 k/mol. Calculation of the free energy based on the predicted initial structure of the wild type instead of the PDB structure lead to a binding energy of only 105 k/mol. This is significantly lower than the value based on the PDB structure, even though the two models only have

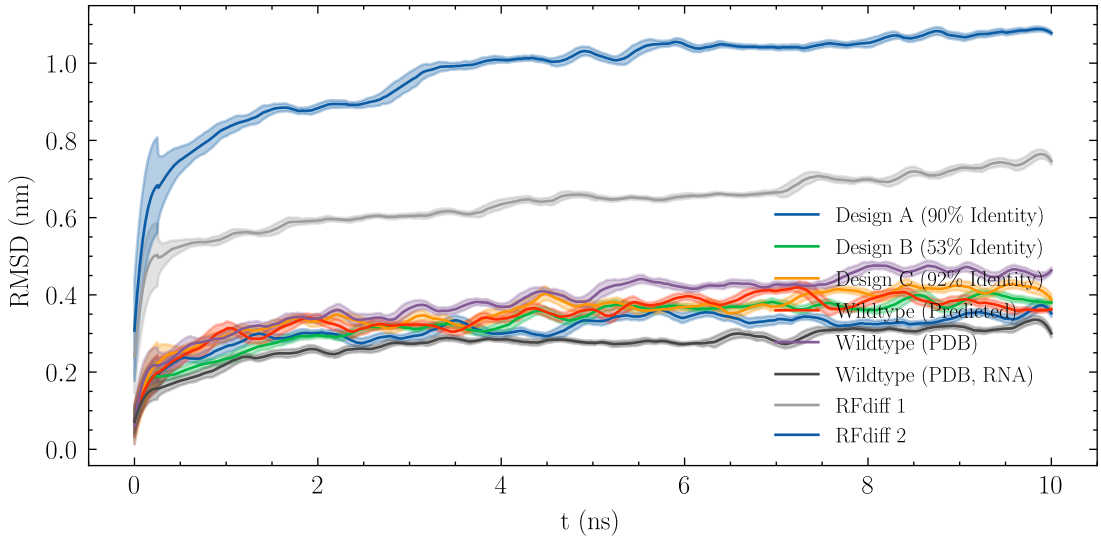


Figure 8: RMSD and Gyrate Plots

an RMSD of 0.9 Å. This suggests a strong dependence of the calculation on the initial state of the system. The free energy calculations were conducted at relatively high pulling rates, which is generally discouraged since it hinders equilibration of the system and estimation of the physically optimal binding trajectory. The high pulling rates were chosen because the proteins tended to unravel in the simulations using lower pulling rates, and the required simulation size for full dissociation would have exceeded the computational resources available for this project. As a result, the estimated binding energies should be interpreted with caution, as they may not fully reflect the true thermodynamic values.

Based on the simulations, designs A and B were chosen for experimental evaluation in the wet lab. As mentioned earlier, the structural design was based on the d29-CP-PVX, since the structure of the flexible tail is not determined. Given the relevance of the N-terminal domain for the wild type structure of PVX [4], both designs were prepended with the N-terminal domain of the wild type. Design B only shares 53 % sequence identity with PVX-CP. Due to this, anti-PVX antibodies might not bind to this design. Based on experience from the Institute of Molecular Biotechnology, an S-Tag was chosen as an N-terminal marker, because it was known to not hinder assembly of PVX, and because it is detectable by an anti-S-Tag antibody. The three designs chosen for wet lab evaluation are thus S-Tag-A, S-Tag-B, and d29-A, the latter being the sequence of design A without any N-terminal modifications. The protein sequences were codon optimized for use in *Nicotiana tabacum* as a host organisms. The designed protein and DNA sequences are described in section 10.2.4.

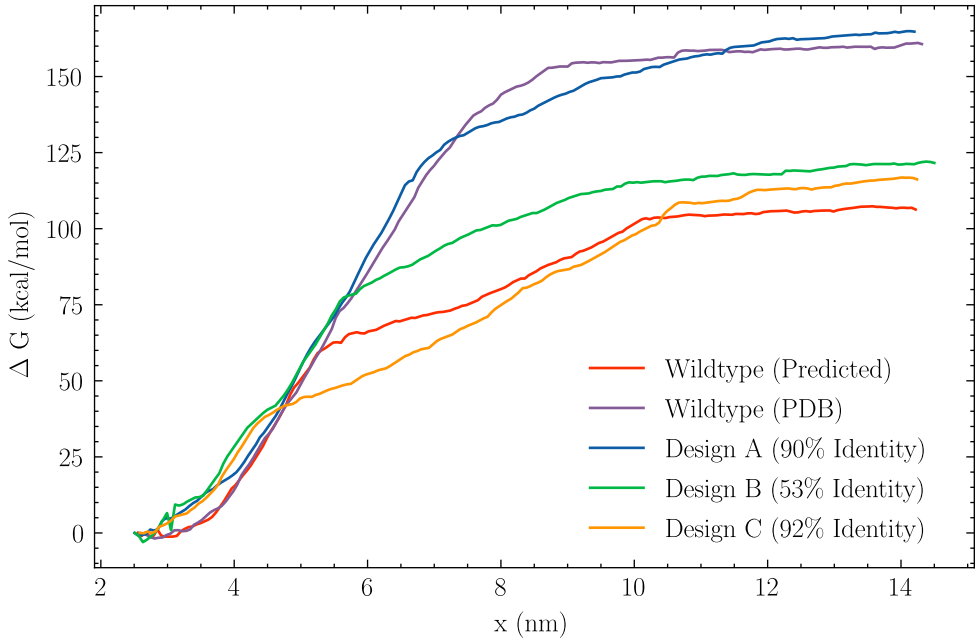


Figure 9: Potential of Mean Force

Part II
Experimental Evaluation

7 Materials

7.1 Laboratory Equipment

7.2 Chemicals

7.3 Media, Buffers, and Solutions

All media, buffers, and solutions that were used in the experiments are shown in Table 3. NaOH and HCl were used to establish the pH.

Table 3: Media, buffers, and solutions that were used in this study. Listed are the component types their respective amounts.

Medium/buffer/solution	Component	Amount
agarose gel	Agarose	1.2 % (w/v)
	Ethidium bromide	$0.5\,\mu g\,mL^{-1}$
	in 1x TAE buffer	
AP buffer (pH 9.6)	Tris-HCl	100 mM
	NaCl	100 mM
	$MgCl_2$	5 mM
blocking solution	powedered milk	$40\,g\,L^{-1}$
	in PBS buffer	

Medium/buffer/solution	**Component**	**Amount**
Coomassie staining solution	Coomassie Brilliant Blue	$2.5\,\mathrm{g\,L^{-1}}$
	Methanol	50 % (v/v)
	Acetic acid	10 % (v/v)
coating buffer (pH 9.6)	—	—
destaining solution	Methanol	5 % (v/v)
	Acetic acid	7.5 % (v/v)
elisa substrate solution	—	—
LB medium	Yeast extract	$5\,\mathrm{g\,L^{-1}}$
	Tryptone	$10\,\mathrm{g\,L^{-1}}$
	NaCl	171 mM
	(Agar)	$15\,\mathrm{g\,L^{-1}}$
LB-Amp medium	Yeast extract	$5\,\mathrm{g\,L^{-1}}$
	Tryptone	$10\,\mathrm{g\,L^{-1}}$
	NaCl	171 mM
	Ampicillin	$100\,\mathrm{mg\,L^{-1}}$
	(Agar)	$15\,\mathrm{g\,L^{-1}}$
NBT/BCIP staining solution	NBT	$33.3\,\mathrm{g\,L^{-1}}$
	BCIP	$16.5\,\mathrm{g\,L^{-1}}$
	in dimethylformamide	
PBS buffer	NaCl	137 mM
KCl	2.7 mM	
Na_2HPO_4	8.1 mM	
KH_2PO_4	1.5 mM	
phosphate buffer (0.01 M, pH 7.2)	Na_2HPO_4	6.67 mM
	NaH_2PO_4	3.33 mM
reducing loading buffer (5x)	Tris-HCl (pH 6.8)	62.5 mM
	Glycerine	$300\,\mathrm{g\,L^{-1}}$
	SDS	$40\,\mathrm{g\,L^{-1}}$
	β-Mercaptoethanol	10 % (v/v)
	Bromophenol blue	$0.1\,\mathrm{g\,L^{-1}}$
Sabu (10x)	Glycine	50 % (v/v)
	Xylene cyanol FF	0.1 % (w/v)
	Bromophenol blue	0.1 % (w/v)
	in 1x TAE buffer (pH 8.0)	
SDS running buffer	Glycine	200 mM
	SDS	$1\,\mathrm{g\,L^{-1}}$
	Tris	25 mM
semi-dry transfer buffer	Glycine	391 mM
	Methanol	20 % (v/v)
	Tris	48 mM
TAE buffer	Tris	40 mM
	Acetic acid	0.1 % (v/v)

Medium/buffer/solution	Component	Amount
	EDTA (pH 8.0)	1 mM

7.4 Reaction Kits

The following reaction kits were used in this study:

- Gibson Assembly® Cloning Kit, NEB, Ipswich, MA
- Mix2Seq Kit, Eurofins Genomics, Ebersberg, Germany
- PureYield™ Plasmid Miniprep System, Promega, Madison, WI
- Wizard® SV Gel and PCR Clean-Up System, Promega, Madison, WI
- NucleoBond Xtra Midi kit for transfection-grade plasmid DNA, Macherey-Nagel, Düren, Germany
- ALiCE® for Research kit, LenioBio, Düsseldorf, Germany

7.5 Enzymes

All enzymes used in this study are shown in Table 4.

