000 CELL-DIFF: UNIFIED DIFFUSION MODELING FOR 001 **PROTEIN SEQUENCES AND MICROSCOPY IMAGES** 002 003

Anonymous authors

Paper under double-blind review

ABSTRACT

Fluorescence microscopy is ubiquitously used in cell biology research to characterize the cellular role of a protein. To help elucidate the relationship between the amino acid sequence of a protein and its cellular function, we introduce CELL-Diff, a unified diffusion model facilitating bidirectional transformations between protein sequences and their corresponding microscopy images. Utilizing reference cell morphology images and a protein sequence, CELL-Diff efficiently generates corresponding protein images. Conversely, given a protein image, the model outputs protein sequences. CELL-Diff integrates continuous and diffusion models within a unified framework and is implemented using a transformer-based network. We train CELL-Diff on the Human Protein Atlas (HPA) dataset and finetune it on the OpenCell dataset. Experimental results demonstrate that CELL-Diff outperforms existing methods in generating high-fidelity protein images, making it a practical tool for investigating subcellular protein localization and interactions.

- 1 INTRODUCTION
- 025 026

004

010 011

012

013

014

015

016

017

018

019

021

023

Protein sequences inherently encode their functions, and predicting these functions solely from se-027 quence information has become a critical area of research. With the development of artificial intelligence, learning-based methods are increasingly employed to predict a wide range of protein 029 properties, including structural conformation (Jumper et al., 2021; Baek et al., 2021), interaction partners (Evans et al., 2021), subcellular localization (Almagro Armenteros et al., 2017; Khwaja 031 et al., 2024b), and binding affinity (Rube et al., 2022). Concurrently, the rapid development of generative models has enabled researchers to design functional proteins (Madani et al., 2023; Dauparas 033 et al., 2022) and drug-like molecules (Isigkeit et al., 2024). These computational methods allow for 034 large-scale virtual screening, significantly reducing the costs and resources associated with experimental validation. The advent of those technologies presents significant opportunities for biomedical 035 research, potentially accelerating advancements in therapeutic target identification, drug discovery, 036 and the investigation of biochemical pathways (Palma et al., 2012). 037

In this work, we focus on the relationship between protein sequences and their cellular functions as characterized by microscopy images. Specifically, we focus on fluorescence microscopy which is ubiquitously used in nearly all cell biology research. Fluorescence microscopy images provide 040 extremely rich information for proteins of interest in the cellular context, such as their expression 041 level, subcellular distribution, and molecular interactions as can be measured by spatial colocaliza-042 tion. Such information characterizes protein functional activities as well as the physiological and 043 pathological state of cells. Disease-causing genetic mutations can alter the amino acid sequence 044 of proteins, resulting in changes in image phenotypes by modifying gene expression patterns, re-045 shaping molecular interaction profiles, or globally perturbing cellular states. As a first step towards 046 building a model that connects the sequence of proteins and their cellular images, recently, Khwaja 047 et al. (2024b) proposed CELL-E, a text-to-image transformer that predicts fluorescence protein im-048 ages from sequence input and cell morphology condition. Furthermore, CELL-E2 (Khwaja et al., 2024a) was developed to enhance the image generation speed of CELL-E by utilizing the idea from MaskGIT (Chang et al., 2022). Additionally, CELL-E2 facilitates the bidirectional transformation 051 between sequences and images. However, their image model only allowed output of highly blurred images lacking fine details to discern any of the subcellular structures other than the most promi-052 nent one (i.e. the nucleus), restricting their applicability only to the study of a very limited set of sequences features (i.e. the nuclear localization signal).

054 To expand the application of sequence-to-cell-image 056 generative models, we in-057 troduce CELL-Diff, a uni-058 fied diffusion model that enables bidirectional transformation between protein 060 sequences and their corre-061 sponding microscopy im-062 ages. Specifically, by uti-063 lizing cell morphology im-064 ages including the nucleus 065 and cytoplasmic markers



Figure 1: Given cell images as conditional input, CELL-Diff facilitates bidirectional generation between protein sequences and images.

066 such as endoplasmic reticulum (ER) and microtubule as conditional input, CELL-Diff can gener-067 ate detailed protein images from given protein sequences. Conversely, it can also output protein 068 sequences when provided with microscopy images, as shown in Figure 1. To enable this bidirectional transformation, CELL-Diff employs the continuous diffusion model for generating mi-069 croscopy images and the discrete diffusion model for redesigning protein sequences, which can be further integrated within a unified framework. Inspired by Unidiffuser(Bao et al., 2023), we adopt 071 separate diffusion time steps for the continuous and discrete diffusion models, enabling conditional 072 generation. The final objective function comprises the noise prediction loss for the continuous dif-073 fusion model and the masked value prediction loss for the discrete diffusion model. Moreover, we 074 design an attention-based U-Net model (Ronneberger et al., 2015; Peebles & Xie, 2023) to inte-075 grate information from both modalities efficiently. We evaluate CELL-Diff on HPA dataset (Digre 076 & Lindskog, 2021), which provides cellular microscopy images of human proteins based on fixed 077 immunofluorescence staining. Subsequently, we fine-tune the model on the OpenCell dataset (Cho 078 et al., 2022), which offers live microscopy images of different human cell lines, each tagged with a 079 single protein via CRISPR/Cas9 gene editing.

- We present CELL-Diff, a diffusion-based generative model that enables conditional bidirectional generation of protein sequences and their corresponding microscopy images. By integrating the continuous diffusion and discrete diffusion models, CELL-Diff can be trained within a unified framework. We propose an attention-based U-Net model for implementing CELL-Diff, which effectively integrates information from images and sequences.
 - We train CELL-Diff on the HPA dataset using different conditional cell images and finetune it on the OpenCell dataset. Experimental results show that our model generates more detailed and sharper protein images compared to previous methods.
- 2
- 090 091

081

082

084

087

RELATED WORKS

092 Multi-modal generative modeling can be formalized as learning the conditional or joint distribution between modalities. Representative applications include text-to-image generation (Ramesh et al., 2021; Ding et al., 2021; Nichol et al., 2022), image-to-text generation (image captioning) (Mokady 094 et al., 2021; Chen et al., 2023), text-to-video generation (Ho et al., 2022), and text-to-speech (Chen 095 et al., 2021; Popov et al., 2021). Most of these approaches rely on diffusion models or auto-096 regressive models for the generation and typically focus on unidirectional transformation. However, our goal is to achieve bidirectional generation, which requires the learning of joint distributions. 098 To achieve this, Hu et al. (2023) proposed a discrete diffusion-based model for learning the joint distribution between images and text, though its scalability remains unexplored. Bao et al. (2023) 100 introduced Unidiffuser, a unified diffusion model capable of unconditional, conditional, and joint 101 generation. The key observation of Unidiffuser is that the learning objective of the diffusion score 102 function can be unified in a general framework with multiple diffusion time steps. Furthermore, 103 Zhou et al. (2024) developed Transfusion, which integrates auto-regressive and diffusion models 104 for both single and cross-modality generation. Transfusion combines the auto-regressive loss with 105 diffusion, training a single transformer model using an extended causal mask. These methods generally depend on large pre-trained encoders for images and text, such as CLIP (Radford et al., 2021) 106 and VQGAN (Esser et al., 2021). However, for microscopy images, the variability in equipment 107 and experimental conditions limits the availability of such robust image encoders, making the direct application of these models challenging. Indeed, the two previous protein-sequence-to-microscopy generators, CELL-E (Khwaja et al., 2024b) and CELL-E2 (Khwaja et al., 2024a), which both used VQGAN, only produce coarse-grain images that have too much blur to distinguish fine-scale sub-cellular structures such as cytoskeleton. As for CELL-Diff, we combine continuous and discrete diffusion to enable bidirectional transformation between protein images and sequences. The model is trained on the pixel space, offering a straightforward and efficient approach.

115 3 TECHNICAL BACKGROUND

Before delving into our unified diffusion model, we briefly introduce the background of diffusion models applicable to continuous and discrete state spaces. Specifically, we employ the continuous diffusion model for microscopy images and the discrete diffusion model for protein sequences.

120 121 122

131 132

140

151

114

117

118

119

3.1 DIFFUSION MODEL FOR CONTINUOUS STATE SPACES

Let I_0 be a continuous random variable in \mathbb{R}^d , where d denotes the dimension, and let $I_{1:T} = \{I_t\}_{t=1}^T$ be a sequence of latent variables, with t as the index for diffusion steps. The diffusion model involves two processes: the forward process and the reverse process. In the forward process, the diffusion model progressively injects noise into the initial data I_0 , transforming it into a Gaussian random variable I_T . In the reverse process, the model learns to invert the diffusion process through a denoising model and generate new data by gradually eliminating the noise.

Forward process. The forward process involves injecting noise into the initial data. Given a variance schedule $\{\beta_t\}_{t=1}^T$, the forward process is defined as:

$$q(\mathbf{I}_t|\mathbf{I}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\mathbf{I}_{t-1},\beta_t\mathbf{I}_d), \quad t = 1,\dots,T.$$
(1)

Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. Then, for any arbitrary step t, it holds that $q(\mathbf{I}_t | \mathbf{I}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{I}_0, (1 - \bar{\alpha}_t)\mathbf{I}_d)$. Consequently, for a sufficiently large T, this process will transform \mathbf{I}_0 into an isotropic Gaussian variable.

Reverse process. The goal of the reverse process is to generate new samples from $p(\mathbf{I}_0)$ starting from a Gaussian random variable $\mathbf{I}_T \sim \mathcal{N}(0, \mathbf{I}_d)$. The reverse process is defined by a Markov Chain with trainable transitions:

$$p_{\theta}(\mathbf{I}_{t-1}|\mathbf{I}_t) = \mathcal{N}(\mu_{\theta}(\mathbf{I}_t, t), \sigma_t^2 \mathbf{I}_d), \quad t = 1, \dots, T.$$
(2)

Here, μ_{θ} represents parameterized neural networks designed to estimate the means from the current state, and σ_t^2 denotes the variance.

Training objective. The training objective function can be derived using variational inference. Instead of optimizing the intractable log-likelihood function $\log p(\mathbf{I}_0)$, the diffusion model maximize its ELBO:

$$\mathbb{E}_{q(\mathbf{I}_{1:T}|\mathbf{I}_0)}\left[\log p_{\theta}(\mathbf{I}_0|\mathbf{I}_1) - D_{\mathrm{KL}}(q(\mathbf{I}_T|\mathbf{I}_0)\|p_{\theta}(\mathbf{I}_T)) - \sum_{t=2}^T D_{\mathrm{KL}}\left(q(\mathbf{I}_{t-1}|\mathbf{I}_t,\mathbf{I}_0)\|p_{\theta}(\mathbf{I}_{t-1}|\mathbf{I}_t)\right)\right],\tag{3}$$

where
$$q(\mathbf{I}_{t-1}|\mathbf{I}_t, \mathbf{I}_0)$$
 has an formulation as $\mathcal{N}(\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{I}^0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{I}^t, \frac{(1-\bar{\alpha}_{t-1})\beta_t}{1-\bar{\alpha}_t}\mathbf{I}_d)$.

To simplify the computation, Ho et al. (2020) used a training objective based on a variant of the ELBO in Equation 3 as

 $L_{\text{DDPM}} = \mathbb{E}_{t,\mathbf{I}_{0},\epsilon} \| \epsilon_{\theta} (\sqrt{\bar{\alpha}_{t}} \mathbf{I}_{0} + \sqrt{1 - \bar{\alpha}_{t}} \epsilon, t) - \epsilon \|_{2}^{2},$ (4)

155 156 157

154

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and ϵ_{θ} is a noise prediction network.

158 3.2 DIFFUSION MODEL FOR DISCRETE STATE SPACES

Several distinct diffusion models are designed for discrete data (Austin et al., 2021; Hoogeboom et al., 2021). This section focuses on the order-agnostic Autoregressive Diffusion Models (OA-ARDM) (Hoogeboom et al., 2022).

Figure 2: Training losses of CELL-Diff. During each training iteration, the protein image I^{port} and sequence \mathbf{S} are transformed using the forward processes of the continuous and discrete diffusion models, with randomly sampled time steps t_{I} and t_{S} , respectively. The network model is tasked with predicting the noise in the protein image and the masked values from the protein sequence, corresponding to the noise prediction loss L_{I} and the masked value prediction loss L_{S} .

Let $\mathbf{S} = (\mathbf{S}_1, \dots, \mathbf{S}_D)$ be a multivariate random variable, where $\forall t \in \{1, \dots, D\}, \mathbf{S}_t \in \{1, \dots, K\}$ with K categories. Denote S_D as the set of all permutations of the integers $1, \ldots, D$, and assume σ represents a random ordering in S_D . Applying Jensen's inequality, we obtain:

$$\log p(\mathbf{S}) = \log \mathbb{E}_{\sigma \sim \mathcal{U}(S_D)} p(\mathbf{S}|\sigma) \ge \mathbb{E}_{\sigma \sim \mathcal{U}(S_D)} \log p(\mathbf{S}|\sigma), \tag{5}$$

where $\mathcal{U}(S_D)$ denotes the uniform distribution over S_D . Following order σ , $\log p(\mathbf{S}|\sigma)$ can be factorized as $\sum_{t=1}^{D} \log p(\mathbf{S}_{\sigma(t)} | \mathbf{S}_{\sigma(< t)})$, where $\mathbf{S}_{\sigma(< t)} = (\mathbf{S}_{\sigma(1)}, \dots, \mathbf{S}_{\sigma(t-1)})$. Combining this with Equation 5, we have:

$$\log p(\mathbf{S}) \ge \mathbb{E}_{\sigma \sim \mathcal{U}(S_D)} \sum_{t=1}^{D} \log p(\mathbf{S}_{\sigma(t)} | \mathbf{S}_{\sigma($$

Therefore, denote f_{θ} as the neural network, C as the categorical distribution, the loss function for OA-ARDM is

$$L_{\text{OA-ARDM}} = \mathbb{E}_{\sigma \sim \mathcal{U}(S_D), t \sim \mathcal{U}(1, \dots, D)} \frac{1}{D - t + 1} \sum_{k \in \sigma(>t)} -\log \mathcal{C}(\mathbf{S}_k | \mathbf{f}_{\theta}(\mathbf{S}_{\sigma((7)$$

This objective function corresponds to the "Masked Language Modeling" training objective pro-posed in BERT (Kenton & Toutanova, 2019) with a reweighting term. At each training step, we first sample a time step t from $\mathcal{U}(1,\ldots,D)$, followed by a random ordering σ from $\mathcal{U}(S_D)$. We then input $\mathbf{S}_{\sigma(< t)}$ into the model, which predicts the remaining values $\mathbf{S}_{\sigma(> t)}$. In the generation step, we first sample a random ordering and then generate the values according to that order. These processes are facilitated through a masking operation, see Appendix A for the details.

