In-silico detection and correction of cryptic splice sites created by codon optimisation of an STXBP1 transgene for gene therapy

Gayatri Dhir, University College London

Supervised by Dr. Arianna Fozzato, Wellcome Sanger Institute

Abstract

STXBP1-related developmental and epileptic encephalopathy (STXBP1-DEE) is an early-onset paediatric neurological disorder caused by haploinsufficiency of the STXBP1 gene, which encodes the presynaptic protein Munc18-1. There is currently no disease modifying therapy; however, those in preclinical development aim to supplement the deficiency by delivering a working copy of the gene. A key step is codon optimisation, where synonymous codons are substituted to improve the transgene and by extension, protein expression. However, altering the DNA sequence may unintentionally create cryptic splice sites, which can be recognised by the spliceosome, causing aberrant splicing that reduces or abolishes the intended protein. This project aimed to detect and correct cryptic splice site creation in a potential STXBP1 therapeutic transgene. Four codon optimisation tools generated five coding sequences that were compared with the wild-type sequence using MaxEntScan and SpliceAI, followed by consequence modelling and subsequent sequence correction to fix any splicing risks while preserving the protein. Although all variants encoded the same 594-amino-acid protein, they differed from wild-type at up to 21% of nucleotide positions. MaxEntScan identified ten newly created candidate splice sites, with each optimisation strategy introducing at least one, and one high risk site flagged by both tools. Consequence modelling predicted seven frameshift events and two in frame deletions; notably, the predicted splice site risk was associated with local sequence change rather than overall GC content, implying these sites cannot be anticipated from composition alone. All ten candidate sites were eliminated through synonymous nucleotide substitutions to preserve the encoded protein. These findings indicate that codon optimisation can introduce potentially consequential splice site risks for STXBP1 transgenes and that sequence-based screening is able to identify them. Experimental validation of the highest risk site, as well as expansion to additional therapeutic transgenes will determine how broadly this approach applies in future gene-therapy design.

Table of Contents

Introduction

Monogenic paediatric neurological disorders are among the most severe neurological disorders of early childhood. A subset of these disorders present as developmental and epileptic encephalopathies (DEE)[1,2] Because they arise from a fault in a single copy of a gene, they represent a tractable and tangible target for genetic medicine and novel therapies. Syntaxin binding protein 1-related disorder, or STXBP1 -related developmental epileptic encephalopathy (DEE) is one such prevalent condition that has become recognised as one of the most common monogenic causes of early-onset epileptic encephalopathies[1,2,6].

STXBP1, located on chromosome 9q34.11[2], encodes Munc18-1, an essential family protein involved in regulating synaptic vesicle priming and broadly vesicular fusion at the presynaptic bouton neuron. Munc18 -1 binds syntaxin-1[3,4,36] in its closed conformation and with Munc13, controls the assembly of the SNARE (Soluble N-ethylmaleimide-sensitive factor Attachment protein REceptor) complex that drives vesicle fusion. STXBP1 is primarily expressed in neurons[5]. Some genetic SNAREopathies, or disorders relating to the correct conformation or function of the SNARE complex involved in presynaptic vesicular priming and release are related to misformation of the Munc18-1 protein[7].

More specifically, the consequences of STXBP1 deficiency are circuit specific. Munc18-1 is essential for neurotransmitter release; models of STXBP1 haploinsufficient mice indicate that cortical GABAergic inhibition was impaired by dysfunctional parvalbumin and somatostatin expressing interneurons (both fast spiking and non spiking interneurons respectively), thus suggesting that STXBP1 dysfunction can tip the excitatory-inhibitory balance of circuits in the brain towards hyperexcitability that manifests as epilepsy[8]. Both are inhibitory (GABAergic) neurons whose normal role is to restrain the excitatory pyramidal neurons around them, but they do so at different sites. Fast-spiking parvalbumin interneurons control the timing and rate at which a pyramidal neuron fires[9], whereas somatostatin interneurons target distal dendrites[9]. When Munc18-1 loss weakens GABA release from these interneurons, both forms of

inhibition are diminished and pyramidal neurons are disinhibited, shifting the excitatory inhibitory balance in the brain. The identification of a clear genetic deficiency that causes defined circuit failures in the brain is what makes STXBP1 a compelling candidate for therapies that supply a working copy of the gene and thus returning the protein to the cells that lack it.