Table 4: Polymerases and restriction enzymes that were used in this study. Listed are the reagent names, the manufacturer, and how the enzymes were used in the experiments.

Reagent	Manufacturer	Use
GoTaq® G2 DNA Polymerase	Promega	Colony PCR
Pfu DNA Polymerase	Promega	Insert Amplification
NcoI-HF®	NEB	Restriction
KpnI-HF®	NEB	Restriction

7.6 Plasmids

7.6.1 pLenEx-Strep-eYFP

The plasmid pLenEx-Strep-eYFP was used for cell-free expression of Strep-eYFP using the ALiCE® expression kit. The plasmid contains an origin of replication for *E. coli*, a gene for ampicillin resistance, and a gene encoding Strep-eYFP. This gene is flanked by 5'- and 3'-UTRs from tobacco mosaic virus (TMV) and under the T7 promoter. The plasmid contains restriction sites for NcoI at the 5' end of Strep-eYFP and for KpnI at the 3' end. In addition to cell-free expression, pLenEx-Strep-eYFP was used for cloning of the genetic elements d29-A, S-Tag-A, and S-Tag-B. The plasmid was provided by the Institute for Molecular Biotechnology.

7.6.2 pLenEx-d29-A, pLenEx-S-Tag-A, pLenEx-S-Tag-B

Like pLenEx-Strep-eYFP, the plasmids pLenEx-d29-A, pLenEx-S-Tag-A, and pLenEx-S-Tag-B contain an origin of replication for *E. coli* and a gene for ampicillin resistance. Instead of Strep-eYFP, the vectors carry the genetic elements d29-A, S-Tag-A, and S-Tag-B, as described in section 6, respectively. These genes are again flanked by 5'- and 3'-UTRs from TMV and are under control of the T7 promoter and in-between the NcoI and KpnI restriction sites. The plasmids were created based on pLenEx-Strep-eYFP using Gibson Assembly.

7.6.3 pLenEx-CP-PVX

The plasmid pLenEx-CP-PVX carries the same genetic elements as pLenEx-Strep-eYFP, except for the Strep-eYFP gene, which is instead replaced by a gene encoding for the PVX coat protein. The plasmid was provided by the Institute for Molecular Biotechnology.

7.7 Antibodies

All antibodies that were used in this study are listed in Table 5. The antibodies were used both for Western Blotting and for ELISA assays.

Table 5: Antibodies that were used in this study. Listed are the characteristics of the antibody and its manufacturer.

Antibody	Type
Rabbit-Anti-PVX	Polyclonal rabbit IgG
S-peptide Epitope Tag Monoclonal Antibody	Monoclonal mouse IgG
Goat-Anti-Rabbit ($GAR^{AP}{}_{FC}$)	Polyclonal goat IgG, alkaline phosphatase cor
Goat-Anti-Mouse ($GAM^{AP}{}_{FC}$)	Polyclonal goat IgG, alkaline phosphatase cor

7.8 Synthetic Oligonucleotides

All syntehtic oligonucleotides that were used in this study are shown in Table 6.

Table 6: Synthetic oligonucleotides that were used in this study. Regions overlapping with the amplified genes are displayed in uppercase.

Name	Sequence (5' → 3')	Use
d29-A fwd	acattttacattctacaactaccATGGCTT CTGGCTTATTCACCATACCTG	Insert amplification
d29-A rev	ccaaaccagaagagcttggtaccTTAAGGG GGGGGAATGGTCAC	Insert amplification
S-Tag-A fwd	acattttacattctacaactaccatggcTA AAGAAACAGCCGCCGCTAAATTC	Insert amplification
S-Tag-B fwd	acattttacattctacaactaccatggctA AGGAGACTGCTGCAGCCAAG	Insert amplification
S-Tag-B rev	ccaaaccagaagagcttggtaccTTAAGCT GCGGGTATGTGTATGATTC	Insert amplification
pLenSeq fwd	% TODO: Ask Juliane, sequence missing	Sequencing
pLenSeq rev	% TODO: sequence missing	Sequencing

7.9 Synthetic Genes

Genes for the constructs S-Tag-A and S-Tag-B were ordered from Integrated DNA Technologies (Coralville, Iowa). The exact sequences can be found in section 10.2.4.

7.10 Organisms

NEB® Turbo Competent *E. coli* (High Efficiency) was used for cloning of the plasmids pLenEx-d29-A, pLenEx-S-Tag-A, and pLenEx-S-Tag-B.

8 Methods

8.1 DNA Cloning

8.1.1 PCR Amplification of Insert DNA

Using the synthetic genes S-Tag-A and S-Tag-B (subsection 7.9) and the primer pairs d29-A fwd / d29-A rev (with template S-Tag-A), S-Tag-A fwd / d29-A rev (with template S-Tag-A), and S-Tag-B fwd / S-Tag-B rev (with template S-Tag-B), polymerase chain reactions (PCRs) were conducted for amplification of the inserts. The primers also introduced overlapping regions with the plasmid pLenEx-Strep-eYFP for the following Gibson Assembly, and in the case of d29-A fwd, a start codon. The PCR was conducted with a Pfu polymerase. The composition of the PCR mix is listed in Table 7. The PCR reaction was conducted in a thermocycler. The exact program is described in Table 8. After running the PCR, the products were frozen at −20 °C until further processing.

8.1.2 Plasmid Restriction Digest

The plasmid pLenEx-Strep-eYFP was digested using the restriction enzymes NcoI-HF and KpnI-HF. Therefore, 10 µg of the plasmid, 5 µL CutSmart® Buffer, 1 µL of

Table 7: Composition of the PCR Mix for Insert Amplification.

Component	Amount
MQ-H_2O	38 µL
10x Pfu Buffer	5 µL
dNTPs (10 mM)	2 µL
Forward Primer (10 µM)	2 µL
Reverse Primer (10 µM)	2 µL
Pfu DNA Polymerase (3 U µL^{-1})	0.5 µL
DNA template (10 ng µL^{-1})	5 µL

Table 8: PCR Program used for Insert Amplification.

Step	Temperature (°C)	Time
Initial Denaturation	94	4 min
Repeat for 30 cycles:		
Denaturation	94	30 s
Annealing	57	30 s
Elongation	72	100 s
Final Elongation	72	5 min
Storage	8	∞ (hold)

NcoI-HF (20 U µL^{-1}), and 1 µL of KpnI-HF (20 U µL^{-1}) were added to MQ-H_2O up to a total volume of 50 µL. The mixture was incubated at 37 °C for three hours and thereafter frozen at −20 °C until further processing.

8.1.3 Agarose Gel Electrophoresis and DNA Recovery

Agarose gel electrophoresis was used to validate and purify the restricted plasmid and the PCR products. 50 µL of each sample were mixed with 5 µL 10x Sabu and applied to the gel, splitting the sample into 25 µL per track. The gels were run at either 100 V for about 45 minutes or at 120 V for about 60 minutes, depending on the size of the gel. After the gel ran through, the samples were cut out under illumination from a UV light source. The DNA was recovered from the gel samples using the Wizard® SV Gel and PCR Clean-Up System. The steps were carried out following the manufacturer's protocol. DNA concentrations of each sample were determined using the NanoDrop™ One.

8.1.4 Gibson Assembly

A Gibson Assembly using 50 ng of purified linear vector DNA was conducted to create the plasmids pLenEx-d29-A, pLenEx-S-Tag-A, and pLenEx-S-Tag-B. The required mass of insert DNA was calculated using Equation 11.

$$\text{Mass}_{\text{insert}} = \left(\frac{\text{desired molar ratio}}{1}\right) \times \text{Mass}_{\text{vector}} \times \left(\frac{\text{Length}_{\text{insert}}}{\text{Length}_{\text{vector}}}\right) \tag{11}$$

The molar ratio was chosen as 2. Given the plasmid length of 2173 bp, and the insert lengths of 679 bp, 802 bp, and 802 bp for the three inserts d29-A, S-Tag-A, and S-Tag-B respectively, the insert DNA masses were calculated as 31.2 ng for d29-A and 37 ng for S-Tag-A and S-Tag-B. The plasmid and inserts and 10 µL Gibson Assembly® Master Mix were added to MQ-H_2O to a total volume of 20 µL and incubated at 50 °C for one hour.

8.1.5 Transformation into Competent Cells

Transformation of the assembled plasmids into NEB® Turbo Competent *E. coli* cells was carried out using the manufacturer's High Efficiency Transformation Protocol, using a reduced volume of cells compared to the original protocol.

After thawing the cells on ice for 10 minutes, the cells were gently mixed and 20 µL of the cells were transferred to a reaction tube on ice. 2 µL of the Gibson Assembly product were added to the cell mixture and carefully flicked to mix cells and DNA. The mixture was placed on ice for 30 minutes. Afterward, a heat shock at exactly 42 °C was conducted for 30 seconds. The cells were thereafter placed on ice for 5 minutes. Afterward, 950 µL of room temperature salt medium was added to the mixture. The cells were rotated at 37 °C for 60 minutes, during which LB-Amp selection plates were warmed to 37 °C. The cells were mixed thoroughly by flicking the tube and inverting it. Afterward, 100 µL were applied to a selection plate and incubated overnight at 37 °C.