METHODOLOGY

In this section, we introduce our unified diffusion model for generating microscopy images and protein sequences. Let I^{prot} represents the protein image, I^{cell} represents the cell morphology image, and S represents the protein sequence. The task of protein image prediction involves sampling from the conditional distribution $p(\mathbf{I}^{\text{prot}}|\mathbf{S}, \mathbf{I}^{\text{cell}})$, while the task of sequence generation involves sampling from $p(\mathbf{S}|\mathbf{I}^{\text{prot}}, \mathbf{I}^{\text{cell}})$. To achieve these goals within a unified diffusion model, we choose to estimate the joint distribution $p(\mathbf{I}^{\text{prot}}, \mathbf{S} | \mathbf{I}^{\text{cell}}))$, which involves model a continues variable \mathbf{I}^{prot} and a discrete variable S.

4.1 **PROPOSED METHOD**

Let I_0 represent the protein image I^{port} and temporarily ignore the cell image I^{cell} , we consider modeling the joint distribution $p(\mathbf{I}_0, \mathbf{S})$. Following the diffusion models described in in Section 3,



we introduce a sequence of latent variables $\mathbf{I}_{1:T} = {\{\mathbf{I}_t\}_{t=1}^T \text{ for } \mathbf{I}_0 \text{ and a random ordering } \sigma \in S_D$ for S. The log-likelihood function satisfies:

$$\log p(\mathbf{I}_0, \mathbf{S}) = \log \mathbb{E}_{\sigma \sim \mathcal{U}(S_D)} \mathbb{E}_{q(\mathbf{I}_{1:T} | \mathbf{I}_0)} \frac{p(\mathbf{I}_{0:T}, \mathbf{S} | \sigma)}{q(\mathbf{I}_{1:T} | \mathbf{I}_0)} \ge \mathbb{E}_{\sigma \sim \mathcal{U}(S_D)} \mathbb{E}_{q(\mathbf{I}_{1:T} | \mathbf{I}_0)} \log \frac{p(\mathbf{I}_{0:T}, \mathbf{S} | \sigma)}{q(\mathbf{I}_{1:T} | \mathbf{I}_0)}.$$
(8)

Assuming the same forward and reverse process in Section 3.1, $q(\mathbf{I}_{1:T}|\mathbf{I}_0)$ can be decomposed as $q(\mathbf{I}_T|\mathbf{I}_0) \prod_{t=2}^{T} q(\mathbf{I}_{t-1}|\mathbf{I}_t, \mathbf{I}_0).$

224 Regrading $\log p(\mathbf{I}_{0:T}, \mathbf{S} | \sigma)$, the decomposition depends on the factorization order between $\mathbf{I}_{0:T}$ and 225 S. Given a specific factorization order, for sequence S, each decomposition term can be represented as $\log p(\mathbf{S}_{\sigma(t_{\mathbf{S}})}|\mathbf{S}_{\sigma(< t_{\mathbf{S}})}, \mathbf{I}_{\geq t_{\mathbf{I}}})$, where $t_{\mathbf{S}} \in \{1, \ldots, D\}$ and $t_{\mathbf{I}} \in \{0, \ldots, T\}^1$. Further-226 more, since the forward process shown in Equation 1 indicates that the information from I_0 to I_T 227 is progressively decreasing, we assume $\log p(\mathbf{S}_{\sigma(t_{\mathbf{S}})}|\mathbf{S}_{\sigma(<t_{\mathbf{S}})}, \mathbf{I}_{\geq t_{\mathbf{I}}}) = \log p(\mathbf{S}_{\sigma(t_{\mathbf{S}})}|\mathbf{S}_{\sigma(<t_{\mathbf{S}})}, \mathbf{I}_{t_{\mathbf{I}}}).$ 228 For image I, using the Markov Chain model, each decomposition term can be represented as 229 $\log p(\mathbf{I}_T | \mathbf{S}_{\sigma(\langle t_{\mathbf{S}})})$ and $\log p(\mathbf{I}_{t_{\mathbf{I}}-1} | \mathbf{I}_{t_{\mathbf{I}}}, \mathbf{S}_{\sigma(\langle t_{\mathbf{S}})})$, where $t_{\mathbf{I}} \in \{1, \ldots, T\}$ and $t_{\mathbf{S}} \in \{1, \ldots, D + t_{\mathbf{I}}\}$ 230 Combining this with $q(\mathbf{I}_{1:T}|\mathbf{I}_0)$, the KL term in Equation 3 can be expressed as 1}. 231 $D_{\text{KL}}\left(q(\mathbf{I}_{t_{1}-1}|\mathbf{I}_{t_{1}},\mathbf{I}_{0})\|p(\mathbf{I}_{t_{1}-1}|\mathbf{I}_{t_{1}},\mathbf{S}_{\sigma(< t_{S})})\right)$, which has a closed-form formulation with Gaussian 232 parameterization. 233

In practice, the choice of factorization order between $I_{0:T}$ and S depends on the downstream purpose. In our case, we aim to generate samples from two conditional distributions $p(\mathbf{I}|\mathbf{S})$ and $p(\mathbf{S}|\mathbf{I})$, which requires simultaneously factorizing $p(\mathbf{I}_{0:T}, \mathbf{S}|\sigma)$ from $\mathbf{I}_{0:T}$ to S and from S to $\mathbf{I}_{0:T}$. To achieve this goal, we adopt the approach from UniDiffuser (Bao et al., 2023), considering all possible factorization combinations. Therefore, we maximize the following objective function:

$$\mathbb{E}_{\sigma \sim \mathcal{U}(S_D)} \mathbb{E}_{q(\mathbf{I}_{1:T} | \mathbf{I}_0)} \sum_{t_{\mathbf{S}}=1}^{D} \sum_{t_{\mathbf{I}}=0}^{T} \log p(\mathbf{S}_{\sigma(t_{\mathbf{S}})} | \mathbf{S}_{\sigma($$

241 242 243

239 240

219 220

221

243 244 245

246

247

251

253

254

255 256 257

258

259 260 261

262 263

264 265

266

267

268

269

 $+\sum_{t_{s}=1}^{D+1} \log p(\mathbf{I}_{0}|\mathbf{I}_{1}, \mathbf{S}_{\sigma(< t_{s})}) - \sum_{t_{s}=1}^{D+1} \sum_{t_{I}=1}^{T} D_{KL} \left(q(\mathbf{I}_{t_{I}-1}|\mathbf{I}_{t_{I}}, \mathbf{I}_{0}) \| p(\mathbf{I}_{t_{I}-1}|\mathbf{I}_{t_{I}}, \mathbf{S}_{\sigma(< t_{s})}) \right),$ (9) where the first term corresponds to the objective function of OA-ARDM in Equation 6, while the remaining terms correspond to the objective function of DDPM in Equation 3.

Utilizing the same parametrization technique as shown in Equation 4 and Equation 7, and considering modeling the joint distribution $p(\mathbf{I}^{\text{port}}, \mathbf{S} | \mathbf{I}^{\text{cell}})$, let \mathbf{f}_{θ} denotes the neural network. The training objective function for protein sequence \mathbf{S} is:

$$L_{\mathbf{S}} = \mathbb{E}_{\sigma \sim \mathcal{U}(S_D), \mathbf{I}^{\text{port}}, t_{\mathbf{I}}, t_{\mathbf{S}}, \epsilon} \frac{\sum_{k \in \sigma(\geq t_{\mathbf{S}})} -\log \mathcal{C}(\mathbf{S}_k | \mathbf{f}_{\theta}(\mathbf{S}_{\sigma(\langle t_{\mathbf{S}})}, \sqrt{\bar{\alpha}_{t_{\mathbf{I}}}} \mathbf{I}^{\text{port}} + \sqrt{1 - \bar{\alpha}_{t_{\mathbf{I}}}} \epsilon, t_{\mathbf{I}}, \mathbf{I}^{\text{cell}}))}{D - t_{\mathbf{S}} + 1}.$$
(10)

The training objective function for protein image **I**^{port} is:

$$L_{\mathbf{I}} = \mathbb{E}_{\sigma \sim \mathcal{U}(S_D), \mathbf{I}^{\text{port}}, t_{\mathbf{I}}, t_{\mathbf{S}}, \epsilon} \| \mathbf{f}_{\theta}(\mathbf{S}_{\sigma(\langle t_{\mathbf{S}})}, \sqrt{\bar{\alpha}_{t_{\mathbf{I}}}} \mathbf{I}^{\text{port}} + \sqrt{1 - \bar{\alpha}_{t_{\mathbf{I}}}} \epsilon, t_{\mathbf{I}}, \mathbf{I}^{\text{cell}}) - \epsilon \|_{2}^{2}.$$
(11)

In summary, combining Equation 10 and Equation 11 and introduce a balancing coefficient λ , the total loss of the proposed CELL-Diff model is:

$$L_{\text{CELL-Diff}} = L_{\mathbf{S}} + \lambda L_{\mathbf{I}}.$$
 (12)

The training strategy is shown in Figure 2.

4.2 MODEL DETAILS

Inference. After training, we can generate samples from two conditional distributions: $p(\mathbf{I}^{\text{port}}|\mathbf{S}, \mathbf{I}^{\text{cell}})$ and $p(\mathbf{S}|\mathbf{I}^{\text{port}}, \mathbf{I}^{\text{cell}})$. Specifically, to generate the protein image \mathbf{I}^{port} , we utilize the conventional reverse diffusion process as shown in Equation 2, conditioning on the unmasked protein

¹Given that the gap between \mathbf{I}_T and the standard Gaussian noise is negligible, we assume $\log p(\mathbf{S}_{\sigma(t_{\mathbf{S}})}|\mathbf{S}_{\sigma(< t_{\mathbf{S}})}) = \log p(\mathbf{S}_{\sigma(t_{\mathbf{S}})}|\mathbf{S}_{\sigma(< t_{\mathbf{S}})}, \mathbf{I}_T).$

274

270

281

283

284

285

287



286 Figure 3: Network architecture of CELL-Diff. Microscopy images are embedded into a latent sequence through residual and attention blocks. The protein sequences are embedded using a pre-288 trained ESM2 model (Lin et al., 2022). These embeddings are concatenated and processed by a 289 transformer model. The U-Net architecture (Ronneberger et al., 2015) is employed to output the noise in the protein image, while a linear projection is utilized to predict the masked values in the 290 protein sequence. Cross-attention mechanisms are implemented to enhance information integration from images and sequences. 292

293

295

296

297 298

299

300

291

sequence S and the cell image \mathbf{I}^{cell} . The network model employed for generation is $\mathbf{f}_{\theta}(\mathbf{S},\cdot,t_{\mathbf{I}},\mathbf{I}^{\text{cell}})$. where $t_{I} = 1, \ldots, T$. For the generation of the protein sequence S, we utilize the reverse process of discrete diffusion OA-ARDM (Hoogeboom et al., 2022). We first sample a random ordering σ , and then generate sequence from $p(\mathbf{S}_{\sigma(t_{\mathbf{S}})}|\mathbf{S}_{\sigma(<t_{\mathbf{S}})}, \mathbf{I}^{\text{port}}, \mathbf{I}^{\text{cell}})$, where $t_{\mathbf{S}} = 1, \ldots, D$. The network model in this scenario is $\mathbf{f}_{\theta}(\cdot, \mathbf{I}^{\text{port}}, 0, \mathbf{I}^{\text{cell}})$. The sampling algorithm for OA-ARDM is shown in Algorithm 2.

Network architecture. As shown in Equation 11 and Equation 10, the network model f_{θ} takes four 301 inputs: the protein sequence S, the protein image I^{port} , the cell image I^{cell} , and the diffusion time step 302 $t_{\rm I}$. To process the protein and cell images, we first concatenate them and then apply the commonly 303 used U-Net architecture (Ronneberger et al., 2015). The concatenated images are fed into a series of 304 downsampling blocks, transforming into image embeddings. The protein sequences are embedded 305 using a pre-trained ESM2 model Lin et al. (2022). Then, the image and protein embeddings are 306 concatenated and processed using an encoder-only transformer model. After passing through the 307 transformer module, the concatenated feature tensors are split into image and sequence feature ten-308 sors. The image feature tensors are then upsampled and combined with the downsampling features 309 to output the noise from the protein image. The sequence feature tensor is processed using a linear projector to predict the masked values. The upsampling and downsampling blocks in the U-Net 310 consist of residual and attention blocks. To enhance the integration of sequence information within 311 the image processing component, we utilize cross-attention mechanisms with the attention blocks. 312 Furthermore, we employ the adaptive layer norm zero (adaLN-Zero) conditioning method (Peebles 313 & Xie, 2023) for incorporating the diffusion time step $t_{\rm I}$. The network architecture is illustrated 314 in Figure 3. 315

316

318

320

317 5 EXPERIMENTS

319 5.1 DATASETS

321 Human Protein Atlas. The Human Protein Atlas (HPA) dataset (Digre & Lindskog, 2021) includes immunofluorescence images across various human cell lines with the proteins of interest stained by 322 antibodies. It provides cellular images for 12,833 proteins, as well as corresponding cell morphology 323 images consisting of staining for the nucleus, ER, and microtubules. For each protein, the dataset



Figure 4: Visual results of protein image generation on HPA dataset.

includes multiple microscopy images from different cell lines. The corresponding protein sequences
can be accessed from the UniProt dataset (UniProt Consortium, 2018). In total, we have collected
88,483 data points, each containing a protein sequence, a protein image, a nucleus image, an ER
image, and a microtubule image.