Furthermore, STXBP1-DEE is one of the more frequently identified monogenic DEEs, accounting for approximately 10% of early onset epileptic cohorts in a study (n=150)[10], with an estimated incidence rate of at least 1 in 30,000 births[1,11]. It is further recognised as one of the top five causative genes across genetic screens for infantile spasms, DEEs, and presentations of global developmental delay[12,14]. Furthermore, patients which are initially diagnosed with other disorders such as Early Infantile Epileptic Encephalopathy (EIEE), Atypical Rett Syndrome, and Dravet syndrome also comprise a population of pathogenic STXBP1 variants, which also indicates that the incident rate could be much higher given that genetic testing is not performed universally and could be attributed to alternative disorders.

Clinically, the disorder is pleiotropic as well as severe. The phenotypic spectrum is wide, and the severity of the disease and underlying symptoms is variable depending on the mutation to the STXBP1 gene and underlying protein formation. In the largest cohort assembled to date in 2022 (n=534)[1], intellectual disability was present in approximately 90% of patients over 11, of whom 64% were severely affected and 38% of that group were nonverbal. Seizures affect the great majority of patients and is the primary symptom, with onset in the first year of life in almost 88% of patients[1] and focal-onset seizures the most common type. Furthermore, in a screen of 663 patients with DEE, disorders related to the STXBP1 gene were among the genes found to have an association with infantile spasming[14]. Beyond seizure, patients can also express a variety of movement-associated disorders such as ataxia, Parkinsonian like tremors, or tetraplegia[2].

Crucially for this disease, genetic screens have pointed to splicing errors as a potential disease mechanism. In the same large cohort (n=534), splice-site variants were the third most common disease causing genotypic change (n=79)[1]. Overall, this establishes that disrupting splicing is a proven route to STXBP1 disorders in patients, which is crucial because if disrupting splicing causes disease in the natural gene, a splice signal accidentally created in a therapeutic copy could also result in errors, producing a similar truncated Munc18-1 protein. It is therefore extremely critical to determine whether any engineered sequence intended to treat this disorder are themselves free of such signals.

All treatments currently available for STXBP1-related disorders are symptomatic medications such as prednisone, clobazam, and levetiracetam rather than disease-modifying[2]; essentially, they aim to control the main symptom — seizures — but leave the underlying genetic protein deficit untouched. Yet, even levetiracetam, one of the most prescribed antiseizure medications for STXBP1 patients, was not associated with improved seizure freedom in the largest cohort studied (n=534)[1,15]. Critically, even when seizures are controlled for STXBP1-DEE patients, developmental, motor, and communication impairments persist. Repurposed agents under investigation, including the chemical chaperone agent and the antiseizure drug fenfluramine are in clinical stages[2], but similarly fail to address the underlying protein deficit rather than the cause. It is clear that the shared and key limitation of these approaches is that even when seizures are controlled, motor and communication deficiencies in patients still persist because the root cause of insufficient Munc18-1 has not been addressed. Given that the disorder arises from haploinsufficiency (it is caused by too little of Munc18-1 rather than the addition of a harmful mutant protein)[1,2,6], restoring the missing Munc18-1 protein should therefore correct the root deficiency in patients. Thus, this gap points clearly towards gene therapies that are able to deliver a working copy of the gene to restore the missing protein.