8.1.6 Colony PCR

After incubation overnight, 9 colonies of each construct were picked for Colony PCR. The master mix for the PCR was created following the composition in Table 9.

Table 9: PCR Master Mix Composition

Component	Amount
MQ-H_2O	15 µL
5x Green GoTaq® Reaction Buffer	4 µL
dNTPs (10 mM)	0.4 µL
Forward Primer	0.2 µL
Reverse Primer	0.2 µL
Taq DNA Polymerase	0.2 µL

The colonies were picked using sterile toothpicks and applied to an LB-Amp reference plate. Afterward, they were placed into the PCR tubes for about 30 seconds. The reference plate was incubated at 37 °C overnight. The PCR was run using the program from Table 8 and was followed by an agarose gel electrophoresis as previously described for analysis of the PCR products.

8.1.7 Plasmid Mini-Preparation and Sequencing

For each construct, two clones showing successful amplification during the colony PCR were selected for a plasmid mini preparation and sequencing. 6 mL of LB-AMP medium were inoculated and incubated overnight at 37 °C.

The next day, the culture was centrifuged over three rounds of 2 mL each for 3 min at 6500 ×g, disposing of the supernatant. The pellet was fully resuspended by vortexing with 600 µL of MQ-H_2O. Then, 100 µL of cell lysis buffer were added and the reaction tube was carefully inverted for mixing. Afterward, 350 µL of neutralization solution were added, and the tube was inverted until a full color change to yellow occurred.

The solution was centrifuged for 10 min at 21 300 ×g. Afterward, the supernatant was pipetted onto a PureYield™ mini column. The column was centrifuged for 15 s at 21 300 ×g, and the flow-through was discarded.

200 µL of endotoxin removal wash were added to the column, followed by centrifugation for 15 s at 21 300 ×g, again discarding the flow-through. Then, 400 µL of column wash solution were added to the column, followed by centrifugation for 30 s at maximum speed. The flow-through was discarded.

The column was transferred to a new reaction tube. Elution was performed using 30 µL of nuclease-free water. After application of the water and incubation for 1 min at room temperature, the column was centrifuged for 15 s at 21 300 ×g. The concentration of the plasmid in the flow-through was determined using the NanoDrop™ One.

The purified plasmids were diluted to a final concentration of 100 ng μL^{-1}. In a Mix2Seq kit, 5 µL of the plasmid was mixed with 5 µL of pLenSeq fwd (10 µM) in one compartment and with 5 µL of pLenSeq rev (10 µM) in another compartment. The kit was sent for sequencing.

8.1.8 Plasmid Midi-Preparation

After successful sequencing, the clones were used in a midi-preparation for use of the plasmids in cell-free protein expression. As recommended by the cell-free expression kit's manufacturer, the Macherey-Nagel® NucleoBond® Xtra kit was used for the preparation.

The midi-preparation was conducted following the manufacturer's protocol for low-copy plasmids. Concretely, 200 mL of LB-AMP medium were inoculated and incubated overnight at 37 °C. The next morning, the cell culture was centrifuged at 4500 ×g at 4 °C for 15 min. The pellet was resuspended in 16 mL RES buffer by vortexing. Lysis was conducted in 16 mL of LYS buffer at room temperature for 5 min. Neutralization was performed using 16 mL of NEU buffer.

The NucleoBond column, including the filter, was equilibrated with 12 mL EQU buffer. For clarification of the lysate, centrifugation at 5000 ×g and 4 °C for 10 min was conducted. Some precipitate remained in the reaction tube and was carefully decanted before applying the lysate to the column on top of the column's filter. Washing was done with 5 mL of EQU buffer. Afterward, the filter was removed and washing was repeated using 8 mL of Wash buffer. Elution was conducted using 5 mL of ELU buffer.

For precipitation, 3.5 mL of isopropyl alcohol were added to the mixture. A centrifugation at 5000 ×g and 4 °C for 1 h was conducted. The supernatant was discarded. Then, 2 mL of 70 % ethanol were added, followed by a second centrifugation at 5000 ×g for 5 min at 20 °C. The ethanol was carefully removed by pipetting, and the reaction tube was dried for 10 min.

Afterward, the DNA was resuspended in 500 µL MQ-H_2O. The concentration was determined using the NanoDrop™ One.

8.2 Protein Expression and Purification

8.2.1 Cell-Free Protein Expression

Cell-Free protein expression of the constructs d29-A, S-Tag-A, and S-Tag-B was conducted using an ALiCE® for Research kit. Before starting the reaction, the plasmids pLenEx-d29-A, pLenEx-S-Tag-A, pLenEx-S-Tag-B, pLenEx-CP-PVX, and pLenEx-Strep-eYFP (all from purification using a NucleoBond Xtra Midi kit) were concentrated using speed vacuuming at 30 °C to concentrations between 1400 ng µL^{-1} and 3300 ng µL^{-1}. The volumes of plasmid DNA to be used for the cell-free reaction were calculated according to Equation 12. The final plasmid concentration was chosen as 50 nM.

$$V_{\mathrm{DNA}}\,[\mu\mathrm{L}] = \left(L_{\mathrm{plasmid}}\,[\mathrm{bp}] \cdot 618\ \frac{\mathrm{g}}{\mathrm{mol}\cdot\mathrm{bp}}\right) \cdot\ V_{\mathrm{reaction}}\,[\mu\mathrm{L}] \cdot\ \left(\frac{c_{\mathrm{final}}\,[\mathrm{nM}]}{c_{\mathrm{stock}}\,[\mathrm{ng}/\mu\mathrm{L}]}\right) \cdot\ 10^{-6} \tag{12}$$

Before starting the reaction, the 50 µL ALiCE tubes were fully thawed in a heating block at 25 °C. The solution was mixed by inverting the tubes and centrifuged for 5 seconds at 100 ×g to collect the liquid. After centrifugation, the tubes were placed on ice. The lids were perforated with a single hole using a needle (0.9 mm diameter). The appropriate volume of the plasmids pLenEx-d29-A, pLenEx-S-Tag-A, pLenEx-S-Tag-B, pLenEx-CP-PVX, and pLenEx-Strep-eYFP were added to the respective reaction tubes. Additionally, a non-template control was set up by adding 2 µL MQ-H_2O. The reaction was incubated at 25 °C on an Eppendorf ThermoMixer at 700 rpm for 48 h. Afterward, the reaction tubes were placed on ice to stop the reaction, before being frozen at −20 °C.

8.2.2 Protein Purification Using Capto Core 700

Following cell-free protein expression, size exclusion chromatography using Capto™ Core 700 multimodal chromatography resin was conducted to purify large particles. 1 mL Capto™ Core matrix was suspended in 3 mL of 0.1 M phosphate buffer (pH 7.2) within a column and full sedimentation was awaited. 30 µL of the cell-free expression solution were applied to the column and incubated for 5 min. Afterward, the flow-through was collected. The column was cleaned using 3 mL of a solution out of 1.5 mL 30 % isopropyl alcohol and 1.5 mL 1 M NaOH. The column was stored in 20 % ethanol at 4 °C and reused multiple times. The chromatography was used on the samples d29-A, S-Tag-A, S-Tag-B, and PVX-CP.

8.3 Protein Analysis

8.3.1 SDS-PAGE

Samples from cell-free protein expression, both before and after purification with the Capto™ Core 700 column, were used in an discontinuous SDS polyacrylamide gel electrophoresis (SDS-PAGE) for further use with Coomassie Staining and Western Blot.

The composition of the resolving gel and the stacking gel are listed in Table **??**. All reagents used for the resolving gel, except for the ammonium persulfate (APS), were mixed together by vortexing. After addition of APS, the solution was shortly vortexed and about 5 mL of the gel were transferred into a 0.75 mm thick chamber for polymerization. Directly afterward, isopropyl alcohol was added to the top of the gel.

After polymerization, the isopropyl alcohol was removed using Whatman paper. The components for the stacking gel were mixed, APS was added, and the stacking gel was poured on top of the resolving gel. A comb was inserted into the stacking gel to create sample pockets.

Table 10: Composition of discontinuous SDS-PAGE gels. The amounts are shown for the preparation of two gels.

Component	**Resolving Gel (T = 12%)**	**Stacking Gel (T = 4%)**
MQ-H_2O	2.115 mL	3.645 mL
Tris-HCl stock (1 M)	3.75 mL (pH 8.8)	625 µL (pH 6.8)
AA stock (30%)	4 mL	830 µL
SDS (10%)	100 µL	50 µL
TEMED	10 µL	5 µL
APS (20%)	30 µL	15 µL
Total Volume	10 mL	5 mL

After polymerization of the stacking gel, the gel was either directly used in electrophoresis, or wrapped into wet paper and stored at 4 °C for up to a week.

For electrophoresis, the gel was placed vertically in a chamber containing SDS running buffer. The samples were mixed with 5x reducing loading buffer in a 4:1 ratio and boiled for 5 min. 10 µL sample volume was transferred into the gel pockets, and the marker Color Prestained Protein Standard, Broad Range (10-250 kDa) by New England Biolabs was used as marker. The gel was run at 180 V for about one hour.

8.3.2 Coomassie Staining

After completion of the SDS-PAGE, the gels were placed in Coomassie Staining solution for 30 minutes, while being gently swiveled on an orbital shaker. The Coomassie Staining solution was removed, and destaining solution was added to the gel, still being swiveled. The destaining solution was replaced multiple times, before destaining was complete.