OpenCell. The OpenCell (Cho et al., 2022) dataset provides a library of 1,311 CRISPR-edited
 HEK293T human cell lines, each with a target protein fluorescently tagged using the split mNeonGreen2 system. For each target protein, OpenCell provides 4–5 confocal images along with
 a reference nucleus image. The cells were imaged live, offering a more accurate representation of
 protein distribution than the immunofluorescence images from HPA. Notably, 1,102 proteins are
 common between the HPA and OpenCell datasets. In total, we collected 6,301 data points, each
 containing a protein sequence, a protein image, and a nucleus image.

Given the size limitations of the HPA and OpenCell datasets, particularly in the diversity of protein
 sequences, we randomly selected 100 proteins from the shared subset between the two datasets as
 the test set, leaving the remainder for training. The test set for HPA and OpenCell contains 766 and
 470 data points, respectively.

371

373

353

354 355

372 5.2 IMPLEMENTATION DETAILS

We first train CELL-Diff models on the HPA dataset and then fine-tune on the OpenCell dataset. Both pre-training and fine-tuning are conducted for 100,000 iterations using the Adam optimizer Kingma & Ba (2014). The learning rate is initialized using a linear warm-up strategy, increasing from 0 to 3×10^{-4} over the first 1,000 iterations, followed by a linear decay to zero. The batch size is set to 192. For images from the HPA dataset, we apply the random crop of size 1024, Table 1: Comparison of protein image generation performance on HPA and OpenCell datasets. "Nucl" denotes the nucleus image, "ER" denotes the endoplasmic reticulum image, and "MT" denotes the microtubule image. "FID-T" indicates the FID computed using the thresholded protein image, and "FID-O" indicates the FID computed using the original protein image.

Dataset	Method	Cell image	\mid MSF-resolvability (nm) \downarrow	$\text{IoU}\uparrow$	$\text{FID-T}\downarrow$	$\text{FID-O} \downarrow$
	CELL-E2	Nucl	1872	0.461	77.0	167.0
		Nucl	646	0.448	35.9	31.9
HPA	CELL D:#	Nucl, ER	642	0.619	32.4	25.2
	CELL-DIII	Nucl, MT	642	0.601	33.0	27.1
		Nucl, ER, MT	641	0.623	34.1	24.1
OpenCell	CELL-E2	Nucl	1239	0.515	70.4	248.1
OpenCen	CELL-Diff	Nucl	628	0.524	40.4	20.0

followed by resizing to 256. For the OpenCell dataset, images are randomly cropped to a size of 256. Data augmentation is performed using random flips and rotations. The sequence embedding dimension is 640, and the transformer module consists of 24 layers with 8-head attention. The U-Net architecture includes three groups of downsampling and upsampling modules, each containing two residual and attention blocks, with channel sizes increasing from 64 to 512. To convert images into sequences, we use the patchify operation from DiT (Peebles & Xie, 2023) with a patch size of 8. CELL-Diff is trained with 1,000 diffusion steps using the shifted cosine noise schedules (Hoogeboom et al., 2023), and use DDIM (Song et al., 2020) with 100 steps to accurate the sampling speed. The weighting coefficient λ in Equation 12 is set to 100, and the maximum protein sequence length is 2,048. All models are trained using two Nvidia H100 GPUs.

5.3 PROTEIN IMAGE GENERATION

We evaluate the protein image generation performance of CELL-Diff. Given that the protein image prediction problem is relatively new, we compare CELL-Diff with the most closely related method, CELL-E2 (Khwaja et al., 2024a). To provide a quantitative comparison, we introduce the Maximum Spatial Frequency (MSF) resolvability for microscopy images to measure its capability to discern fine structural details. Given a microscopy image I, we define the Fourier Ring Power Spectral Density (FRPSD) as FRPSD(r) = $\sum_{r_i \in r} |\hat{\mathbf{I}}_1(r_i)|^2$, where $\hat{\mathbf{I}}$ denotes the Fourier transform of I and r_i denotes the pixel element at radius r. The MSF-resolvability is then defined as:

$$\text{MSF-resolvability} = \frac{1}{f}, \ f = \frac{i}{\text{Image Size} \times \text{Pixel Size}}, \ \text{where} \ \begin{cases} \text{FRPSD}(r) > 10^{-3}, & r < i \\ \text{FRPSD}(r) < 10^{-3}, & r = i \end{cases}.$$
(13)

416 We also employ the Intersection over Union (IoU) metric, which measures the similarity between 417 two masks and is commonly used in image segmentation tasks. To calculate IoU, we apply median 418 value thresholding to the original protein images to generate binary masks, while for CELL-E2, we use the predicted thresholding images. Additionally, we compute the Fréchet Inception Distance 419 (FID) Heusel et al. (2017) score to evaluate the similarity between the real and predicted images. 420 FID is a learning-based metric that evaluates the quality of images generated by generative models. 421 It measures the similarity between the generated and real images regarding their feature distribu-422 tions. Lower FID scores indicate that the generated images are more similar to the real images. To 423 compute FID, we concatenate the protein and nucleus images as input. In practice, we compute 424 FID-T and FID-O, representing the FID score based on thresholding and original protein images, 425 respectively. The results are shown in Table 1. The results show that CELL-Diff generated images 426 exhibit better MSF-resolvability than CELL-E2. In particular, the MSF-resolvability for the orig-427 inal HPA and OpenCell data are 640 nm and 426 nm, respectively. The results from CELL-Diff 428 are approaching the resolvability of the original training data, allowing us to discern finer details in protein distribution such as various cytoplasmic organelles. Regarding the prediction accuracy met-429 ric IoU, CELL-Diff and CELL-E2 achieve comparable performance when using only the nucleus 430 image as the conditional cell image, which can be greatly improved by incorporating additional cell 431 morphology images, such as those of the ER and microtubules. Regarding the learning-based metric

380

381

382

394

396

397

398

399

400

401

402 403

404

433	Table 2: Ablation analysis of	cross attention	module on	HPA d	ataset.	The nucleus	image is u	sed as
434	the cell morphology image.							
435	Method	MSF-resolvat	oility (nm)	J. Io	U↑	FID-T ↓	FID-O 🗍	

Me	thod MS	F-resolvability (nm)	\downarrow IoU \uparrow	FID-T \downarrow	FID-O \downarrow	
w/o cross	s attention	648	0.431	33.4	40.1	
w cross	attention	646	0.448	35.9	31.9	



Figure 5: Protein localization signal screening. Test sequences are tagged to the C-terminus of the protein BLVRA. The ratio "IoO" represents the median protein intensity inside the nucleus relative to that outside the nucleus.

FID, CELL-Diff significantly outperforms CELL-E2, further demonstrating the superiority of the proposed method. Visual results are illustrated in Figure 4. From the figure, we find that CELL-Diff accurately predicts protein images from unseen protein sequences. Compared with CELL-E2, CELL-Diff generates more resolvable images, enabling the extraction of more detailed information from the generated images. More results are provided in Appendix B.

6 DISCUSSIONS

6.1 ABLATION ON CROSS ATTENTION MODULE

We employ the cross-attention mechanism to more effectively integrate information from sequences to images. To evaluate its efficiency, we conduct an ablation analysis of this module on the HPA dataset, see Table 2. The results show that incorporating this module improves most of the quantitative metrics, demonstrating the effectiveness of the cross-attention mechanism.

470 471

6.2 POTENTIAL APPLICATIONS

In this section, we present three potential applications of the proposed CELL-Diff method for biological discovery. Given that validation relies on biological knowledge and the dataset size is limited,
we retrain all models using all the protein sequences from both the HPA and OpenCell datasets.

475 Virtual screening of protein localization signal. CELL-Diff can be applied for the virtual screen-476 ing of protein localization signals, such as Nuclear Localization Signals (NLS) and Nuclear Export 477 Signals (NES). The NLS is a short amino acid sequence that directs the import of proteins into the 478 nucleus, while the NES facilitates their export from the nucleus. In this approach, the test peptide se-479 quence is tagged to the C-terminus of the protein BLVRA, which is uniformly distributed both inside 480 and outside the nucleus, see the first column of Figure 5. CELL-Diff is then employed to predict 481 the images of the modified protein. The resulting predicted images are analyzed to identify potential 482 localization signals. As illustrated in Figure 5, we compute the median fluorescence intensity inside the nucleus relative to that outside the nucleus, referred to as the IoO ratio. For the original BLVRA 483 protein, the IoO ratio is 1.12. If the IoO ratio of the modified protein exceeds 1.12, the test sequence 484 is likely to function as an NLS, conversely, if the ratio is lower, the sequence is more likely to act as 485 an NES. In Figure 5, we tested known three NLSs and three NESs from the literature. CELL-Diff

436 437 438

451

452

453

454 455 456

457

458

459

460 461 462

463 464

Microtubules Protein: TUBB4A Protein: TUBA1A Protein: KIF18A Protein: EML4 Overlaid image Cell morphology TUBA1A Red: images Green: KIF18A Blue: EML4 Nucleus ER

Figure 6: Virtual staining using HPA data. From identical cell morphology images, CELL-Diff generates staining results for various proteins.

successfully recognized these signals, proving its capability as a computational tool for screening potential protein localization signals.

Virtual staining. Typical fluorescence microscopes can only fit no more than four color channels in 505 the visible spectrum. Because of this physical limitation, both HPA and OpenCell acquire the images 506 of only one protein of interest per sample, with the other color channels occupied by morphological 507 reference images. Consequently, it is challenging to identify the intracellular spatial relationships 508 among multiple proteins of interest because their images are from different cells. With CELL-Diff, 509 we solve this problem by generating images of these proteins conditioned on the same morphology 510 reference images. These virtual staining images allow the subcellular distributions of an arbitrary 511 number of proteins to be directly compared and potential molecular interactions identified from 512 colocalization, while entirely circumventing the color channel limitation of fluorescence microscopy 513 experiments. We demonstrate that from cell morphology images not in the training data set, CELL-Diff can accurately simulate the imaging results for target protein sequences, see Appendix C. We 514 further demonstrate the use of CELL-Diff to identify molecular interaction by virtually staining two 515 microtubule components (TUBA1A and TUBB4A) and two other proteins, KIF18A and EML4, 516 from the same morphology image. The overlaid image clearly shows the association of KIF18A and 517 EML4 with microtubules in the cell, consistent with their known biological function of microtubule 518 binding, see Figure 6. 519

520 Localization signal generation. Utilizing image-to-sequence generation, CELL-Diff can be applied to generate novel protein localization signals, such as NLS and NES. Given a cell morphology 521 image and a corresponding protein image, CELL-Diff generates the protein sequences that should be 522 located at the position indicated by the protein image. We started from the Green Fluorescent Protein 523 (GFP) which has no sequence homology with any human proteins and does not contain localization 524 signals by itself (Köhler et al., 1997; Seibel et al., 2007; Kitamura et al., 2015). Conditioned on 525 an image of either a nucleus-localized protein or a nucleus-excluded protein, we used CELL-Diff 526 to append a short peptide either on the N- or C-terminus of GFP. In this way, we generated 200 527 potential NLS and NES sequences, see Appendix D.

528 529

486

487

488

489

490 491

499

500 501 502

504

- 7 CONCLUSION
- 530 531

532 This paper proposes CELL-Diff, a unified diffusion model that facilitates the transformation between 533 protein sequences and microscopy images. Given cell morphology images as conditional inputs, 534 CELL-Diff generates protein images from protein sequences. Conversely, it can generate protein sequences based on microscopy images. The objective function of CELL-Diff is constructed by in-536 tegrating continuous and discrete diffusion models. Experimental results on the HPA and OpenCell datasets demonstrate that CELL-Diff produces accurate protein images with higher resolvability than previous methods. Potential applications, including virtual screening of protein localization 538 signals, virtual staining, and protein localization signal generation, make CELL-Diff a valuable tool for investigating subcellular protein localization and interactions.

540 REFERENCES

578

579

580 581

582

583

- José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning.
 Bioinformatics, 33(21):3387–3395, 2017.
- Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.
- 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.
- Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang
 Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.
- Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative
 image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.
- Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wave-grad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NsMLjcFa080.
- Ting Chen, Ruixiang ZHANG, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3itjR9QxFw.
- 567 Nathan H. Cho, Keith C. Cheveralls, Andreas-David Brunner, Kibeom Kim, André C. Michaelis, 568 Preethi Raghavan, Hirofumi Kobayashi, Laura Savy, Jason Y. Li, Hera Canaj, James Y. S. Kim, 569 Edna M. Stewart, Christian Gnann, Frank McCarthy, Joana P. Cabrera, Rachel M. Brunetti, Bryant B. Chhun, Greg Dingle, Marco Y. Hein, Bo Huang, Shalin B. Mehta, Jonathan S. Weiss-570 man, Rafael Gómez-Sjöberg, Daniel N. Itzhak, Loïc A. Royer, Matthias Mann, and Manuel D. 571 Leonetti. Opencell: Endogenous tagging for the cartography of human cellular organization. 572 Science, 375(6585):eabi6983, 2022. doi: 10.1126/science.abi6983. URL https://www. 573 science.org/doi/abs/10.1126/science.abi6983. 574
- Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles,
 Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–
 based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.
 - Andreas Digre and Cecilia Lindskog. The human protein atlas—spatial localization of the human proteome in health and disease. *Protein Science*, 30(1):218–233, 2021.
 - Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image
 synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.
- Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green,
 Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with
 alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter.
 Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 30, 2017.