Gene therapies are already underway, a phase 1/2a clinical trial that delivered a functional copy of the STXBP1 gene to patients was paused in 2025 following the death of the first participant[16,17]. Currently,

several other gene therapies for the STXBP1 gene are in preclinical development[17]. It is clear from the aforementioned example and that when it comes to genetic therapies and bioengineering, the method the therapeutic transgene can be sequence engineered becomes important.

Before manufacture and testing, a therapeutic coding sequence can be codon-optimised: synonymous codons are substituted to keep the encoded protein unchanged, while the DNA sequence is altered to improve expression, reduce immunogenic responses, and enhance translation. Optimised sequences can substantially increase the amount of protein produced per delivered gene copy,[18] so that the therapeutic level of expression can be reached at a lower dose — something that is critical when dealing with viral vectors such as AAV, where too much of a vector dose can cause dose-related toxicity and serious adverse events[19]. Thus, an optimised sequence that achieves sufficient expression for a therapeutic effect improves the safety margin for patients. Optimisation is also utilised to deplete CpG dinucleotides, which are recognised by the patient's innate immune system and can drive inflammatory responses against the therapeutic[20]. Furthermore, optimisation is used to reduce the similar sequence identity between the transgene and the patients own copy of the gene[18]. Because the therapeutic transgene and the native gene encode the same protein, their DNA sequences would otherwise be highly similar and at risk to undergo homologous recombination[18] — which could disrupt or alter either the transgene or the native locus. For these reasons, codon optimisation of genetic payloads is a key step in bringing gene therapies towards patients.

The same process that makes these payloads/transgenes can also introduce a hidden implication. Codon optimisation is known to have unintended consequences, including altered protein folding/function or changed post-translational modification in the cell. Because codon optimisation alters the nucleotide sequence, it can inadvertently create short motifs that resemble splice donor or acceptor sites[18]. As such, when the transgene is transcribed in the nucleus, any such cryptic splice site would be read by the spliceosome which could excise parts of the mRNA sequence and thus produce a non-functional protein -

the very outcome that the therapy is designed to prevent. As noted above, splice site variants are an established cause of STXBP1-DEE, and computational tools such as SpliceAI have indicated that disrupting splicing in this gene which occurs naturally with the disease can produce frameshift or deletion effects[23]. However, all publicly available designs of the STXBP1 payload have used the native sequence, not optimised sequences[16,17].

Taken together, the issues of STXBP1-related disorder, its prevalence and symptoms, the lack of a clear therapeutic solution, and the importance of codon optimisation, it is clear that splice-site screening should be considered. Splicing disruption is a proven cause of STXBP1 disease in patients, yet the tools used to optimise the therapeutic gene were never designed to consider splicing at all. These tools utilize various algorithms and produce different sequences, but whether they differ in their tendency to create cryptic splice site motifs has received comparatively little attention. In a disorder where the delivered protein must be full length to work, the possibility that the lack of optimisation step could lead to the transgene being compromised is a workflow and pipeline that should be developed and used. It is this gap that the present work addresses.

This project therefore sets out to create a computational workflow, identify, and correct cryptic splicing signals in codon-optimised STXBP1 therapeutic transgenes Rather than examining naturally occurring patient mutations (since gene supplementation for haploinsufficient disorders deliver a full copy regardless of the underlying variant), it follows a process of generating several codon optimised versions of the coding sequence, scanning for cryptic splice sites with two complementary methods (MaxEntScan and SpliceAI), and identifying and fixing the protein-level consequences of the most significant sites identified by the algorithm.