8.3.3 Western Blot

For immunologic detection of specific epitopes on the SDS gel, Western Blotting was conducted. The gel was placed in semi-dry transfer buffer, and eight layers of Whatman paper as well as a nitrocellulose membrane were soaked in semi-dry transfer buffer for 5 min. Then, a stack of four layers of Whatman paper, the nitrocellulose membrane, the gel, and four layers of Whatman paper, was assembled in the blotting chamber of a Trans-Blot® Turbo™ machine. Transfer to the membrane took place at a constant voltage of 25 V and a current of maximally 1 A.

After blotting, the membrane was cut to the size of the gel, placed into 10 mL blocking buffer and incubated for 30 min under swiveling. The blocking buffer was removed, and the membrane was washed three times with PBS buffer, waiting 5 min between each exchange of the buffer. The primary antibody (either Rabbit-Anti-PVX in a 1:5000 ratio or Mouse-Anti-S-Tag in a 1:10000 ratio) was dissolved in 10 mL PBS buffer and added to the membrane. Incubation with the primary antibody was conducted overnight at room temperature or over the weekend at 4 °C. Afterward, the three washing steps were repeated and the secondary antibody was added (either Goat-Anti-Rabbit FC AP or Goat-Anti-Mouse FC AP, both in a 1:5000 ratio in PBS).

8.3.4 ELISA

8.3.5 Electron Microscopy

9 Results

9.1 DNA Cloning

The transformation with the new constructs was validated by sequencing following the mini-preparation. All clones showed 100% sequence identity with the designs. The concentrations of the plasmids after the midi preparation were measured as 605 ng µL^{-1} (d29-A), 335 ng µL^{-1} (S-Tag-A), and 300 ng µL^{-1} (S-Tag-B), corresponding to yields of 302 µg, 167 µg, and 150 µg.

9.2 Protein Analysis

9.2.1 eYFP Yield by Fluorescence

For an estimate of the protein yield in the lysate, a fluorescence analysis of a cell-free expression setup expressiong Strep-eYFP was conducted. The calibration showed a linear trend with a coefficient of determination of $R^2 = 0.98$ (Supplementary Figure S1). Applying the linear model to the sample values, the concentration of Strep-eYFP in the lysate was estimated as 295 µg mL^{-1}.

9.2.2 Coomassie Staining and Western Blot

Coomassie staining and Western Blot were conducted both directly after the cell-free expression and after the following CaptoCore purification. For each setting, two blots were developed, one using an anti-PVX antibody, and one using an anti-S-Tag antibody.

For the samples from cell-free expression (Figure 10), the Coomassie Staining shows a wide range of bands. For the samples from ALiCE setups expressing d29-A, S-Tag-A, and S-Tag-B, the pattern looks very similar to that of the non-template control. The sample expressing Strep-eYFP has a notable band at a height corresponding to 30 kDa, similar to the single band in the Strep-eYFP control. The track containing the S-Tag-CP-PVX control also displays a single band at 29 kDa, slightly lower than that of the Strep-eYFP control. The sample from the ALiCE setup expressing PVX CP presents a band 29 kDa as well.

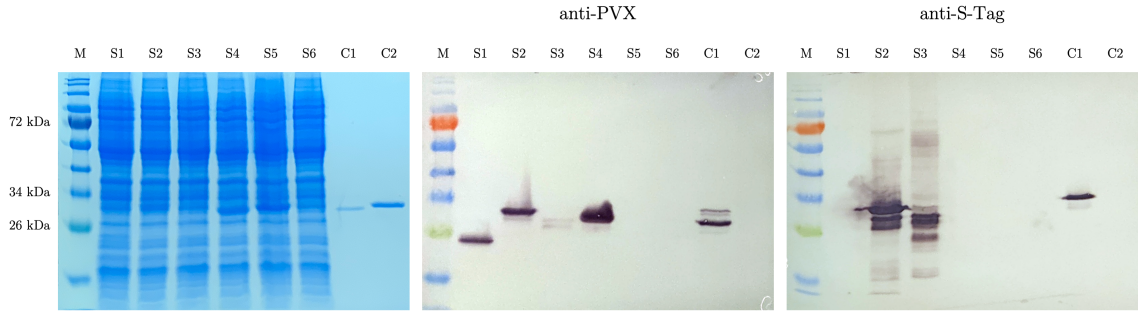


Figure 10: Coomassie Gel and Blot images following ALiCE expression.

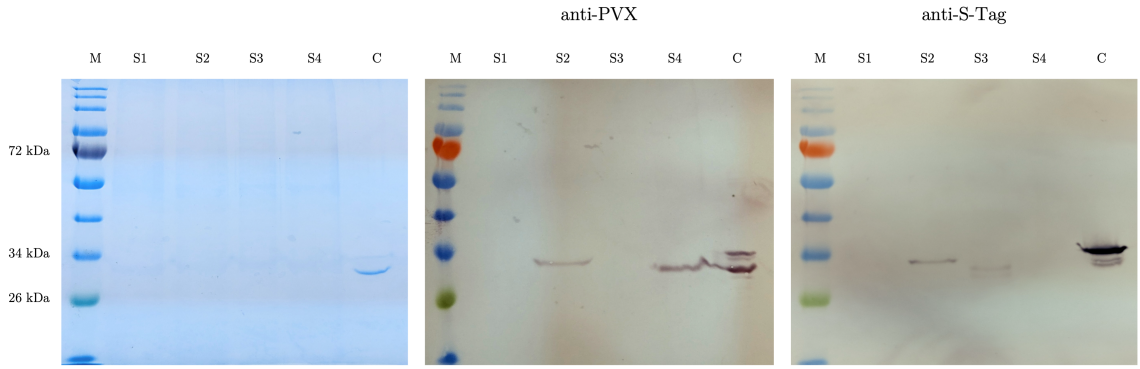


Figure 11: Coomassie Gel and Blot images following CaptoCore Purification.

In the Western Blot based on the anti-PVX antibody, the samples from the ALiCE expressions of d29-A, S-Tag-A, and PVX CP show strong bands at 25 kDa, 31 kDa, and 29 kDa. The band for the expressed PVX CP is particularly strong. The sample from S-Tag-B shows a faint band at 28 kDa. The control S-Tag-CP-PVX has two bands visible in the blot, a stronger band at 28 kDa, and a less prominent one at 30 kDa. In all samples, the observed bands migrated to higher molecular weights than the theoretical molecular weights of 22.6 kDa (d29-A), 26.9 kDa (S-Tag-A), 26.8 kDa (S-Tag-B), 25.1 kDa (CP-PVX), and 26.9 kDa (S-Tag-CP-PVX).

In the blot using the anti-S-Tag antibody, the tracks containing samples from the expression of S-Tag-A and S-Tag-B show a strong signal. For S-Tag-A, the blot shows strong lines at 31 kDa, 29 kDa, and 28 kDa, but with a fainter, diffuse trail to higher and lower molecular weights. In the track containing S-Tag-B, there are prominent lines visible at molecular weights of 29 kDa, 28 kDa, and 25 kDa, as well as a faint, diffuse trail similar to that of S-Tag-A. The control containing S-Tag-CP-PVX shows a strong band at 32 kDa and a faint band at 30 kDa.

The Coomassie Gel of the samples after the CaptoCore chromatography (Figure 11) shows no bands except for the control S-Tag-PVX-CP, which has a band at 30 kDa. The blot using anti-PVX antibodies shows a faint band for the sample of purified S-Tag-A at 31 kDa and for the sample of purified PVX CP at 30 kDa. The control containing S-Tag-PVX-CP particles has bands two bands at 30 kDa and 34 kDa. In the blot using the anti-S-Tag antibody, the sample from S-Tag-A shows a faint line at 31 kDa. The track of the sample from S-Tag-B has a very faint line at 30 kDa. The control S-Tag-PVX-CP has two bands, a strong one at 35 kDa and a fainter one at 33 kDa.

9.2.3 ELISA

For a quantitative analysis of the cell-free protein expression, an ELISA was performed on the samples, using both the anti-PVX antibody, and the anti-S-Tag antibody.

The calibration measurements for the anti-PVX-antibody show saturation at protein concentrations larger than $100\,\mathrm{ng\,mL^{-1}}$. The four calibration samples below that threshold follow a linear trend

$$\mathrm{OD} = 6.5\,\mu\mathrm{L\,ng^{-1}} \cdot c \tag{13}$$

with a coefficient of determination of $R^2 = 0.90$ (Supplementary Figure S2). The measurements from the samples yielded significant OD values for d29-A, S-Tag-A, and CP-PVX, and no significant signal for S-Tag-A and the non-template H^2O from the ALiCE expressions (Supplementary Table S4). For d29-A, S-Tag-A, and CP-PVX, the measured ODs of the different dilutions were almost identical, with an OD of 0.2 for d29-A, 0.22 for S-Tag-A, and 0.87 for CP-PVX. Due to this, the back-calculated concentrations using the linear model and the dilution factors vary over the two dilutions. Calculations based on the stronger diluted samples yield to concentrations of $30(5)\,\mu\mathrm{g\,mL^{-1}}$ for d29-A, $33(4)\,\mu\mathrm{g\,mL^{-1}}$ for S-Tag-A, and $134(4)\,\mu\mathrm{g\,mL^{-1}}$ for CP-PVX.