594 Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Advances 595 in Neural Information Processing Systems, volume 33, pp. 6840–6851, 2020. 596 Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P 597 Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition 598 video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2022. 600 Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows 601 and multinomial diffusion: Learning categorical distributions. Advances in Neural Information Processing Systems, 34:12454-12465, 2021. 602 603 Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and 604 Tim Salimans. Autoregressive diffusion models. In International Conference on Learning Rep-605 resentations, 2022. URL https://openreview.net/forum?id=Lm8T39vLDTE. 606 Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for 607 high resolution images. In International Conference on Machine Learning, pp. 13213–13232. 608 PMLR, 2023. 609 610 Minghui Hu, Chuanxia Zheng, Zuopeng Yang, Tat-Jen Cham, Heliang Zheng, Chaoyue Wang, 611 Dacheng Tao, and Ponnuthurai N. Suganthan. Unified discrete diffusion for simultaneous vision-612 language generation. In The Eleventh International Conference on Learning Representations, 613 2023. URL https://openreview.net/forum?id=8JqINxA-2a. 614 Laura Isigkeit, Tim Hörmann, Espen Schallmayer, Katharina Scholz, Felix F Lillich, Johanna HM 615 Ehrler, Benedikt Hufnagel, Jasmin Büchner, Julian A Marschner, Jörg Pabel, et al. Automated 616 design of multi-target ligands by generative deep learning. Nature Communications, 15(1):7946, 617 2024. 618 John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, 619 Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate 620 protein structure prediction with alphafold. nature, 596(7873):583-589, 2021. 621 622 Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep 623 bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pp. 4171-624 4186, 2019. 625 Emaad Khwaja, Yun Song, Aaron Agarunov, and Bo Huang. Celle-2: Translating proteins to pic-626 tures and back with a bidirectional text-to-image transformer. Advances in Neural Information 627 Processing Systems, 36, 2024a. 628 629 Emaad Khwaja, Yun S Song, and Bo Huang. Cell-e: A text-to-image transformer for protein image prediction. In International Conference on Research in Computational Molecular Biology, pp. 630 185-200. Springer, 2024b. 631 632 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, 633 2014. 634 Akira Kitamura, Yusaku Nakayama, and Masataka Kinjo. Efficient and dynamic nuclear localization 635 of green fluorescent protein via rna binding. Biochemical and Biophysical Research Communica-636 tions, 463(3):401-406, 2015. 637 638 Rainer H Köhler, Warren R Zipfel, Watt W Webb, and Maureen R Hanson. The green fluorescent 639 protein as a marker to visualize plant mitochondria in vivo. The plant journal, 11(3):613–621, 640 1997. 641 Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan 642 dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of 643 protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. 644 645 Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models 646 generate functional protein sequences across diverse families. Nature Biotechnology, 41(8):1099-647 1106, 2023.

- Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv* preprint arXiv:2111.09734, 2021.
- Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob
 Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and
 editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.
- Carlos-Andres Palma, Marco Cecchini, and Paolo Samorì. Predicting self-assembly: from empirism to determinism. *Chemical Society Reviews*, 41(10):3713–3730, 2012.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.
- Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pp. 8599–8608. PMLR, 2021.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
 models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen,
 and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.
- ⁶⁷¹ Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention-*⁶⁷³ *MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceed-*⁶⁷⁴ *ings, part III 18*, pp. 234–241. Springer, 2015.
- H Tomas Rube, Chaitanya Rastogi, Siqian Feng, Judith F Kribelbauer, Allyson Li, Basheer Becerra, Lucas AN Melo, Bach Viet Do, Xiaoting Li, Hammaad H Adam, et al. Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nature biotechnology*, 40(10):1520–1527, 2022.
- Nicole Maria Seibel, Jihane Eljouni, Marcus Michael Nalaskowski, and Wolfgang Hampe. Nuclear
 localization of enhanced green fluorescent protein homomultimers. *Analytical biochemistry*, 368 (1):95–99, 2007.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv* preprint arXiv:2010.02502, 2020.
- The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46 (5):2699–2699, 2018.
- ⁶⁸⁸
 ⁶⁸⁹
 ⁶⁸⁹ Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.
- 692 693 694

657

663

- 695
- 696
- 697
- 698
- 699
- 700
- 701

A IMPLEMENTATION OF DISCRETE DIFFUSION MODEL

The training and sampling process of the discrete diffusion model OA-ARDM (Hoogeboom et al., 2022) can be facilitated through a masking operation. Denote C as the categorical distribution, the training and sampling algorithms are shown in Algorithm 1 and Algorithm 2, respectively. For each training iteration, we first sample a time step t from $U(1, \ldots, D)$, and a random ordering σ from $U(S_D)$. Subsequently, we generate a mask m based on the index i such that $\sigma(i) < t$. We then apply the network \mathbf{f}_{θ} , which takes $\mathbf{m} \odot \mathbf{S}$ as input, and predicts the masked values $(1 - \mathbf{m}) \odot \mathbf{S}$.

Algorithm 1 Training OA-ARDM

Require: Network \mathbf{f}_{θ} , datapoint **S**. **Ensure:** $L_{\text{OA-ARDM}}$. 1: Sample $t \sim \mathcal{U}(1, \dots, D), \sigma \sim \mathcal{U}(S_D)$.

1: Sample $t \sim \mathcal{U}(1, ..., D), \delta \sim \mathcal{U}(S_D)$. 2: Compute $\mathbf{m} \leftarrow (\sigma < t)$. 3: Compute $\mathbf{l} \leftarrow -(1 - \mathbf{m}) \odot \log \mathcal{C}(\mathbf{S} | \mathbf{f}_{\theta}(\mathbf{m} \odot \mathbf{S}))$. 4: $L_{\text{OA-ARDM}} \leftarrow \frac{1}{D - t + 1} \text{sum}(\mathbf{l})$.

Algorithm 2 Sampling from OA-ARDM

720 **Require:** Network f_{θ} . 721 Ensure: Sample S. 722 1: Initialize $\mathbf{S} = \mathbf{0}$, sample $\sigma \sim \mathcal{U}(S_D)$. 723 2: for $t = 0, 1, 2, \dots, D$ do 724 $\mathbf{m} \leftarrow (\sigma < t)$ and $\mathbf{n} \leftarrow (\sigma = t)$. 3: 725 $\mathbf{S}' \sim \mathcal{C}(\mathbf{S} | \mathbf{f}_{\theta}(\mathbf{m} \odot \mathbf{S})).$ 4: 726 5: $\mathbf{S} \leftarrow (1-\mathbf{n}) \odot \mathbf{S} + \mathbf{n} \odot \mathbf{S}.$ 727 6: end for 728

B PROTEIN IMAGE GENERATION

We present more protein image generation results. The results on the HPA and OpenCell datasets
are shown in Figure 7 and Figure 8, respectively. From these results, we observe that CELL-Diff
is capable of generating realistic protein images with high accuracy, enabling the discernment of
fine details. Compared to CELL-E2, CELL-Diff produces images with higher resolvability, which
provides better clarity of detailed localization structures.

C VIRTUAL STAINING RESULTS

Here, we present additional virtual staining results. Figure 9 shows virtual staining using data from the HPA dataset, and Figure 10 shows results using data from the OpenCell dataset. These results demonstrate that CELL-Diff generates accurate staining images compared to real ones, offering an efficient approach for simultaneous visualization of multiple biological features within the same sample.

744 745 746

747

710

711 712

713

714

715

729 730

731

737 738

739 740

741

742

743

D LOCALIZATION SIGNAL GENERATION

We use CELL-Diff to generate protein localization signals. Specifically, we select the images in Figure 11 as the conditional input for generating NLS and NES signals. Using the CELL-Diff model, we generate short amino acid sequences positioned at the N-terminus (before the GFP sequence) and the C-terminus (after the GFP sequence). A total of 100 potential sequences are generated for each signal type, consisting of 50 N-terminus and 50 C-terminus sequences. Generated NLS and NES sequences are summarized in Table 3 and Table 4, respectively.

754



Figure 7: Visual results of protein image generation on HPA dataset.



Figure 8: Visual results of protein image generation on OpenCell dataset.



Figure 9: Virtual staining using HPA dataset.



Figure 10: Virtual staining using OpenCell dataset.