Methods

The *Homo sapiens* wild-type STXBP1 coding sequence (1785 nt, 594 amino acids) was obtained from the MANE select transcript ENST00000373302[24], and the protein and its domains were cross-referenced using UniProt entry P61764[25]. Five codon optimised versions of the coding sequence were generated utilizing four independent, publicly available tools: JCat[26], OPTIMIZER[27], GenScript Gen Smart[28], and the IDT Codon optimisation Tool[29] (two settings were utilised, IDT_legacy and IDT_pilot).These tools apply different algorithms to select synonymous codons: JCat maximises the codon-adaptation index by always choosing the most frequent codon expressed in the organism of interest[26], OPTIMIZER can instead match the overall codon-usage distribution of the organism of interest[27], and commercial tools such as GenSmart[28] along with IDT[29] utilizes a multi-pronged method involving a balance between GC content and the codon-adaptation index. The same canonical wild-type sequence was used as input for each tool, so all differences between variants come only from the respective optimisation algorithm . Every variant was confirmed to be 1785 nt long and to translate to the exact wild-type protein using Biopython.

Sequences were then aligned with MAFFT to confirm codon level correspondence[30]. For each sequence, GC content, GC content at the third codon position, the number of CpG dinucleotides, and percent identity to wild-type were calculated with Biopython[31]. Four properties were measured for each sequence to describe how far optimisation had moved the CDS from the natural gene. Identity to wild-type is the fraction of unchanged nucleotides; because the protein is fixed, a lower percentage means a tool rewrote more of the sequence. GC content is the percentage of G or C bases, which tools tend to raise because GC-rich messages are often better expressed. GC3 is the GC content at the third codon position — the base most synonymous swaps change, therefore it serves as the most direct measure of how hard a tool pushed toward its preferred codons. CpG count is the number of CG dinucleotides, which some tools deliberately reduce to avoid immune sensing and silencing. This is due to unmethylated CpG motifs in

the transgene having a tendency to be recognised by Toll-like receptor 9 in the innate immune system[20], which can provoke an inflammatory response and long-term silencing of the transgene. Depleting CpG therefore helps the transgene of interest persist and stay expressed in the body.

Two complementary tools were used to detect splice signals. Firstly, SpliceAI, a 32-layer neural network that uses dilated convolutions to analyse up to 10,000 nucleotides of flanking sequences around each position[32]. It was trained using GENCODE-annotated pre-mRNA transcripts to distinguish splice donor and acceptor sites[32]. SpliceAI was run in custom-sequence mode where each coding sequence was padded with flanking N (unknown) bases to meet the model's required 10,000 nt context window3[32]. Each optimised sequence was scored with its five-model ensemble, and gave a donor and acceptor probability between 0 and 1 at every position - where a higher probability means that the model is more confident that a position reflects a real splice site. To isolate the effect of optimisation, the wild-type probability was subtracted from each variant's probability at the same position to give a delta (Δ) score: a positive Δ marks a splice signal that optimisation created, and a negative Δ value reflects one that was removed. Positions with $\Delta \geq 0.2$ were flagged and $\Delta \geq 0.5$ treated as high confidence, reflecting the recommended thresholds from prior literature.

The second tool, MaxEntScan uses a non-machine learning approach. It uses the maximum entropy principle to model splice sites by creating the least-biased probability distribution that not only matches the base frequencies, but also certain dependencies and non-adjacent relationship between bases[33]. MaxEntScan scores short sequences against a statistical model of real human splice sites; where the window for donors and acceptors are 9 and 23 nucleotide sequences respectively[33]. MaxEntScan reports a score in bits, where genuine human splice sites tend to score around 8-9[33]. A higher bit score therefore means the window reflects an authentic splice site[33]. Every possible window across each sequence was scored. A 3.0 bit baseline was used to define a recognisable motif and an 8.0 bit threshold to mark a strong, usable one[34]. A site was created when its score rose above the 3.0 bit floor relative to wild-type,

and strengthened when an existing motif increased substantially. These bit score conventions were adapted from published variant criteria using MaxEntScan[34].

The two tools were utilised together, where SpliceAI reads the broader sequence context and is the more accurate of the two given its training on real annotated models of human splice sites, but it is harder to determine exactly why a site was flagged as it utilizes a more complex model — a deep neural network of many layers and millions of learned parameters rather than from a single clear rule or motif[35]. Conversely, MaxEntscan looks only at a shorter window, making it more sensitive but prone to producing false positives. Thus, a site flagged by both tools is therefore a stronger candidate than one flagged by only one.