For the anti-S-Tag antibody, calibration was sub-linear for low concentrations, but followed a linear trend for the whole sample range up to $1500\,\mathrm{ng\,mL^{-1}}$. The linear model was calculated as

$$\mathrm{OD} = 0.9\,\mu\mathrm{L\,ng^{-1}} \cdot c \tag{14}$$

and yielded a coefficient of determination of $R^2 = 0.98$ (Supplementary Figure S2). As for the anti-PVX-antibody, the sample OD measurements were highly similar for the 1:500 and 1:1000 dilutions (Supplementary Table S5), with significant ODs of 2.4/2.2 (1:500/1:1000) for S-Tag-A and 1.5/1.1 (1:500/1:1000) for S-Tag-B. The samples of d29-A, CP-PVX, and the non-template control showed no significant absorbtion. Following back-calculation of the stronger dilution, the concentration for S-Tag-A was evaluated as $2400(90)\,\mu\mathrm{g\,mL^{-1}}$, and for S-Tag-B as $1200(100)\,\mu\mathrm{g\,mL^{-1}}$.

9.2.4 Electron Microscopy

10 Discussion

10.1 DNA Cloning

DNA cloning was concluded by sequencing of the plasmids and the calculation of the yield from the midi preparation. The full agreement of the sequencing result with the designed sequences enables further use of the plasmids in cell-free expression. The yields of the midi preparation are slightly lower than the manufacturer's reported typical yield of $500\,\mu\mathrm{g}$, but still high enough to allow for use in cell-free expression after concentration.

10.2 Protein Expression

10.2.1 eYFP Yield by Fluorescence

Following cell-free expression, the protein yield of the batch was estimated by fluorescence measurements of the Strep-eYFP reaction setup. The determined concentration of $0.3\,\mathrm{mg\,mL^{-1}}$ is significantly lower than the manufacturer's reported typical yield of at least $2\,\mathrm{mg\,L^{-1}}$. Typical reasons for low lysate performance, as stated by the manufacturer, are an inappropriate amount of plasmid DNA, poor oxygenation, or expired lysate. As for the plasmid DNA, the manufacturer explains that the optimal amount varies for each protein and should be determined by testing. Potentially, use of less DNA could improve the yield. Oxygenation in this setup was provided through a hole in the reaction tube's lid. An optimal oxygen supply could be established by using the dedicated perforated lids shipped with newer versions of the lysate. Use of newer lysate might also lead to a higher yield, since deterioration of the kit's components is a known issue

10.2.2 Coomassie Staining and Western Blot

Evaluation of the cell-free expression's qualitative success, as well as of the CaptoCore chromatography, was conducted through Western Blots using an anti-PVX antibody and an anti-S-Tag antibody.

After the Coomassie staining of the ALiCE lysate, the samples d29-A, S-Tag-A, and S-Tag-B, showed no notable difference from the non-template control, while the tracks containing lysate from the PVX-CP and the Strep-eYFP setup display bands at molecular weights matching the control. This could imply a lower yield in these setups. However, the coloring from other proteins in the lysate provides bad contrast, and visual differences might simply be caused by the fact that the bands are at slightly different heights than those of Strep-eYFP, and possibly worse to distinguish from the background. Further, the control S-Tag-CP-PVX only displays a faint band as well, reinforcing that there is no sure implication on the yield.

The anti-PVX Western Blot shows defined bands for d29-A, S-Tag-A, and PVX-CP, but only a very faint band for S-Tag-B. This could be due to a lower yield for S-Tag-B, or less avidity of the anti-PVX antibody against the protein, given that in only has 0.53 % sequence identity to the wild type. For all samples and the control S-Tag-CP-PVX, the migration of distance suggested a molecular weight higher than the theoretical one. For PVX particles from plants, this is a well-known phenomenon and partly due to specific glycosylations of the proteins [22]. However, glycosylation of the ALiCE samples is unlikely, since they were expressed through the cytosolic pathway. The discrepancy with the theoretical weights might imply partial aggregation.

In the Western Blot using the anti-S-Tag antibody, both samples S-Tag-A and S-Tag-B show a strong signal, implying that the antibody has high avidity against them. The signal is not constrained to the migration distance as seen in the anti-PVX blot, but is faintly visible through most of the track. Signal at larger molecular weights might stem from aggregates, while signal at lower molecular weights suggests the existence of smaller protein parts containing an S-Tag, possibly through partial degradation of the protein while boiling the samples.

Following CaptoCore purification, no discernible bands were visible in the Coomassie

gel. However, the Western Blot showed bands for S-Tag-A and PVX-CP using the anti-PVX antibody, as well as a faint band for S-Tag-A and a very faint band for S-Tag-B when using the anti-S-antibody. This implies the presence of particles or protein aggregates with a molecular weight larger than 700 kDa in the samples, even though in low amounts. Notably, the blot also shows a band for PVX-CP, even though the wild type coat protein is known to not assemble into VLPs in absence of a suitable RNA [11]. The faint bands could thus stem either from assembled particles, or from protein aggregates with a high molecular weight. This was further analyzed by electron microscopy.

10.2.3 ELISA

For a quantitative analysis of the cell-free expression of d29-A, S-Tag-A, S-Tag-B, and PVX-CP, ELISAs using both anti-PVX and anti-S-Tag antibodies were conducted.

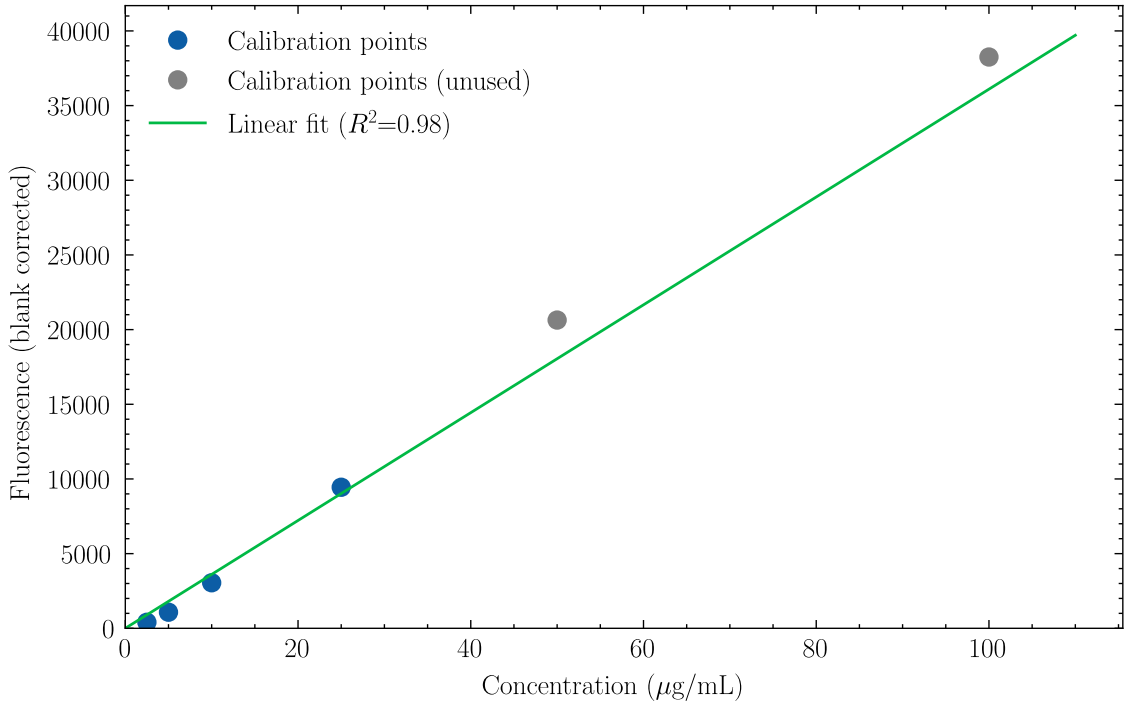
In both cases, the optical densities of the samples were almost identical for the 1:500 and the 1:1000 dilution. This might be caused by other proteins in the lysate saturating the binding capacity of the microwell plate, even at the stronger dilution, and thus rendering only the proportion of the specific protein in the lysate relevant. Under this assumption, the result from the 1:1000 dilution would be more significant, even though in general stronger dilutions should be tested to get a better estimate.

Using the 1:1000 dilution, the highest protein concentration in the anti-PVX ELISA was observed for the PVX-CP, with a calculated concentration of $134\,\mu g\,mL^{-1}$. This is less than half the value measured for Strep-eYFP expression by the lysate. A possible cause is lower expression or quicker degradation of the protein in the lysate. A different reason could be that the antibody has higher avidity for the assembled viral particles than for the PVX-CP in the ALiCE lysate, which are known to not assemble without specific RNA present [11]. This is plausible, since the antibody was generated by rabbit immunization against viral particles. For the constructs d29-A, S-Tag-A, and S-Tag-B, the anti-PVX ELISA leads to lower calculated concentrations of about $30\,\mu g\,mL^{-1}$ for d29-A and S-Tag-A, and no discernible signal for S-Tag-B. This could also be due to lower avidity of the antibody towards the adapted sequences. The missing signal for S-Tag-B is consistent with the anti-PVX Western Blot, which showed only a very faint line for S-Tag-B.