Index N-terminus C-terminus 1 AKSEK PSPFVM 2 KKVFS LVTLAERP 3 KNPTDS LVKLAERD 4 ENFTAS LPALAERR 5 ENPTAR GVKLAERD 6 ANLTAS PVKLAERK 7 ENRTAR FIGVPFGGFIF 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPVEEGDEKRQE 10 ENGGAS AKGUEGDLKFEE 11 SDNSSAGF KKKVLKGDLKFEE 12 SDIENAEK LRPVAEGGEKFEE 13 KKKERIEQF RTFLRPRKVKMEERE 14 KSCERQVLF RTFLRPRKVKMEERE 15 GDIDCSEKF RTFLRPRKVKMEERE 16 SDRERITTH IEKIKRPRSSNAETLY 17 IIFDPGRQKTRLIN IEKIKRPRSSNAETLY 10 LAFIAGRGERKKK IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 23 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 24 S	73			
Table 3: Generated NLS sequences. Index N-terminus C-terminus 1 AKSEK PSPFVM 2 KKVES LVTLAERP 3 KNPTDS LVTLAERP 4 ENFTAS LPALAERR 5 ENPTAR GVKLAERD 6 ANLTAS PVKLAERK 7 ENRTAR FIGVFPGGFIF 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDINSSAGF KKKVLKGGLKFEE 13 KKKERIEQF RTFLRPRKVKMEQRE 14 KSCERQVLF RTFLRPRKVKMEQRE 15 GDIDCSEKF RTFLRPRKVKMEQRE 16 SDRERITTH IEKIKRPRSSNAETTY 17 IFDORGKQKRLKK IEKIKRPRSSNAETTY 18 ASFYETRYERLINN IEKIKRPRSSNAETTY 20 CIFDIRQKTRIJN IEKIKRPRSSNAETTY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETTY 22 SULKVDQEVKLKVDS <th>74</th> <th></th> <th></th> <th></th>	74			
Index N-terminus 1 AKSEK PSPFVM 2 KKVES LVTLAERP 3 KNPTDS LVKLAERD 4 ENFTAS LPALAERR 5 ENPTAR GVKLAERD 6 ANLTAS PVKLAERK 7 ENRTAR FIGVPF0GFIF 8 ENRTDS PLAFDRMKFIL 9 KNRTAL AVPVEEGDEKRQE 10 ENGGAS AKQUEGDLKFEK 11 SDNSSAGF KKKVLKGDLKFEK 12 SDIENIAEK LRPVAEGGEKFEE 13 KKKERLEQF RTFLRPRVVKMEQRE 14 KSCERQVLF RTFLRPRVKMEQRE 15 GDIDCSEKF RTFLRPRKVKMEQRE 16 SDRERITTH EKIKRPRSSNAETTY 19 LAFLACRGERKKK IEKIKRPRSSNAETTY 19 LAFLACRGERKKK IEKIKRPRSSNAETTY 20 CIFDIRQKTLIN IEKIKRPRSSNAETTY 21 SSLLKVDQEVKKKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKKKVDS I	75		Table 3: Generated N	LS sequences
Index C-terminus 1 AKSEK PSPFVM 2 KKVES LVTLAERP 3 KNPTDS LVKLAERD 4 ENFTAS LPALAERR 5 ENPTAR GVKLAERD 6 ANLTAS LPALAERR 7 ENRTAR FIGVFPGGFIF 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPVEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDNSSAGF KKKVLKGDLKFEK 12 SDIERIAEK LRPVAEGDEKFQE 13 KKERREQF RTFLRPRVKMEQRE 14 KSCERQVLF RTFLRPRVKMEQRE 15 GDIDCSEKF RTFLRPRSNAETEY 16 SDRERITH EKIKRPRSSNAETEY 17 IIFDPGRQKRLKK IEKIKRPRSSNAETWY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETWY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETWY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETUY 23 SLLKVDQEVKLKVDS IE	76	T. 1.	N tomin	C tamina
1 AKSEK PSPFVM 2 KKVES LVTLAERP 3 KNPTDS LVKLAERD 4 ENFTAS LPALAERR 5 ENPTAR GVKLAERD 6 ANLTAS PVKLAERK 7 ENRTAR FIGVPGGFIF 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPVEEGDEKRQE 10 ENGGAS AKGUEGDLKFEE 11 SDNSSAGF KKKVLKGDLKFEK 12 SDIERIAEK LRPVAEGGEKFEE 13 KKKERLEQF RTFLRPKVKMEERE 14 KSCERQVLF RTFLRPKVKMEERE 15 GDDCSEKF RTFLRPKSNAETRY 16 SDRERITTH IEKIKRPRSSNAETRY 17 IIFDPGRQKRLKK IEKIKRPRSSNAETRY 18 ASFYETRYERLIN IEKIKRPRSSNAETRY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 23	77	Index	N-terminus	C-terminus
2 KKVES LVTLAERP 3 KNPTDS LVKLAERD 4 ENFTAS LPALAERR 5 ENPTAR GVKLAERD 6 ANLTAS PVKLAERK 7 ENRTAR FIGVFPGGFIF 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPVEEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDNSAAGF KKKVLKODLKFEK 12 SDIERIAEK LRPVAEGGEKFEE 13 KKKEREOF RTFLRPKVKMEERE 14 KSCERQYLF RTFLRPKVKMEERE 15 GDIDCSEKF RTFLRPKVKMEERE 16 SDRERITTH IEKIKRPRSSNAFTEY 17 IIFDPGRQKKLKK IEKIKRPRSSNAFTEY 18 ASFYETRYERLIN IEKIKRPRSSNAFTY 19 LAFIAGGERKKK IEKIKRPRSSNAFTY 19 20 CIFDIRQKTRLIN IEKIKRPRSSNAFTY 12 SSLLKVDQEVKLKVDS IEKIKRPRSSNAFTY 12 SSLLKVDQEVKLKVDS IEKIKRPRSSNAFTY 12 SSLLKVDQEVKLKVDS IEKIKRPRSSNAFTY 12 <th>78</th> <th>1</th> <th>AKSEK</th> <th>PSPFVM</th>	78	1	AKSEK	PSPFVM
3 KNPTDS LVKLAERD 4 ENFTAS LPALAERR 5 ENPTAR GVKLAERD 6 ANLTAS PVKLAERK 7 ENRTDS PIAFDRMKFIL 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPVEEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDNSSAGF KKKVLKGDLKFEE 12 SDIERIAEK LRPPKVKMEQRE 13 KKKERIEOF RTFLRPKVKMEQRE 14 KSCERQYLF RTFLRPKVKMEQRE 15 GDIDCSEKF RTFLRPKVKMEQRE 16 SDRERITTH IEKIKRPRSSNAFTEY 18 ASFYETRYERLTN IEKIKRPRSSNAFTY 19 LAFIAGGERKKK IEKIKRPRSSNAFTLY 20 CIFDIRQKTRLIN IEKIKRPRSSNAFTLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAFTLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAFTLY 23 SSLLKVDQEVKLKVDS IEKIKRPRSSNAFTLY 24 SSLLKVDQEVKLKVDS IEKIKRPRSSNAFTLY 25 SKVNRVTIVRERKGOVRNYYE KSRACTNPKEEERSNILKK <t< th=""><th>79</th><th>2</th><th>KKVES</th><th>LVTLAERP</th></t<>	79	2	KKVES	LVTLAERP
4 ENFTAS LPALAERR 5 ENPTAR GVKLAERD 6 ANLTAS PVKLAERK 7 ENRTAR FIGVFPGGFIF 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPVEEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDNSSAGF KKKVLKGDLKFEK 12 SDIERIAEK LRPVAEGGEKFEE 13 KKKERIEOF RTFLRPKVKMEERE 14 KSCERQYLF RTFLRPKVKMEERE 15 GDIDCSEKF RTFLRPKVKMEERE 16 SDRENITTH IEKIKPRSSNAETLY 17 IIFDPGRQKRLKK IEKIKPRSSNAETLY 18 ASFYETRYERLIN IEKIKPRSSNAETLY 20 CIFDIRQKTRLIN IEKIKPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 23 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 24 SKLKVDEVELKVDS YACKEEDEEDEDDDNM 25 SSFEICRLVFLVFGMLCPA NVKYCRENPLEEPSPJAKTK 26 SKNDVIRLQRKRGVRDYYNE KSACCVTNTPEGEASILNS	30	3	KNPTDS	LVKLAERD
5 ENPTAR GVKLAERD 6 ANLTAS PVKLAERK 7 ENRTAR FIGVPFGGFF 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPVEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDINSSAGF KKKVLKGDLKFEK 12 SDIERIAEK LRPPKVKMEQRE 13 KKKERIEQF RTFLRPKVKMEQRE 14 KSCERQYLF RTFLRPKVKMEQRE 15 GDIDCSEKF RTFLRPKVKMEQRE 16 SDERITTH IEKIKPRSSNAETEY 17 IIFDGRQKRLKK IEKIKPRSSNAETVR 19 LAFIAGRGERKKK IEKIKPRSSNAETLY 20 CIFDRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 23 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 24 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 25 SSFLICKUPGVRSRDPEM NVRARIVNCLEVEENPSNKLE 26 SKNDVILQRKRRGVSRDPEM NVKARIVNCLEVEENPSNKLE 27 PLIEVLREAVGRSGVRRDYYE	21	4	ENFTAS	LPALAERR
6 ANLTAS PVKLAERK 7 ENRTAR FIGVFPGGFIF 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPVEEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDNSSAGF KKKVLKGDLKFEK 12 SDIERIAEK LRPVAEGGEKFEE 13 KKKERLEQF RTFLRPPKVKMEQRE 14 KSCERQYLF RTFLRPPKVKMEQRE 15 GDIDCSEKF RTFLRPPKVKMEQRE 16 SDRERTTH IEKIKRPRSINAETEY 17 IIFDPGRQKRLKK IEKIKRPRSINAETTY 19 LAFIAGRGERKKK IEKIKRPRSSNAETLY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSINAETLY 23 SSLLKVDQEVKLKVDS VKYCEENPLEEPESDIDNM 24 SSLLKVDQEVKLKVDS VKXCELPEEDESDDDNM 25 SSFEICRLVFLVFGMLCPA NVKYCEENPLEPEPESINKLE 26 SKNDVIRLQRKRPGVSRDPEM NVKYCEENPLEPESENLNKAE 27 PLIEVLREAVGRSGVRRDYYE KSACVNTFEGEASILNSLL 28 </th <th>20</th> <th>5</th> <th>ENPTAR</th> <th>GVKLAERD</th>	20	5	ENPTAR	GVKLAERD
7 ENRTAR FIGVFPGGFIF 8 ENRTDS PIAFDRMKFIL 9 KNRTAL AVPVEEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDNSAGF KKKVLKGDLKFEK 12 SDIENAEK LRPVAEGGEKFEE 13 KKKERIEQF RTFLRPPKVKMEQRE 14 KSCERQVLF RTFLRPPKVKMEQRE 15 GDIDCSEKF RTFLRPRKVKMEQRE 16 SDRERITTH IEKIKRPRSSNAETEY 18 ASFYETRYERLTN IEKIKRPRSSNAETVR 19 LAFIAGRGEKKK IEKIKRPRSSNAETLY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSINAETLY 23 SSLLKVDQEVKLKVDR IEKIKRPRSINAETLY 24 SSLLKVDQEVKLKVDS YAKEELEEDEDSDDDNM 25 SSFEICRLVFLVFGMLCPA NVRARIVNGLEVEENPSNKLE 26 SKNDVIRLQRKRPGVSRDPEM NVRARIVNGLEVEENPSNKLE 27 PLIEVLREAVGRSGVRNDYE KVRACVINTTVEREKKGGKXVSNE 28 SKVNRVTTVVERKGQVYNSNE KVYACRPWKFEERESINKLAE		6	ANLTAS	PVKLAERK
8 ENRIDS PIAFDRMKHL 9 KNRTAL AVPVEEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDNSSAGF KKKVLKGDLKFEE 12 SDIERIAEK LRPVAEGGEKFEE 13 KKKERIEQF RTFLRPPKVKMEQRE 14 KSCERQYLF RTFLRPPKVKMEQRE 15 GDIDCSEKF RTFLRPPKVKMEQRE 16 SDRERITTH IEKIKRPRSINAETEY 18 ASFYETRYERLIN IEKIKRPRSINAETY 19 LAFIAGRGEKKKK IEKIKRPRSINAETIY 20 CIFDIRQKTRLIN IEKIKRPRSINAETIY 21 SSLLKVDQEVKLKVDS IEKIKRPRSINAETIY 22 SDLLKVDQEVKLKVDS IEKIKRPRSINAETIY 23 SSLLKVDQEVKLKVDR IEKIKRPRSINAETIY 24 SSLLKVDQEVKLKVDS YAKEELEEEDSDDDNM 25 SSFEICRLVFLOKGMLCPA NVKACRENPLEEPSPIAKTK 26 SKNDVIRLQRKRPGVSRDPEM NVRARIVNGLEVEENSNLL 27 PLIEVLREAVGRSOKROPYYE KSACYNKGDNLDIDNLVLE 28 SKVNRVTTVRERKGVRUYSNE KVYACRDLWKEGER	33	7	ENRTAR	FIGVFPGGFIF
9 KNRTAL AVPVEEGDEKFQE 10 ENGGAS AKGQLEGDLKFEE 11 SDNSSAGF KKKVLKGDLKFEK 12 SDIERIAEK LRPVAEGGEKFEE 13 KKKERIEQF RTFLRPFNVKMEERE 14 KSCERQVLF RTFLRPRVKMEERE 15 GDIDCSEKF RTFLRPGKVKMEQRE 16 SDRERITTH IEKIKRPRSINAETEY 17 IIFDGRQKRLKK IEKIKRPRSINAETY 19 LAFIAGRGERKKK IEKIKRPRSSNAETVR 19 LAFIAGRGERKKK IEKIKRPRSSNAETLY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSINAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSINAETLY 23 SSLLKVDQEVKLKVDS IEKIKRPRSINAETLY 24 SSLLKVDQEVKLKVDS IEKIKRPRSINAETLY 25 SSFEICRLVFLVFGMLCPA NVKYCRENPLEPEPESPIAKTK 26 SKNDVILQRKRPGVSRDPEM NVRARIVNERKEERSINLKAE 27 PLIEVLREAVGRSGVRNYYE KSRACVTNTPEGEASILNSLL 28 SKVNRVTTVRERKGVRYVSNE KVYACRPWKFEERESNLNKAE 29 DVSVLOLNFGKTGRYDYH	34	8	ENRTDS	PIAFDRMKFIL
10 ENGGAS AKGQLEGDLKFËE 11 SDNSSAGF KKKVLKGDLKFEK 12 SDIERIAEK LRPVAEGGEKFEE 13 KKKERIEQF RTFLRPKVKMEQRE 14 KSCERQYLF RTFLRPKVKMEQRE 15 GDIDCSEKF RTFLRPKVKMEQRE 16 SDRERITH IEKIKRPRSINAETEY 17 IIPDPGRQKRLKK IEKIKRPRSSNAETEY 18 ASFYETRYERLTN IEKIKRPRSSNAETWY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETDY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETDY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETDY 23 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETDY 24 SSLLKVDQEVKLKVDS YAEELEEDESDDDNM 25 SFEICRLYFLVFGMLCPA NVKYCRENPLEEPESPIAKTK 26 SKNDVIRLQRKRPGVSRDPEM NVRARIVNGLEVEEPSNAKLE 27 PLEVLREAVGRSGVRRDYYE KSACYTNPEGEASILNSLL 28 SKVNRVTTVRERKGVRYDYH DKYACRDNLDRKCGNKKCACKKEKK 29 IDVSVLDLNFGKTGVRYDYHE GXYQVSSKGGGKKCACKKEKK 29 IDVSVLDLNFGKTGVRYDYHE GAKYQVSSKGGGKKCACKKEKK 29 <t< th=""><th>5</th><th>9</th><th>KNRTAL</th><th>AVPVEEGDEKFQE</th></t<>	5	9	KNRTAL	AVPVEEGDEKFQE
11SDNSSAGFKKKVLKGDLKFEK12SDIERIAEKLAPVAEGGEKFEE13KKKERLEQFRTFLRPKVKMEQRE14KSCERQYLFRTFLRPKVKMEQRE15GDIDCSEKFRTFLRPKVKMEQRE16SDREITTHIEKIKRPRSSNAETEY17IIFDPGRQKRLKKIEKIKRPRSSNAETVR18ASFYETRYERLTNIEKIKRPRSSNAETUY20CIFDIRQKTRLINIEKIKRPRSSNAETUY21SSLLKVDQEVKLKVDSIEKIKRPRSSNAETUY22SDLLKVDQEVKLKVDSIEKIKRPRSSNAETUY23SSLLKVDQEVKLKVDSIEKIKRPRSINAETUY24SSLLKVDQEVKLKVDSYAKEELEEEDESDDNM25SSFEICRLVFLVFGMLCPANVKYCRENPLEEPESPIAKTK26SKNDVIRLQRKRPGVSRDPEMNVRARIVNGLEVEENPSNKLE27PLEVLREAVGRSGVRRDYYEKSRACVTNTPEGEASILNSLL28SKVNRVTTVRERKGVRYVSNEKVYACRENKFEERSNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKXACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACQUKKCAKKEKK33PDNRIEGVLRDDVHLVYHEGKKVQVSSRGGGDKSALKVEKK34PQNEYIEEHKRKNDLYLVGEKGKYYCVSNRFOGKKCAALKELK35SDLIEVFRADYRNDLHEKGXYYCVSNRFOGKKCAALKELK36PQNEYIEEVRRADYRNLDLGVRGAKYQVSSKGGDKSALKVEKK37PQNEYQEEVRGRYDLYLVGEKGKYYQVSSNPEGKKCANHPLPK38PUNEYIELVFKADYRNLHEKGAYYCVSSNPEGKKCAALKEKK39GLLEYIEEVRRADYRNLHEKGAYYQVSSNPEGKKCALKKELK39GLLEYIEEVRRADYRNLHEKGAYQVSSKGEDNKKSINEKEKK39GLLEYIEEVRGTYDYRNVLHEKGAKYQVSSKGENKKSINEKEKK	6	10	ENGGAS	AKGQLEGDLKFEE
12 SDIERIAEK LRPVAEGGEKFEE 13 KKKERIEQF RTFLRPPKVKMEQRE 14 KSCERQYLF RTFLRPPKVKMEQRE 15 GDIDCSEKF RTFLRPGKVKMEQRE 16 SDRERITTH IEKIKRPRSINAETKY 17 IIFDPGRQKRLKK IEKIKRPRSSNAETVR 18 ASFYETRYERLIN IEKIKRPRSSNAETVR 20 CIFDIRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 23 SSLLKVDQEVKLKVDR IEKIKRPRSINAETLY 24 SSLLKVDQEVKLKVDR IEKIKRPRSINAETLY 25 SSFEICRLVFLVFGMLCPA NVKYCRENPLEEPESPIAKTK 26 SKNDVIRLQRKRGVSRDPEM NVRARIVNGLEVEENPSNKLE 27 PLEVLREAVGRSGVRRDYYE KSRACVTNTPEGEASILNSLL 28 SKVNRVTTVRERKGVRYDYHI DKYACRDLNFRKEGCRYNKTI 29 IDVSVLDLNFGKTGVRYDYHI DKYACRDLNFRKEGCRYNKTI 29 IDVSVLDLNFGKTGVRYDYHI GKKYQVSSKGGGKKCAALKEKK 30 PLNNVQRLHVEERGHRLDYAN GKSXSSGGGKKCAALKEKK 31 PDUNNURLGVLRMAPYYYY GKKYQVSSNEGGKKCANKEKKK	7	11	SDNSSAGF	KKKVLKGDLKFEK
13 KKKERIEOF RTFLRPPKVKMEQRE 14 KSCERQYLF RTFLRPKVKMEQRE 15 GDIDCSEKF RTFLRPGKVKMEQRE 16 SDRERITTH IEKIKRPRSINAETEY 17 IIFDPGRQKRLKK IEKIKRPRSSNAETEY 18 ASFYETRYERLTN IEKIKRPRSSNAETUY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETUY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETUY 22 SDLLKVDQEVKLKVDS IEKIKRPRSINAETUY 23 SSLLKVDQEVKKKVDS IEKIKRPRSINAETUY 24 SSLLKVDQEVKKKVDS IEKIKRPRSINAETUY 25 SSFEICRLVFLVFGMLCPA NVKYCRENPLEEPESPIAKTK 26 SKNDVIRLQRKRPGVSRDPEM NVKYCRENPLEEPESPIAKTK 27 PLEVLREAVGRSGVRRDYYE KSRACVTNTPEGEASILNSLL 28 SKVNKTTVRERKGVRYVSNE KYACRPWKFEERESNLNKAE 29 IDVSVLDLNFGKTGVRYDYHI DKYACRDLNFRKEECRYNKTI 28 SUNSYTFRERMDVRYDYE GAYYCKSSKGGGKKCALKKEK 29 IDVSVLDLNFGKTGVRYDYHI DKYACRDLNRKCAKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKK	38	12	SDIERIAEK	LRPVAEGGEKFEE
14 KSCERQYLF RTFLRPKVKMEERE 15 GDIDCSEKF RTFLRPGKVKMEQRE 16 SDRERITTH IEKIKRPRSLNAETKY 17 IIFDPGRQKRLKK IEKIKRPRSSNAETVR 18 ASFYETRYERLTN IEKIKRPRSSNAETWY 19 LAFIAGRGERKKK IEKIKRPRSSNAETWY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 23 SSLLKVDQEVKLKVDS IEKIKRPRSLNAETLY 24 SSLLKVDQEVKLKVDS IEKIKRPRSLNAETLY 25 SSFEICRLVFLVFGMLCPA NVKYCRENPLEEPESPIAKTK 26 SKNDVIRLQRKRPGVSRDPEM NVRARIVNGLEVEENPSNKLE 27 PLIEVLREAVGRSGVRRDYYE KSRACVTNTPEGEASILNSLL 28 SKVNRVTTVRERKGVRYVSNE KVYACRPWKFEERESNLNKAE 29 IDVSVLDLNFGGTGVRYDYHI DKYACRDLNFRKEECRYNKTI 30 PLNENVQELHVERGRHLDYAN GKSACRYNKGDNLDIDLVLE 531 GDLDVSYTFRERMDVRYDYEE GAYYCXSSKGGGKKCAALKEKK 36 PONRIELGVLRADURLVYGER GVYYCVSNKPGGKKCAALKEKK 37 PONRIELGVRRADURLVYGER <	9	13	KKKERIEOF	RTFLRPPKVKMEORE
15 GDIDCSEKF RTFLRPGKVKMEQRE 16 SDRERITTH IEKIKRPRSLNAETKY 17 IIFDPGRQKRLKK IEKIKRPRSSNAETY 18 ASFYETRYERLTN IEKIKRPRSSNAETVR 19 LAFIAGRGERKKK IEKIKRPRSSNAETLY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 23 SSLLKVDQEVEKKVDS IEKIKRPRSSNAETLY 24 SSLLKVDQEVEKKVDS YAKEELEEEDESDDDNM 25 SSFEICLVFLVFGMLCPA NVKYCRENPLEEPESPIAKTK 26 SKNDVIRLQRKRPGVSRDPEM NVRARIVNGLEVEENPSNKLE 27 PLIEVLREAVGRSGVRRDYYE KSVACVTNTGECASILNSLL 28 SKVNNVTTVRERKGVRYVSNE KVACRPWKFEERESNLNKAE 29 IDVSVLDLNFGGTGVRYDYHI DKYACSBGGKCAGKKEKK 30 PLNNVQRLHVEERGHRLDYAN GKSACRYNKGDNLDIDNLVLE 31 GDLOVSYTFRERMDVRYDYEE GAYYCXSSKGGGBKCAALKEKK 33 PDVNRIELGVLRDVHLYHY GKKLQVSNKHDGKKCALKEKK 34 PQNEYIEEHRKRYDLYLVYGEK GVYYCVSNKPGGKKCAALKEKK 35 PQNEYGEVRGRYDLRLVYGER<	0	14	KSCEROYLE	RTFLRPPKVKMEERE
16 SDRERITTH IRKIRPRSLNAETKY 17 IIFDPGRQKRLKK IEKIKRPRSLNAETKY 18 ASFYETRYERLTN IEKIKRPRSSNAETVR 19 LAFIAGRGERKKK IEKIKRPRSSNAETLY 20 CIFDIRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 23 SSLLKVDQEVELKVDS IEKIKRPRSSNAETLY 24 SSLLKVDQEVELKVDS YAKEELEEDESDDDNM 25 SSFEICRLVFLVFGMLCPA NVKYCRENPLEEPESPIAKTK 26 SKNDVIRLQRKRFGVSRDPEM NVRARIVNGLEVENPSNKLE 27 PLIEVLREAVGRSGVRDYYE KSRACVTNTPEGEASILNSLL 28 SKVNRVTTVRERKGVRYVSNE KVYACRPWKFEERESNLNKAE 29 IDVSVLDLNFGKTGVRYDYHI DKYACRDLNFRKECRYNKTI 30 PLNNVQRLHVEEGGRGPGVRIVYY GKKYQVSSKGGGKCAALKVEKK 31 GDLDVSYTTRERMDWRYDYEE GAYYCXSNKHDGKKCAALKEKK 34 PQNEYIEEHRKRYDLYLVYGEK GKYYRVSSNPEGKKCAALKEKK 35 PQNEYQEEVRGRYDLYLVYGEK GKYYCVSNNPGGKKCAALKEKK 36 PQNEYIEEHRKRYDLYLVYGEK GKYYCVSNNPGGKKCAALKEKK 37		15	GDIDCSEKE	RTFLRPGKVKMEORE
1010111111111111111111121111131212141414151212151212161212171212181412191412191414101414111412151413141414151415141614171417141814191419141914111411141214141415161614171417141814191419141914191411141114111412151316141614161511161417161716171618161916191619161116111611161116<)1	16	SDRFRITTH	IFKIKRPRSI NAFTKY
11 In Dr Niklenki Infiktion 13 ASFYETRYERLTN IEKIKRPRSSNAETTVR 19 LAFIAGRGERKKK IEKIKRPRSSNAETTVR 19 LAFIAGRGERKKK IEKIKRPRSSNAETTVR 20 CIFDIRQKTRLIN IEKIKRPRSSNAETLY 21 SSLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 22 SDLLKVDQEVKLKVDS IEKIKRPRSSNAETLY 23 SSLLKVDQEVKLKVDS YAKEELEEEDESDDDDNM 24 SSLLKVDQEVELKVDS YAKEELEEEDESDDDNM 25 SSFEICRLVFLVFGMLCPA NVKYCRENPLEEPSNKLE 26 SKNDVIRLQRKRPGVSRDPEM NVRARIVNGLEVEENPSNKLE 27 PLIEVLREAVGRSGVRRDYYE KSRACVTNTPEGEASILNSLL 28 SKVNRVTTVRERKGVRYDSNE KVYACRPLKFEERESNLNKAE 29 IDVSVLDLNFGKTGVRYDYHI DKYACRDLNFRKEECRYNKTI 30 PLNENVQRLHVEERGHRLDYAN GKSLQVSNKHODKKCALKKEK 31 GDLDVSYTFRERMDVRYDYEE GAYYCKSSKGGGDKSALKVEKK 32 PLNERNEGQRGRGVRIVYYY GKKLQVSNKHDGKKCALKKEK 33 PDVNRIELGVLRDDVHLVYGEK GKYYVSSNPEGKKCANHPLK 34 PQNEYIEEHRKRYDLYLVYGEK GVYYCVSNKPGGKKCALKKEK 35)2	17		IEKIKRPRSSNAFTEV
10LAFIAGRGERKKKLEKIKRPRSSNAETWY20CIFDIRQKTRLINIEKIKRPRSSNAETUY21SSLLKVDQEVKLKVDSIEKIKRPRSSNAETUY22SDLLKVDQEVKLKVDSIEKIKRPRSSNAETUY23SSLLKVDQEVKLKVDRIEKIKRPRSSNAETUY24SSLLKVDQEVKLKVDRIEKIKRPRSINAETUY25SSFEICRLVFLVFGMLCPANVKYCRENPLEEPESPIAKTK26SKNDVIRLQRKRPGVSRDPEMNVRARIVNGLEVEENPSNKLE27PLIEVLREAVGRSGVRRDYYEKSRACVTNTPEGEASILNSLL28SKVNRVTTVRERKGVRYVSNEKVYACRPWKFEERESNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK34PONNURLHVEERGHRLDYANGKKYQVSSNGGGDKSALKVEKK35PDNNRIELGVLRDDVHLYYHEGKKYQVSSNPEGKKCINKPLLK36PONEYIEEHRKRYDLYLVGEKGVYYCVSNKHDGKKCALKKELK37PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSGGDKCLLLKEKV39GLLEYIEEVRGRADLYLVLHERGAKYQVSSNGENKKSINEKEKK39GLLEYIEEVRGRADLYLVLHERGAKYQVSSNGENKKSINEKEKK34PQNEYIEELRKRYDYRVLHYKGAYYCVSSNPTGQKCINAVEKK36PLNEYIELVFKRADYRRDLHEKEAYYQVSSKEGHAKKSINEKEKK39GLLEYIEEVRGRADLYLVLHERGAKYQVSSNGENKKSINEKEKK34PUNEYIELFYFKADYRRVLHYRGAYYCVSSNPTGQKCINAVEKK35PLNEYIELFYFKADYRRVLGVRYKEAEEYKLKYYLAPACHTEEIDS36FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH	3	18	A SEVETRVERI TN	IEKIKRPRSSNAETVR
1919101020CIFDIRQKTRLINIEKIKRPRSSNAETLY21SSLLKVDQEVKLKVDSIEKIKRPRSSNAETLY22SDLLKVDQEVKLKVDRIEKIKRPRSSNAETLY23SSLLKVDQEVKLKVDRIEKIKRPRSSNAETLY24SSLLKVDQEVKLKVDSYAKEELEEDESDDDNM25SSFEICRLVFLVFGMLCPANVKYCRENPLEEPESPIAKTK26SKNDVIRLQRKRPGVSRDPEMNVRARIVNGLEVEENPSNKLE27PLIEVLREAVGRSGVRRDYYEKSRACVTNTFEGEASILNSLL28SKVNRVTTVRERKGVRYVSNEKVYACRPWKFEERESNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAKKEKK32PLNERNEGQRGRGVRIVYYYGKKYQVSSKGGGBKSALKVEKK34PQNEYIEEHRKRYDLYLVYGEKGKYYRVSSNPEGKKCINKPLLK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKKELK36PQNELIEEVRRRYWYRDLGVRGAKYQVSSSGGGDKCILLKEKK37PONEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKELK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKGGDKCILLKEKK39GLLEYIEEVRGRADLYLVLHERGAKYQVSSNFIGQKCINAVEKK34PQNELIEQVFGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK35PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS36PANELIEQVFGRADLYLVLHERGAKYQVSSNPIGQKCINAVEKK36GULEYIEEVRGADFRVLGVRYKEAEEYKLKYYLAPKHTEEIDS37GULEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS36PARDSQKSIISPVINHYKFYSRPQRKAQPAQPADEVAEKADEPMEH37 <th>)4</th> <th>10</th> <th>I AFIAGPGEPKKK</th> <th>IEKIKPPPSSNAETWV</th>)4	10	I AFIAGPGEPKKK	IEKIKPPPSSNAETWV
20CITCHINGIEKIKRPRSIGNAETLY21SSLLKVDQEVKLKVDSIEKIKRPRSGNAETLY22SDLLKVDQEVKLKVDSIEKIKRPRSGNAETLY23SSLLKVDQEVKLKVDRIEKIKRPRSINAETLY24SSLLKVDQEVKLKVDRIEKIKRPRSINAETLY25SSEEICRLVFLVFGMLCPANVKYCRENPLEEPESPIAKTK26SKNDVIRLQRKRPGVSRDPEMNVRARIVNGLEVEENPSNKLE27PLIEVLREAVGRSGVRDYYEKSRACVTNTPEGEASILNSLL28SKVNVTTVRERKGVRYVSNEKVYACRPWKFEERESNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK3632PLNERNEGQRGRPGVRIVYYYGKKLQVSNKHDGKKCALKKELK3733PDVNRIELGVLRDDVHLVYGEKGKYYRVSSNPEGKKCANLKELK3644PQNEYIEEHRKRYDLYLVYGERGVYYCVSNKPGGKKCAALKEKK3959PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK39GLLEYIEEVRRADYRRDLHEKEAYYQVSSKSGGDKCILLKEKK39GLLEYIEEVRGRADLYLVKGEKGAKYQVTSSNPTGQKCINAVEKK40VLNEYEEEVRGTDYDYRRVLHPKGAKYQVSSKGENKKSINEKEKK41PQNELIEEVRGADLYLVLHERGAKYQVSSKGENKKSINEKEKK4234PLNEYIELHFKRADYRRDLHEKEAYYQVSSKSEGNKCILLKEKV39GLLEYIEEVRGADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTDYDYRLVLHPKGAKYQVSSKGENKKSINEKEKK41PQNELIEEVFGRADLYLVLHERGAKYQVSSNPTGQKCINAVEKK4241PQNELIEEVFGRADLYKENFRNQYDNVTVWSPQGRIHQIEYAM	5	20	CIEDIDOVTDI IN	IEVIVDDOSNAETI V
21SSLLKVDQEVKLKVDSIEKIKRPRSSNAETDY22SSLLKVDQEVKLKVDSIEKIKRPRSSNAETDY23SSLLKVDQEVKLKVDSIEKIKRPRSSNAETDY24SSLLKVDQEVELKVDSYAKEELEEEDESDDDNM25SSFEICRLVFLVFGMLCPANVKYCRENPLEEPESPIAKTK26SKNDVIRLQRKRPGVSRDPEMNVRARIVNGLEVEENPSNKLE27PLIEVLREAVGRSGVRDYYEKSRACVTNTPEGEASILNSLL28SKVNRVTTVRERKGVRYVSNEKVYACRPWKFEERESNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDLNIDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK32PLNERNEGQRGRPGVRIVYYYGKKLQVSSKGGGKKCALKVEKK33PDVNRIELGVLRDDVHLVYHEGKKLQVSSNFDGKKCANKPLLK34PQNEYIEEHRKRYDLYLVYGEKGVYYCVSSNPEGKKCANHPLPK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSSNPEGKKCANHPLPK36PQNEYQEEVRGRTDLRLVYGEKGKKYQVSSKGGDKCILLKEKK37PQNEYQEEVRGRTDLRLVYGEKGKKYQVSSKGGDKCILLKEKK38PLNEYIELVFKRADYRDLHEKEAYYQVSSKSGGMCCILLKEKK39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRRVLHPKGAKYQVSSKGENKKSINEKEKK41PQNELIEEVRGRADLYLVLHERGAKYQVSSNPEGKCANHPLPK42PLNEYIELVFKRADYRRVLHPKGAKYQVSSKGENKKSINEKEKK4440VLNEYEEEVRGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ44PATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSIVLRASSDELTRPQQKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRI	6	20		
22SDLLKVDQEVKLKVDSIEKIKNPRSINAETD123SSLLKVDQEVKLKVDRIEKIKNPRSINAETD124SSLLKVDQEVELKVDSYAKEELEEEDESDDDNM25SSFEICRLVFLVFGMLCPANVKYCRENPLEEPESPIAKTK26SKNDVIRLQRKRPGVSRDPEMNVRARIVNGLEVEENPSNKLE27PLIEVLREAVGRSGVRRDYYEKSRACVTNTPEGEASILNSLL28SKVNRVTTVRERKGVRYVSNEKVYACRPWKFEERESNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK32PLNERNEGQRGRGRGVRIVYYYGKKLQVSSKGGGDKSALKVEKK34PQNEYIEEHRKRYDLYLVYGEKGKKLQVSSNKHDGKKCALKKELK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK36PQNELIEEVRKRYWYRLDLGVRGAKYQVSSSGGDKCILLKEKK37PQNEYQEEVRGRYDLYLVKGEKGKKYQVSSSGGDKCILLKEKK38PLNEYIELPKRADYRRDLHEKEAYYQVSSKGENKKSINEKEKK39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYRVWLEVKEDEPELKREEIEKATKELDS42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS44PATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH4748 <th>7</th> <th>21</th> <th>SOLLANDQENKKKVDS</th> <th></th>	7	21	SOLLANDQENKKKVDS	
25SSLLKVDQEVELKVDKIEKIKRRSLNABILI24SSLLKVDQEVELKVDSYAKEELEEEDESDDDNM25SSFEICRLVFLVFGMLCPANVKYCRENPLEPESPIAKTK26SKNDVIRLQRKRPGVSRDPEMNVRARIVNGLEVEENPSNKLE27PLIEVLREAVGRSGVRRDYYEKSRACVTNTPEGEASILNSLL28SKVNRVTTVRERKGVRYDSNEKVYACRDNFRKEECRYNKTI39DVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEGAYYCKSSKGGGKKCAGKKEKK32PLNERNEGQRGRPGVRIVYYYGKKYQVSSKGGGKKCALKKELK34PQNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK35PQNRIELGVLRDDVHLVYGEKGKYYRVSSNPEGKKCINKPLLK36PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK37PQNEYQEEVRGRYDLYLVKGEKGKKYQVSSSGGDKCILLKEKK39GLLEYIEEVRKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV39GLLEYIEEVRGRADLYLVHFKGAKYQVSSNPTGQKCINAVEKK40VLNEYEEEVRGTYDYRRVLHPKGAKYQVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRRVLHPKGAKYQVSSNPTGQKCINAVEKK42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS44PATDSQKSIISPVINHYKFIYSRPQRKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH48RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH <t< th=""><th>0</th><th>22</th><th></th><th>IENING NASINAE I DI</th></t<>	0	22		IENING NASINAE I DI
24SSLLKVDQE VELKVDSTAKEELEEEDESDDDINM25SSFEICRLVFLVFGMLCPANVKYCRENPLEEPESPIAKTK26SKNDVIRLQRKRPGVSRDPEMNVRARIVNGLEVEENPSNKLE27PLIEVLREAVGRSGVRRDYYEKSRACVTNTPEGEASILNSLL28SKVNRVTTVRERKGVRYVSNEKVYACRPWKFEERESNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKECCRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK32PLNERNEGQRGRGRCVRIVYYYGKKYQVSSKGGGDKSALKVEKK33PDVNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK34PQNEYIEEHRKRYDLYLVYGEKGVYYCVSNKPGGKKCAALKEKK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK36PQNELIEEVRKRYWYRLDLGVRGAKYQVTSSPEGKKCANHPLPK37PQNEYQEEVRGRTDLRLVYGEKGKKYQVSSSGGDKCILLKEKK38GLLEYIEEVRGRADLYLVKGEKGKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGRADLYLVKGEKGAKYQVSSSREGKKCILLKEKV39GLLEYIEEVRGRADLYLVLHERGAKYQVSSNPTGQKCINAVEKK40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPRQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS4440VLNEYEEEVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS44PATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAE		25		
25SSFEICRLVFLVFGMLCPANVRARIVNGLEVEENPESPIAKTK26SKNDVIRLQRKRPGVSRDPEMNVRARIVNGLEVEENPSNKLE27PLIEVLREAVGRSGVRRDYYEKSRACVTNTPEGEASILNSLL28SKVNRVTTVRERKGVRYVSNEKVYACRPWKFEERESNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE53GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK5429PLNERNEGQRGRPGVRIVYYGKKLQVSSKGGGDKSALKVEKK5531GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGDKSALKVEKK5632PLNERNEGQRGRPGVRIVYYGKKLQVSNKHDGKKCALKKELK5733PDVNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK5834PQNEYIEEHRKRYDLYLVGEKGVYYCVSNNFPGGKKCAALKEKK5935PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK507PQNEYQEEVRGRYDLYLVKGEKGKYQVSSKSEGKKCILLKEKV539GLLEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV549GLLEYIEEVRGRADLYLVLHERGAKYQVSSNPTGQKCINAVEKK5541PQNELIEQVFGRYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK5643GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS7744PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH7643PONDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH7744PPATDSQKSIISPVINHYKFIYSRPQRKAQPAQPADEVAEKADEPMEH7648RENQYDNDVTWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH7744PPATDSQKSIISPVINHYKFIYSRPQRKAQPAQPAD	9	24	SSELCDI VELVECNI CDA	
126SKNDVIRLQKKRPGVSKDPEMNVKARIVNGLEVEENPSNKLE27PLIEVLREAVGRSGVRRDYYEKSRACVTNTPEGEASILNSLL28SKVNRVTTVRERKGVRYVSNEKVYACRPWKFEERESNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGDKSALKVEKK32PLNERNEGQRGRPGVRIVYYYGKKYQVSSKGGGDKSALKVEKK33PDVNRIELGVLRDDVHLVYHEGKKYQVSSNFDGKKCALKKELK34PQNEYIEEHRKRYDLYLVYGEKGVYYCVSNKPGGKKCAALKEKK35PQNEYIEEHRKRYDLRLVYGERGVYYCVSNKPGGKKCAALKEKK36PQNELIEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK37PQNEYQEEVRGRTDLRLVYGEKGKKYQVSSSGGGDKCILLKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV39GLLEYIEEVRGRADLYLVLHERGAKYQVSSNPTGQKCINAVEKK40VLNEYEEEVRGTDYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49FRNQYDNDVTVWSPQG	00	25	SSFEICKLYFLYFGMLCPA	
2121PLIEVLREAVGRSGVRKDYYEKSRACVINIPEGEASILNSLL28SKVNRVTTVRERKGVRYVSNEKVYACRPWKFEERESNLNKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK32PLNERNEGQRGRPGVRIVYYGKKYQVSSKGGGDKSALKVEKK33PDVNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK34PQNEYIEEHRKRYDLYLVYGEKGVYYCVSNKPGGKKCAALKKELK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK36PQNELIEEVRKRYWYRLDLGVRGAKYQVTSSPEGKKCILLKEKK37PQNEYQEEVRGRYDLYLVKGEKGKYQVSSKGGDKCILLKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSGGGKCILLKEKK39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS44PAATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH40RPQRKAQPAQPADEVAEKADEPMEHRPQRKAQPAQPADEVAEKADEPMEH41PATDSQKSIISPVINHYKFIYSRPQRKAQPAQPADEVAEKADEPMEH42PANDVTVWSPQGR	01	20		N V KAKI V NGLE V EENPSINKLE
28SKVNKVTTVREKKGVKTVSNEKVTACRPWKFEERESNLINKAE29IDVSVLDLNFGKTGVRYDYHIDKYACRDLNFRKEECRYNKTI30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK32PLNERNEGQRGRPGVRIVYYYGKKYQVSSKGGGDKSALKVEKK33PDVNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK34PQNEYIEEHRKRYDLYLVYGEKGKYYRVSSNPEGKKCINKPLLK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKKEKK36PQNELIEEVRKRYWYRLDLGVRGAKYQVTSSPEGKKCANHPLPK37PQNEYQEEVRGRDLHLVYGEKGKKYQVSSKGGDKCILLKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKK39GLLEYIEEVRGRADLYLVHFRGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS44PATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH40FRNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH41PARNOVDNVVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH42FRNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH44PERNQYDNDVTVSPQG	02	27	PLIEVLKEAVGKSGVKKDY YE	KSRAC V IN I PEGEASILINSLL