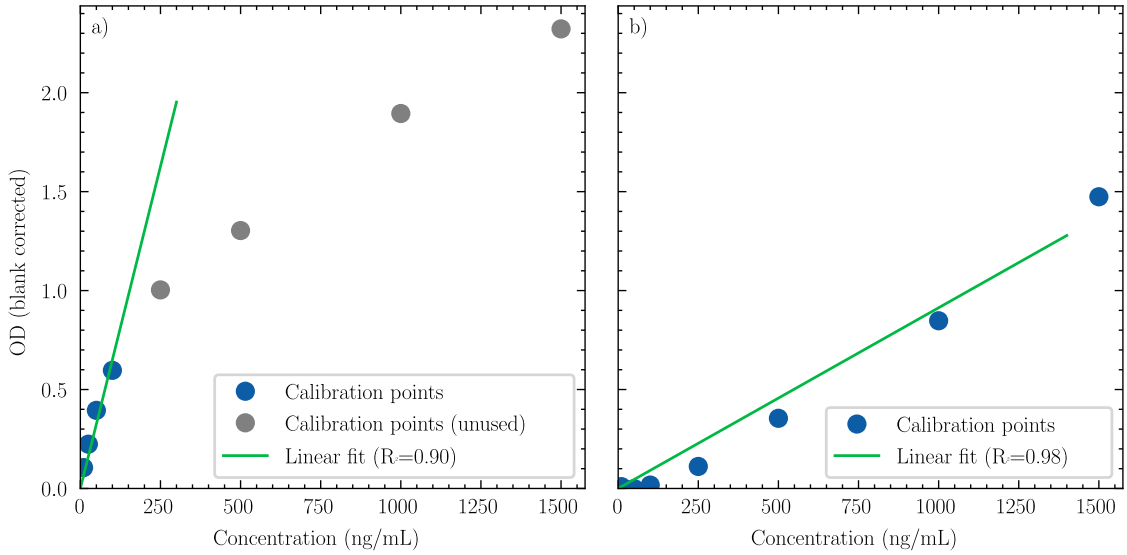
Using the anti-S-Tag antibody, only the samples from S-Tag-A and S-Tag-B had significant optical densities. The back-calculated concentrations of $2.4\,mg\,mL^{-1}$ and $1.2\,mg\,mL^{-1}$ are very high, and unlikely given the measured expression of only $0.3\,mg\,mL^{-1}$ for the Strep-eYFP ALiCE control setup. Possibly, the anti-S-Tag antibody has considerably higher avidity against protein monomers than for the assembled particles in the calibration samples. This could explain the high optical density, since the faint signal in the CaptoCore after the Western Blot implies that little to no protein in the ALiCE setup assembled. Notably, the anti-S-Tag antibody is not explicitly labeled as being suitable for ELISA, so the high signal might also be due to this, even though the calibration showed a linear trend with high determination.

10.2.4 Electron Microscopy

Appendix A: Supplementary Figures



Supplementary Figure S1: Calibration curve for eYFP measurement.



Supplementary Figure S2: Calibration curve for ELISA using anti-PVX and anti-S-Tag antibodies

Appendix B: Supplementary Tables

Supplementary Table S1: Calibration data for eYFP standard

Concentration ($\mu g\,mL^{-1}$)	Fluorescence
100.0	38 253 ± 1972
50.0	20 639 ± 618
25.0	9445 ± 388
10.0	3049 ± 181
5.0	1078 ± 56
2.5	405 ± 18

Supplementary Table S2: Measured fluorescence and estimated eYFP concentrations for Strep-eYFP (ALiCE) and Non-template (ALiCE)

Dilution	**Fluorescence**	**Conc.** ($\mu g\,mL^{-1}$)	**Back-calculated** ($\mu g\,mL^{-1}$)
Strep-eYFP (ALiCE)			
1:50	2149 ± 87	5.95 ± 0.24	298 ± 12
1:100	1107 ± 47	3.07 ± 0.13	307 ± 13
1:200	505 ± 42	1.40 ± 0.12	280 ± 23
1:500	184 ± 3	0.51 ± 0.01	255 ± 5
Non-template (ALiCE)			
1:50	1 ± 1	0.00	0.0 ± 0.1
1:100	-1	0.00	0.0
1:200	0 ± 1	0.00	0.0 ± 0.3
1:500	-1 ± 1	0.00	-1.0 ± 0.8

Supplementary Table S3: Calibration data for anti-PVX and anti-S-Tag ELISAs

Concentration ($ng\,mL^{-1}$)	OD (anti-PVX)	OD (anti-S-Tag)
1500	2.322 ± 0.037	1.474 ± 0.054
1000	1.895 ± 0.067	0.847 ± 0.032
500	1.303 ± 0.026	0.355 ± 0.021
250	1.003 ± 0.037	0.111 ± 0.019
100	0.597 ± 0.031	0.017 ± 0.018
50	0.395 ± 0.026	-0.001 ± 0.017
25	0.224 ± 0.032	-0.005 ± 0.017
10	0.105 ± 0.026	0.009 ± 0.018

Supplementary Table S4: Estimated concentrations from anti-PVX ELISA

Sample (Dilution)	**OD**	**Back-calculated** ($\mu g\,mL^{-1}$)
d29-A		
1:500	0.208 ± 0.028	16.0 ± 2.2
1:1000	0.200 ± 0.032	30.7 ± 4.9
S-Tag-A		
1:500	0.234 ± 0.033	18.0 ± 2.5
1:1000	0.217 ± 0.026	33.4 ± 4.0
S-Tag-B		
1:500	0.005 ± 0.026	0.4 ± 2.0
1:1000	0.020 ± 0.030	3.1 ± 4.6
CP-PVX		
1:500	0.871 ± 0.028	66.9 ± 2.2
1:1000	0.873 ± 0.026	134.2 ± 4.0
Non-template		
1:500	0.036 ± 0.032	2.8 ± 2.5
1:1000	0.012 ± 0.040	1.9 ± 6.2

Supplementary Table S5: Estimated concentrations from anti-S-Tag ELISA

Sample (Dilution)	**OD**	**Back-calculated** ($\mu g\,mL^{-1}$)
d29-A		
1:500	0.041 ± 0.023	22 ± 12
1:1000	0.041 ± 0.020	44 ± 22
S-Tag-A		
1:500	2.406 ± 0.134	1317 ± 73
1:1000	2.191 ± 0.079	2399 ± 86
S-Tag-B		
1:500	1.484 ± 0.073	813 ± 40
1:1000	1.094 ± 0.098	1198 ± 107
CP-PVX		
1:500	0.022 ± 0.018	12 ± 10
1:1000	0.046 ± 0.020	50 ± 22
Non-template		
1:500	0.039 ± 0.018	21 ± 10
1:1000	0.036 ± 0.019	38 ± 21

Appendix B: Algorithms

Supplementary Algorithm S1 Sample Diffusion

def SampleDiffusion($\{\mathbf{f}^*\}$, $\{\mathbf{s}_i^{\text{inputs}}\}$, $\{\mathbf{s}_i^{\text{trunk}}\}$, $\{\mathbf{z}_{ij}^{\text{trunk}}\}$, Noise Schedule $[c_0, c_1, \ldots, c_T]$, $\gamma_0 = 0.8$, $\gamma_{\min} = 1.0$, noise scale $\lambda = 1.003$, step scale $\eta = 1.5$):

$\vec{\mathbf{x}}_l \sim c_0 \cdot \mathcal{N}(\vec{\mathbf{0}}, \mathbf{I}_3)$ $\quad \vec{\mathbf{x}}_l \in \mathbb{R}^3$

for all $c_\tau \in [c_1, \ldots, c_T]$ **do**

  $\{\vec{\mathbf{x}}_l\} \leftarrow$ CentreRandomAugmentation$(\{\vec{\mathbf{x}}_l\})$

  $\gamma \leftarrow \gamma_0$ if $c_\tau > \gamma_{\min}$ else 0

  $\hat{t} \leftarrow c_{\tau-1}(\gamma + 1)$

  $\vec{\boldsymbol{\xi}}_l \leftarrow \lambda\sqrt{\hat{t}^2 - c_{\tau-1}^2} \cdot \mathcal{N}(\vec{\mathbf{0}}, \mathbf{I}_3)$ $\quad \vec{\boldsymbol{\xi}}_l \in \mathbb{R}^3$

  $\vec{\mathbf{x}}_l^{\text{noisy}} \leftarrow \vec{\mathbf{x}}_l + \vec{\boldsymbol{\xi}}_l$

  $\{\vec{\mathbf{x}}_l^{\text{denoised}}\} \leftarrow$ DiffusionModule$(\{\vec{\mathbf{x}}_l^{\text{noisy}}\}, \hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{inputs}}\}, \{\mathbf{s}_i^{\text{trunk}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\})$

  $\vec{\boldsymbol{\delta}}_l \leftarrow (\vec{\mathbf{x}}_l - \vec{\mathbf{x}}_l^{\text{denoised}})/\hat{t}$

  $dt \leftarrow c_\tau - \hat{t}$

  $\vec{\mathbf{x}}_l \leftarrow \vec{\mathbf{x}}_l^{\text{noisy}} + \eta \cdot dt \cdot \vec{\boldsymbol{\delta}}_l$

end for

return $\{\vec{\mathbf{x}}_l\}$

Supplementary Algorithm S2 Sample Diffusion with Symmetrization

def SampleDiffusion($\{\mathbf{f}^*\}$, $\{\mathbf{s}_i^{\text{inputs}}\}$, $\{\mathbf{s}_i^{\text{trunk}}\}$, $\{\mathbf{z}_{ij}^{\text{trunk}}\}$, Noise Schedule $[c_0, c_1, \ldots, c_T]$, $\gamma_0 = 0.8$, $\gamma_{\min} = 1.0$, noise scale $\lambda = 1.003$, step scale $\eta = 1.5$, Monomer Transforms $\{T_j\}$):