29IDVSVLDLNFGK1GVRYDYHIDKYACRDLNFRKEECKYNKII30PLNNVQRLHVEERGHRLDYANGKSACRYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK32PLNERNEGQRGRPGVRIVYYGKKYQVSSKGGGDKSALKVEKK33PDVNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK44PQNEYIEEHRKRYDLYLVYGEKGKYYRVSSNPEGKKCINKPLLK5590PQNEYIEEHRKRYDLYLVYGERGVYYCVSNKPGGKKCAALKEKK569090PUELIEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK5790PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK5891PQNEYQEEVRGRYDLYLVKGEKGKKYQKSSSGGGDKCILLKEKK5993PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKK5093GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK509494PLNEYIELVFKRADYRRVLHPKGAYYCVSSNPTGQKCINAVEKK5494PUNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ5542PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS5643GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS5744PPATDSQKSIISPVINHYKFIYSRPQRKAQPAQPADEPAEKADEPMEH5845PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEPAEKADEPMEH5946RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH48RENQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49ERNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH40FRNQYDNDVTVWSPQGRIHQIEYA	03	28	SKVNKVIIVREKKGVKYVSNE	K V YACKPW KFEERESNLINKAE
30PLNNVQRLHVEEKGHRLDYANGKSACKYNKGDNLDIDNLVLE31GDLDVSYTFRERMDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK32PLNERNEGQRGRPGVRIVYYGKKYQVSSKGGGKKCAGKKEKK33PDVNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK34PQNEYIEEHRKRYDLYLVYGEKGKYYRVSSNPEGKKCINKPLLK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK36PQNELIEEVRKRYWYRLDLGVRGAKYQVTSSPEGKKCANHPLPK37PQNEYQEEVRGRYDLYLVKGEKGKKYQKSSSGGGDKCILLKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKGENKKSINEKEKK39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ5643GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS7744PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH8645PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH9946RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH40FRNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH41RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH4244RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH4445FRNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH46RENQYDNDVTVSPQGRIHQI	04	29		
31GDLDVSYTFRERMIDVRYDYEEGAYYCKSSKGGGKKCAGKKEKK32PLNERNEGQRGRPGVRIVYYGKKYQVSSKGGGDKSALKVEKK33PDVNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK44PQNEYIEEHRKRYDLYLVYGERGKYYRVSSNPEGKKCINKPLLK55PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK6036PQNELIEEVRKRYWYRLDLGVRGAKYQVTSSPEGKKCANHPLPK7137PQNEYQEEVRGRYDLYLVKGEKGKKYQKSSSGGDKCILLKEKK7238PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSGGNKCILLKEKK739GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK749GLLEYIEEVRGRADLYLVLHERGAYYCVSSNPTGQKCINAVEKK7541PQNELIEQVFGRYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK7642PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKPKQEEEELQ7744PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH7845PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH7946RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH748RENQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH748RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH748RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH7548RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH7647RRNQYENDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH7748RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH7850FRNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH7	05	30	PLNNVQRLHVEERGHRLDYAN	GKSACRYNKGDNLDIDNLVLE
32PLNERNEGQRGRPGVRIVYYYGKKYQVSSKGGGDKSALKVEKK33PDVNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK34PQNEYIEEHRKRYDLYLVYGEKGKYYRVSSNPEGKKCINKPLLK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK36PQNELIEEVRKRYWYRLDLGVRGAKYQVTSSPEGKKCANHPLPK37PQNEYQEEVRGRYDLYLVKGEKGKKYQKSSSGGDKCILLKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTADLYLVLHERGAKYQVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKPKQEEEELQ44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVWSPQGRIHQIEYAMRPQQKAQPAQPADEVAEKADEPMEH40FRNQYDNDVTVWSPQGRIHQIEYAMRPQQKAQPAQPADEVAEKADEPMEH41SENQYDNDVTVWSPQGRIHQIEYAMRPQQKAQPAQPADEVAEKADEPMEH42FRNQYDNDVTVWSPQGRIHQIEYAMRPQQKAQPAQPADEVAEKADEPMEH44FORQYDNDVTVWSPQGRIHQIEYAMRPQQKAQPAQPADEVAEKADEPMEH45FRNQYDNDVTVWSPQGRIHQIEYAMRPQQKAQPAQPADEVAEKADEPMEH46FNQYDNDVTVWSPQGRIHQIEYAMRPQQKAQPAQPADEVAEKADEPMEH <tr< th=""><th>06</th><th>31</th><th>GDLDVSY IFRERMDVRY DYEE</th><th>GAY YCKSSKGGGKKCAGKKEKK</th></tr<>	06	31	GDLDVSY IFRERMDVRY DYEE	GAY YCKSSKGGGKKCAGKKEKK
33PDVNRIELGVLRDDVHLVYHEGKKLQVSNKHDGKKCALKKELK34PQNEYIEEHRKRYDLYLVYGEKGKYYRVSSNPEGKKCINKPLLK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK36PQNELIEEVRKRYWYRLDLGVRGAKYQVTSSPEGKKCANHPLPK37PQNEYQEEVRGRYDLYLVKGEKGKKYQKSSSGGGDKCILLKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS43GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEPAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH40FRNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH41RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH4249ERNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH4350FRNQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH44444444	07	32	PLNERNEGQRGRPGVRIVYYY	GKKYQVSSKGGGDKSALKVEKK
34PQNEYIEEHRKRYDLYLVYGEKGKYYRVSSNPEGKKCINKPLLK35PQNEYQEEVRGRTDLRLVYGERGVYYCVSNKPGGKKCAALKEKK36PQNELIEEVRKRYWYRLDLGVRGAKYQVTSSPEGKKCANHPLPK37PQNEYQEEVRGRYDLYLVKGEKGKKYQKSSSGGGDKCILLKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS43GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH40FRNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH41RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH4249ERNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH4350FRNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH44444444	00	33	PDVNRIELGVLRDDVHLVYHE	GKKLQVSNKHDGKKCALKKELK
35PQNEYQEEVRGRIDLRLVYGERGVYYCVSNKPGGKKCAALKEKK36PQNELIEEVRKRYWYRLDLGVRGAKYQVTSSPEGKKCANHPLPK37PQNEYQEEVRGRYDLYLVKGEKGKKYQKSSSGGGDKCILLKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS43GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEPAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH40FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH4148RENQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH40FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH4148RENQYDNDVTVSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH4249FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH4450FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH	00	34	PQNEYIEEHRKRYDLYLVYGEK	GKYYRVSSNPEGKKCINKPLLK
36PQNELIEEVRKRYWYRLDLGVRGAKYQVISSPEGKKCANHPLPK37PQNEYQEEVRGRYDLYLVKGEKGKKYQKSSSGGGDKCILLKEKK38PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS43GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH40FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH4148RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH49FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH40FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH4148RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH4249FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH4444FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH	09	35	PQNEYQEEVRGRIDLRLVYGER	GVYYCVSNKPGGKKCAALKEKK
1137PQNEYQEEVRGRYDLYLVKGEKGKKYQKSSSGGGDKCILLKEKK1238PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV1339GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK1440VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK1541PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ1642PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS1744PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH1845PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH1946RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH1047RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH14RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH15FRNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH16SRNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH1748RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH1820FRNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH1950FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH10FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH1148RENQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH1249FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH1350FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH	10	36	PQNELIEEVRKRYWYRLDLGVR	GAKYQVTSSPEGKKCANHPLPK
1238PLNEYIELVFKRADYRRDLHEKEAYYQVSSKSEGKKCILLKEKV1339GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK1440VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK1541PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ1642PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS1744PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH1845PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEVAEKADEPMEH1946RRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH2047RRNQYENDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH2148RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH2249ERNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH2350FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEVAEKADEPMEH2444FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH	11	37	PQNEYQEEVRGRYDLYLVKGEK	GKKYQKSSSGGGDKCILLKEKK
39GLLEYIEEVRGRADLYLVLHERGAKYQVSSKGENKKSINEKEKK40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS43GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEPAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRCAVPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIELAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH	12	38	PLNEYIELVFKRADYRRDLHEK	EAYYQVSSKSEGKKCILLKEKV
40VLNEYEEEVRGTYDYRRVLHPKGAYYCVSSNPTGQKCINAVEKK41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS43GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADEPAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRCAVPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIELAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQQKAQPAQPADEPAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH	13	39	GLLEYIEEVRGRADLYLVLHER	GAKYQVSSKGENKKSINEKEKK
41PQNELIEQVFGRYDYRLDYGEKQEEAPESELPPELKPKQEEEELQ42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS43GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS7744PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADGPAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRCAVPAQPADEVAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIELAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQQKAQPAQPTDEVAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH	14	40	VLNEYEEEVRGTYDYRRVLHPK	GAYYCVSSNPTGQKCINAVEKK
42PLNEYIELHRKRYDYRFVYWLKEVKEDEPELKREEIEKATKELDS43GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS744PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADGPAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRCAVPAQPADEPAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIELAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQQKAQPAQPADEPAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADEAEKADEPMEH2450FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH	15	41	PQNELIEQVFGRYDYRLDYGEK	QEEAPESELPPELKPKQEEEELQ
43GQNDYCELVRGSADFRRVLGVRYKEAEEYKLKYYLAPKHTEEIDS44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADGPAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRCAVPAQPADEPAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIELAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVVVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQQKAQPAQPTDEVAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH	16	42	PLNEYIELHRKRYDYRFVYWLK	EVKEDEPELKREEIEKATKELDS
44PPATDSQKSIISPVINHYKFIYSRPQQKAQPAQPADEVAEKADEPMEH45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADGPAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRCAVPAQPADEPAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIELAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVVVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQQKAQPAQPTDEVAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH	17	43	GQNDYCELVRGSADFRRVLGVR	YKEAEEYKLKYYLAPKHTEEIDS
45PERPDESETNPSLVLRASSDELTRPQRKAQPAQPADGPAEKADEPMEH46RRNQYDNDVTVWSPQGRIHQIEYAMRPQRCAVPAQPADEPAEKADEPMEH47RRNQYENDVTVWSPQGRIHQIELAMRPQRKAQPAQPADEVAEKADEPMEH48RENQYDNDVVVWSPQGRIHQIEYAMRPQRKAQPAQPADEPAEKADEPMEH49ERNQYDNDVTVESPQGRIHQIEYAMRPQQKAQPAQPTDEVAEKADEPMEH50FRNQYDNDVTVWSPQGRIHQIEYAMRPQRKAQPAQPADERAEKADEPMEH	10	44	PPATDSQKSIISPVINHYKFIYS	RPQQKAQPAQPADEVAEKADEPMEH
 46 RRNQYDNDVTVWSPQGRIHQIEYAM RPQRCAVPAQPADEPAEKADEPMEH 47 RRNQYENDVTVWSPQGRIHQIELAM RPQRKAQPAQPADEVAEKADEPMEH 48 RENQYDNDVVVWSPQGRIHQIEYAM RPQRKAQPAQPADEPAEKADEPMEH 49 ERNQYDNDVTVESPQGRIHQIEYAM RPQQKAQPAQPTDEVAEKADEPMEH 50 FRNQYDNDVTVWSPQGRIHQIEYAM RPQRKAQPAQPADERAEKADEPMEH 	18	45	PERPDESETNPSLVLRASSDELT	RPQRKAQPAQPADGPAEKADEPMEH
47 RRNQYENDVTVWSPQGRIHQIELAM RPQRKAQPAQPADEVAEKADEPMEH 48 RENQYDNDVVVWSPQGRIHQIEYAM RPQRKAQPAQPADEPAEKADEPMEH 49 ERNQYDNDVTVESPQGRIHQIRYAM RPQQKAQPAQPTDEVAEKADEPMEH 50 FRNQYDNDVTVWSPQGRIHQIEYAM RPQRKAQPAQPADERAEKADEPMEH	19	46	RRNQYDNDVTVWSPQGRIHQIEYAM	RPQRCAVPAQPADEPAEKADEPMEH
48 RENQYDNDVVVWSPQGRIHQIEYAM RPQRKAQPAQPADEPAEKADEPMEH 49 ERNQYDNDVTVESPQGRIHQIRYAM RPQQKAQPAQPTDEVAEKADEPMEH 50 FRNQYDNDVTVWSPQGRIHQIEYAM RPQRKAQPAQPADERAEKADEPMEH 24 44 44	20	47	RRNQYENDVTVWSPQGRIHQIELAM	RPQRKAQPAQPADEVAEKADEPMEH
49 ERNQYDNDVTVESPQGRIHQIRYAM RPQQKAQPAQPTDEVAEKADEPMEH 50 FRNQYDNDVTVWSPQGRIHQIEYAM RPQRKAQPAQPADERAEKADEPMEH 24	21	48	RENQYDNDVVVWSPQGRIHQIEYAM	RPQRKAQPAQPADEPAEKADEPMEH
50 FRNQYDNDVTVWSPQGRIHQIEYAM RPQRKAQPAQPADERAEKADEPME	22	49	ERNQYDNDVTVESPQGRIHQIRYAM	RPQQKAQPAQPTDEVAEKADEPMEH
	23	50	FRNQYDNDVTVWSPQGRIHQIEYAM	RPQRKAQPAQPADERAEKADEPMEH
	24			