$\vec{\mathbf{x}}_l \sim c_0 \cdot \mathcal{N}(\vec{\mathbf{0}}, \mathbf{I}_3)$ $\quad \vec{\mathbf{x}}_l \in \mathbb{R}^3$

\# Modification: Initial Symmetrization

★: $(R_{\text{ref}}, \vec{\mathbf{t}}_{\text{ref}}) \leftarrow (\mathbf{I}, \text{mean}(\{\vec{\mathbf{x}}_l^{(1)}\}))$ $\quad$ Denoted as $T_{\text{ref}} = (R_{\text{ref}}, \vec{\mathbf{t}}_{\text{ref}})$

★: $\vec{\mathbf{x}}_l^{(j)} \leftarrow T_{\text{ref}} \circ T_j \circ T_{\text{ref}}^{-1} \circ \vec{\mathbf{x}}_l^{(1)}$

for all $c_\tau \in [c_1, \ldots, c_T]$ **do**

\# Track Origin of Symmetrization Center

★: $\vec{\mathbf{t}}_{\text{ref}} \leftarrow \text{mean}(\{\vec{\mathbf{x}}_l\}_{l \in I_1})$

$\{\vec{\mathbf{x}}_l\}, T_{\text{aug}} \leftarrow \text{CentreRandomAugmentation}(\{\vec{\mathbf{x}}_l\})$

\# Track Movement by CentreRandomAugmentation

★: $T_{\text{ref}} \leftarrow T_{\text{aug}} \circ T_{\text{ref}}$

$\gamma \leftarrow \gamma_0$ if $c_\tau > \gamma_{\min}$ else 0

$\hat{t} \leftarrow c_{\tau-1}(\gamma + 1)$

$\vec{\boldsymbol{\xi}}_l \leftarrow \lambda\sqrt{\hat{t}^2 - c_{\tau-1}^2} \cdot \mathcal{N}(\vec{\mathbf{0}}, \mathbf{I}_3)$ $\quad \vec{\boldsymbol{\xi}}_l \in \mathbb{R}^3$

$\vec{\mathbf{x}}_l^{\text{noisy}} \leftarrow \vec{\mathbf{x}}_l + \vec{\boldsymbol{\xi}}_l$

$\{\vec{\mathbf{x}}_l^{\text{denoised}}\} \leftarrow \text{DiffusionModule}(\{\vec{\mathbf{x}}_l^{\text{noisy}}\}, \hat{t}, \{\mathbf{f}^*\}, \{\mathbf{s}_i^{\text{inputs}}\}, \{\mathbf{s}_i^{\text{trunk}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\})$

\# Recenter and Symmetrize Denoised Prediction

★: $\vec{\mathbf{x}}_l^{\text{denoised}} \mathrel{+}= \text{mean}(\{\vec{\mathbf{x}}_l^{(1),\text{noisy}}\}) - \text{mean}(\{\vec{\mathbf{x}}_l^{(1),\text{denoised}}\})$

★: $\vec{\mathbf{x}}_l^{(j),\text{denoised}} \leftarrow T_{\text{ref}} \circ T_j \circ T_{\text{ref}}^{-1} \circ \vec{\mathbf{x}}_l^{(1),\text{denoised}}$

$\vec{\boldsymbol{\delta}}_l \leftarrow (\vec{\mathbf{x}}_l - \vec{\mathbf{x}}_l^{\text{denoised}})/\hat{t}$

$dt \leftarrow c_\tau - \hat{t}$

$\vec{\mathbf{x}}_l \leftarrow \vec{\mathbf{x}}_l^{\text{noisy}} + \eta \cdot dt \cdot \vec{\boldsymbol{\delta}}_l$

end for

return $\{\vec{\mathbf{x}}_l\}$

Appendix C: Synthetic Sequences

C.1: Construct A - Protein Sequence (Designed)

ASGLFTIPDGDFFDTVRHIVASNAVATNEDLSKIEALWKDMKVPTDTLFQAAVDLCRHCA
DVGSSAQTEMIGTGPYSNGISFARLAAAIRQVCTLRQFCMKYAPVVWNWMLTNNSPPANW
QARGFKPEHKFAAFDFFDGVTNPAAIMPKEGLLRPPSEAEMIAAHTAAEVKSTKARAQSN
DFASLDAAVTRGRITGQTTAEAVVTIPPP

C.2: Construct B - Protein Sequence (Designed)

ATGLNTVPDGDYFKTVKHKVLSNRVATDAELAAIETKWLAAGVPAATLFQAALDLCFQAA
DIGCGEDTVFVGTGPYTNGVSFQDLAAAIIRQVTTLLKFCRRYAPCVWNYMLTHNLPPADW
LARGFYPDHRYAAFDFFDGVENPAAIQPKLGLLRPPTVAERIAYHTLKTITTTTAAAAGN
DFASLHTAVTRGRLTGQSAAERIIHIPAA

C.3: Construct A - DNA Sequence (d29-A)

GCTTCTGGCTTATTCACCATACCTGACGGGGATTTCTTCGATACTGTAAGGCACATTGTA
GCTAGTAATGCTGTGGCAACAAACGAGGATCTCAGCAAGATCGAGGCTTTGTGGAAAGAC
ATGAAAGTTCCTACTGACACTCTTTTCCAGGCTGCCGTCGATCTGTGCCGACATTGTGCA
GACGTTGGGAGCTCTGCTCAGACAGAAATGATCGGTACTGGACCATATTCAAACGGAATA
TCATTTGCAAGACTGGCCGCTGCCATCCGACAAGTATGCACTTTGCGACAATTTTGTATG
AAGTACGCACCTGTTGTCTGGAATTGGATGTTGACTAATAATTCTCCACCCGCAAACTGG
CAGGCCAGAGGCTTCAAGCCAGAACACAAGTTTGCTGCATTTGACTTCTTCGACGGAGTT
ACTAATCCTGCCGCAATCATGCCTAAAGAAGGATTGTTACGACCCCCATCCGAGGCCGAG
ATGATCGCAGCTCATACTGCAGCCGAGGTCAAGAGCACAAAAGCTCGAGCACAAAGCAAT
GACTTCGCTTCCCTTGACGCAGCCGTCACACGAGGGCGAATCACCGGCCAGACTACAGCA
GAGGCCGTTGTGACCATTCCCCCCCCT

C.4: Construct B - DNA Sequence (d29-B)

GCCACCGGGCTGAATACCGTCCCTGACGGCGACTATTTCAAAACAGTCAAGCACAAGGTT
TTATCCAATAGAGTAGCAACTGATGCTGAGTTGGCCGCTATTGAGACAAAATGGCTGGCT
GCAGGGGTACCAGCAGCAACACTCTTCCAAGCCGCTCTGGACCTTTGTTTTCAGGCTGCC
GACATCGGTTGTGGTGAGGACACAGTATTCGTCGGGACCGGACCTTACACCAATGGCGTT
AGCTTCCAAGACCTCGCTGCCATCATAAGGCAAGTGACTACCTTATTGAAGTTCTGCCGA
CGTTACGCCCCCTGTGTATGGAACTATATGTTGACCCACAATCTTCCACCCGCAGATTGG
CTGGCACGTGGCTTCTATCCTGACCACAGGTATGCTGCCTTCGATTTCTTCGACGGCGTA
GAGAATCCTGCTGCTATCCAACCAAAATTGGGTCTGCTTCGTCCCCCTACTGTTGCTGAG
AGGATCGCTTACCATACCCTGAAGACCATTACAACAACCACCGCAGCCGCCGCAGGTAAC
GATTTTGCTTCACTTCATACTGCTGTAACTCGTGGTAGACTCACAGGTCAGAGCGCCGCT
GAGAGAATCATACACATACCCGCAGCT

C.5: Construct A - DNA Sequence with S-Tag (S-Tag-A)

ATGAGTAAAGAAACAGCCGCCGCTAAATTCGAACGTCAGCATATGGATAGTCCTGCATCA
ACAACCCAACCCATAGGTAGCACCACTAGCACAACTACTAAGACTGCCGGTGCAACCCCT
GCTACCGCTTCTGGCTTATTCACCATACCTGACGGGGATTTCTTCGATACTGTAAGGCAC

ATTGTAGCTAGTAATGCTGTGGCAACAAACGAGGATCTCAGCAAGATCGAGGCTTTGTGG
AAAGACATGAAAGTTCCTACTGACACTCTTTTCCAGGCTGCCGTCGATCTGTGCCGACAT
TGTGCAGACGTTGGGAGCTCTGCTCAGACAGAAATGATCGGTACTGGACCATATTCAAAC
GGAATATCATTTGCAAGACTGGCCGCTGCCATCCGACAAGTATGCACTTTGCGACAATTT
TGTATGAAGTACGCACCTGTTGTCTGGAATTGGATGTTGACTAATAATTCTCCACCCGCA
AACTGGCAGGCCAGAGGCTTCAAGCCAGAACACAAGTTTGCTGCATTTGACTTCTTCGAC
GGAGTTACTAATCCTGCCGCAATCATGCCTAAAGAAGGATTGTTACGACCCCCATCCGAG
GCCGAGATGATCGCAGCTCATACTGCAGCCGAGGTCAAGAGCACAAAAGCTCGAGCACAA
AGCAATGACTTCGCTTCCCTTGACGCAGCCGTCACACGAGGGCGAATCACCGGCCAGACT
ACAGCAGAGGCCGTTGTGACCATTCCCCCCCCTTAA

C.6: Construct B - DNA Sequence with S-Tag (S-Tag-B)