Index	N-terminus	C-terminus
1	VELDEGADA	DKDEDEGEN
2	FELDITOATA	
3		ISLKQSWEVI
4	SSASDAMAKHE	ISUKQAILVII
5	ASASGAMAKHE	IHEKOAPEVH
6	SIETPTROIRIT	ISI CESPI VH
7	VSWIISVI VVI JEG	ISUKOAVGVH
8	VSWIISPLV VLI IFS	ISLOOADEVH
Q	VSWIISRLVLLEG	ISOKOAPEVH
10	GEENEKITI CGTVCI	I KDVI EGDEKEE
11	GNINEKKTTIGEVCV	I KDVEEGDI KEE
12	GSINEKKTTCGTVCI	TPPKKKTSGGTDSA
12	SPENRKSTTCGTVCI	TPPKKKTSGGGDSA
13	DSWEDI VDOVI GATKNE	ISKCI VRCHKI FHV
14		
15	DDCEDCDVOAKDAESAE	IFICELIOSSKEENI IETCCIVDSAVIEHC
10		IFICELI KSAKLEHC
17		IFICCLI KSLKLEHI IETCALADSCVI EHV
10		IFICALARSORLERI IETACI VDSI VVEHV
19	DDREHT V TQASLAFNILE	
20		
21		GEEQNLIALQDYLDDNQ
22	A SCADI EL WEANVCAAER	
23	ASURFLELWEAN VUAAEK	AEEQINKEAIQDVEDENQ
24	ASCREELWEAV VUAAER	IPRPRSNISDGQKLKGKI
23		LPRPRENASDFQSLKSTT
20		LPRPKLNISDFQELKPKA
27		LPRPKLNISDFQKLKLV I
28	ESGDPRELWEAVVGAAER	LPRPKLNISDFQLLKKKE
29	ESGRPPELWEAKVGAAEN	LPRPKLNKSDFQSCKPKI
30	SSEENCRLVVLVFGMCCPA	IPRPKLNASDFQSLKKGY
31	SSEMILKLV VLV IGMSCPA	LPRPRICISDFQKFKEKV
32		IPRPKLNI SDFQKLKKKG
33	HLRGVGAGAIAKKKLIEAKY	IPRPKLN I SEFQELY MKE
34		LFIDDY SQEITAEH Y REALK
35		LFIDLY SQKIIAEEARELLK
30		LFIDLVSQAIIAEEARDDLP
37		LFIDLY SQEITAEEPREAAP
38	HSRCVGGGAIADKALAEDKY	LFIDLY SQEITAEEARELLN
39		LFIDLY SNCIIFEEY REDLP
40	HEREVGLGAIAKKKLAEAKY	LFIDLY SQEIGDEEY REALP
41	HFREVGAGAIAKKKLAELKY	LFYDLYSQEITKEEPREALK
42	ALAKLLLESNIRLW VNRPSIIIT	LFIDLY SQEITKEEPREALK
43	ALPRNLLKSSIVPWVNISSVIQT	LFIDLY SQPITDEEPREALK
44	ALEEALLFSSLVSWNVYPIVIQK	NKPMAKDKEGFTMYKYILQHKIQ
45	AVDKNGLNGNIREVNVIPIIIIT	NLLMPTDLAKIGPHWRSLDTSSS
46	SGPKDMLELGGVIWNNRSQNLYS	EAIMLISIDEGNEFKAELNGKTV
47	AGPKNLLELGIVLVVRLYKFILS	MSGAPDTLGQGGGGGGGGGGGGGG
48	ALNEGLLEGLGQLVVQTVSNIYK	MSGAPDTLGQGGGGGGGGGGGGG
49	AAASAGATRALLLLLMAVAAPSRA	MSGAPDTGSQGGGGGGGGGGGGGGGGGGGG
50	AAASAGATRWLLLLLMAVAAPSRA	MSGAPDTLGQGGGGGGGGGGGGGGG