ATGAGCAAGGAGACTGCTGCAGCCAAGTTTGAGCGTCAGCACATGGACAGCCCAGCTTCA
ACCACTCAACCCATCGGTTCTACTACATCCACAACTACCAAGACAGCAGGGGCTACTCCA
GCCACAGCCACCGGGCTGAATACCGTCCCTGACGGCGACTATTTCAAAACAGTCAAGCAC
AAGGTTTTATCCAATAGAGTAGCAACTGATGCTGAGTTGGCCGCTATTGAGACAAAATGG
CTGGCTGCAGGGGTACCAGCAGCAACACTCTTCCAAGCCGCTCTGGACCTTTGTTTTCAG
GCTGCCGACATCGGTTGTGGTGAGGACACAGTATTCGTCGGGACCGGACCTTACACCAAT
GGCGTTAGCTTCCAAGACCTCGCTGCCATCATAAGGCAAGTGACTACCTTATTGAAGTTC
TGCCGACGTTACGCCCCCTGTGTATGGAACTATATGTTGACCCACAATCTTCCACCCGCA
GATTGGCTGGCACGTGGCTTCTATCCTGACCACAGGTATGCTGCCTTCGATTTCTTCGAC
GGCGTAGAGAATCCTGCTGCTATCCAACCAAAATTGGGTCTGCTTCGTCCCCCTACTGTT
GCTGAGAGGATCGCTTACCATACCCTGAAGACCATTACAACAACCACCGCAGCCGCCGCA
GGTAACGATTTTGCTTCACTTCATACTGCTGTAACTCGTGGTAGACTCACAGGTCAGAGC
GCCGCTGAGAGAATCATACACATACCCGCAGCTTAA

References

[1] Mark James Abraham et al. “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”. In: *SoftwareX* 1-2 (2015), pp. 19–25. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2015.06.001. URL: https://www.sciencedirect.com/science/article/pii/S2352711015000059.

[2] Josh Abramson et al. “Accurate structure prediction of biomolecular interactions with AlphaFold 3”. In: *Nature* 630.8016 (2024), pp. 493–500. DOI: 10.1038/s41586-024-07487-w. URL: https://doi.org/10.1038/s41586-024-07487-w.

[3] Nathaniel R. Bennett et al. “Improving de novo protein binder design with deep learning”. In: *Nature Communications* 14.1 (2023), p. 2625. DOI: 10.1038/s41467-023-38328-5. URL: https://doi.org/10.1038/s41467-023-38328-5.

[4] CAMILLA BETTI et al. “Potato virus X movement in Nicotiana benthamiana: new details revealed by chimeric coat protein variants”. In: *Molecular Plant Pathology* 13.2 (2012), pp. 198–203. DOI: https://doi.org/10.1111/j.1364-3703.2011.00739.x. eprint: https://bsppjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1364-3703.2011.00739.x. URL: https://bsppjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1364-3703.2011.00739.x.

[5] Longxing Cao et al. “Design of protein-binding proteins from the target structure alone”. In: *Nature* 605.7910 (2022), pp. 551–560. DOI: 10.1038/s41586-022-04654-9. URL: https://doi.org/10.1038/s41586-022-04654-9.

[6] J. Dauparas et al. “Robust deep learning-based protein sequence design using ProteinMPNN”. In: *Science* 378.6615 (2022), pp. 49–56. DOI: 10.1126/science.add2187. eprint: https://www.science.org/doi/pdf/10.1126/science.add2187. URL: https://www.science.org/doi/abs/10.1126/science.add2187.

[7] Peter L. Freddolino et al. “Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus”. In: *Structure* 14.3 (2006), pp. 437–449. ISSN: 0969-2126. DOI: https://doi.org/10.1016/j.str.2005.11.014. URL: https://www.sciencedirect.com/science/article/pii/S0969212606000608.

[8] Alessandro Grinzato et al. “Atomic structure of potato virus X, the prototype of the Alphaflexiviridae family”. In: *Nature Chemical Biology* 16.5 (2020), pp. 564–569. ISSN: 1552-4469. DOI: 10.1038/s41589-020-0502-4. URL: https://doi.org/10.1038/s41589-020-0502-4.

[9] Junyao He et al. “Virus-like Particles as Nanocarriers for Intracellular Delivery of Biomolecules and Compounds”. In: *Viruses* 14.9 (2022). ISSN: 1999-4915. DOI: 10.3390/v14091905. URL: https://www.mdpi.com/1999-4915/14/9/1905.

[10] Jochen S. Hub, Bert L. de Groot, and David van der Spoel. "g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates". In: *Journal of Chemical Theory and Computation* 6.12 (Dec. 2010), pp. 3713–3720. DOI: 10.1021/ct100494z. URL: https://doi.org/10.1021/ct100494z.

[11] Sun-Jung Kwon et al. "cis-Acting sequences required for coat protein binding and in vitro assembly of Potato virus X". In: *Virology* 334.1 (2005), pp. 83–97. ISSN: 0042-6822. DOI: 10.1016/j.virol.2005.01.018. URL: https://www.sciencedirect.com/science/article/pii/S0042682205000395.

[12] Jim Lawrence, Javier Bernal, and Christoph Witzgall. "A Purely Algebraic Justification of the Kabsch-Umeyama Algorithm". In: *Journal of Research of the National Institute of Standards and Technology* 124 (Oct. 2019). ISSN: 2165-7254. DOI: 10.6028/jres.124.028. URL: http://dx.doi.org/10.6028/jres.124.028.

[13] Duc H. T. Le, Ulrich Commandeur, and Nicole F. Steinmetz. "Presentation and Delivery of Tumor Necrosis Factor-Related Apoptosis-Inducing Ligand via Elongated Plant Viral Nanoparticle Enhances Antitumor Efficacy". In: *ACS Nano* 13.2 (Feb. 2019), pp. 2501–2510. DOI: 10.1021/acsnano.8b09462. URL: https://doi.org/10.1021/acsnano.8b09462.

[14] Duc H. T. Le et al. "Potato virus X, a filamentous plant viral nanoparticle for doxorubicin delivery in cancer therapy". In: *Nanoscale* 9 (6 2017), pp. 2348–2357. DOI: 10.1039/C6NR09099K. URL: http://dx.doi.org/10.1039/C6NR09099K.

[15] Diane L. Lynch et al. "Understanding Virus Structure and Dynamics through Molecular Simulations". In: *Journal of Chemical Theory and Computation* 19.11 (2023). PMID: 37192279, pp. 3025–3036. DOI: 10.1021/acs.jctc.3c00116. eprint: https://doi.org/10.1021/acs.jctc.3c00116. URL: https://doi.org/10.1021/acs.jctc.3c00116.

[16] Elaine C. Meng et al. "UCSF ChimeraX: Tools for structure building and analysis". In: *Protein Science* 32.11 (2023), e4792. DOI: https://doi.org/10.1002/pro.4792. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4792. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4792.

[17] Mona O. Mohsen and Martin F. Bachmann. "Virus-like particle vaccinology, from bench to bedside". In: *Cellular & Molecular Immunology* 19.9 (2022), pp. 993–1011. DOI: 10.1038/s41423-022-00897-8. URL: https://doi.org/10.1038/s41423-022-00897-8.

[18] Saghi Nooraei et al. "Virus-like particles: preparation, immunogenicity and their roles as nanovaccines and drug nanocarriers". In: *Journal of Nanobiotechnology* 19.1 (2021), p. 59. DOI: 10.1186/s12951-021-00806-7. URL: https://doi.org/10.1186/s12951-021-00806-7.

[19] Juliane Röder, Christina Dickmeis, and Ulrich Commandeur. "Small, Smaller, Nano: New Applications for Potato Virus X in Nanotechnology". In: *Frontiers in Plant Science* Volume 10 - 2019 (2019). ISSN: 1664-462X. DOI: 10.3389/fpls.2019.00158. URL: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2019.00158.

[20] Soheil Sarabandi and Federico Thomas. “Solution methods to the nearest rotation matrix problem in : A comparative survey”. In: *Numerical Linear Algebra with Applications* 30.5 (2023), e2492. DOI: https://doi.org/10.1002/nla.2492. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/nla.2492. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/nla.2492.

[21] Nicole F. Steinmetz et al. “Potato Virus X as a Novel Platform for Potential Biomedical Applications”. In: *Nano Letters* 10.1 (Jan. 2010), pp. 305–312. DOI: 10.1021/nl9035753. URL: https://doi.org/10.1021/nl9035753.

[22] Alejandro C. Tozzini et al. “Potato Virus X Coat Protein: A Glycoprotein”. In: *Virology* 202.2 (1994), pp. 651–658. ISSN: 0042-6822. DOI: https://doi.org/10.1006/viro.1994.1386. URL: https://www.sciencedirect.com/science/article/pii/S0042682284713869.

[23] Jeanmarie Verchot. “Potato virus X: A global potato-infecting virus and type member of the Potexvirus genus”. In: *Molecular Plant Pathology* 23.3 (2022), pp. 315–320. DOI: https://doi.org/10.1111/mpp.13163. eprint: https://bsppjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/mpp.13163. URL: https://bsppjournals.onlinelibrary.wiley.com/doi/abs/10.1111/mpp.13163.

[24] Joseph L. Watson et al. “De novo design of protein structure and function with RFdiffusion”. In: *Nature* 620.7976 (2023), pp. 1089–1100. DOI: 10.1038/s41586-023-06415-8. URL: https://doi.org/10.1038/s41586-023-06415-8.