Because the transgene is an intronless sequence, no splicing is intended and cryptic donors and acceptors are not typically paired as they would be across a natural intron-exon sequence of mRNA. To test whether a created site could still drive unintended splicing, each was searched for a compatible partner elsewhere in the sequences (a downstream acceptor for a created donor or an upstream donor for a created acceptor). Where a partner existed, the intervening sequence was removed in Python to simulate the splice The length of the removed piece was checked to identify if the reading frame was preserved or if a frameshift and resulting stop codon appeared soon after. The resulting transcript was then translated to locate where the protein would end, and the lost region was mapped on the domains of Munc18-1 (Domain 1, Domain 2, Domain 3a, and Domain 3b) sourced from UniProt[25].

The created sites that then led to protein loss were repaired by replacing the codon responsible for the splice motif with a synonym codon. Each change had to satisfy three conditions: firstly, drop the target MaxEntScan score below the 3.0 lower threshold; secondly, leave the encoded protein completely unchanged; and, lastly create no new splice site. Every rescued sequence was then re-scored with both tools and a whole sequence re-scan confirmed that no new site has appeared anywhere and that none of the wild-type motifs previously removed by optimisation had been accidentally reintroduced.

All analyses were performed in Python 3.10 using Biopython[31], pandas, the maxentpy wrapper, MaxEntScan[33] and the SpliceAI package[32].

Results

Table 1

Sequence composition of wild-type and codon optimised STXBP1 variants

Sequence	GC (%)	GC3 (%)	CpG	Identity to WT (%)
wildtype	52.9	68.6	68	100.0
OPTIMIZER	63.8	99.8	124	83.1
JCat	63.8	99.8	133	83.1
IDT_legacy	50.3	63.2	88	78.6
IDT_pilot	55.1	76.8	31	82.1
GenSmart	54.7	76.0	79	80.4

All five variants preserved the 594-amino-acid protein exactly but differed from the wild-type sequence at 17-21% of nucleotides, confirming that each tool rewrote roughly one in five bases. The tools diverged most in how far they pushed composition. OPTIMIZER and JCat increased GC content up to 63.8% and GC3 to 99.8%, meaning almost every third-codon position was loaded with G or C, which is generally a key indicator of aggressive optimisation. IDT_legacy did the opposite, sitting below wild-type at 50.3%

GC and 63.2% GC3. The tools also split on CpG strategy: IDT_pilot depleted CpGs to 31 (well below the wild-type 68), while JCat nearly doubled them to 133.

Table 2.a

Cryptic splice sites flagged above thresholds (MaxEntScan)

Note: *Positions are numbered by the MaxEntScan motif start (for example 1553); the corresponding GT or AG dinucleotide lies one base downstream (1554), so the site described as 1553 in Table 2 and 1554 subsequent tables are the same site.*

Tool	**Site type**	**Position (nt)**	**Status**
GenSmart	Acceptor	322	Created
IDT_legacy	Acceptor	322	Created
IDT_legacy	Donor	955	Created
IDT_legacy	Donor	1553	Created
IDT_legacy	Donor	569	Strengthened
IDT_legacy	Acceptor	987	Strengthened
IDT_legacy	Donor	1681	Strengthened
IDT_pilot	Acceptor	326	Created
JCat	Acceptor	326	Created
JCat	Donor	894	Created

OPTIMIZER	Acceptor	326	Created
OPTIMIZER	Donor	894	Created
OPTIMIZER	Donor	1553	Created
OPTIMIZER	Donor	569	Strengthened

Table 2.b

Cryptic splice sites flagged above thresholds (SpliceAI)

Tool	Site type	Position (nt)	Δ score
IDT_legacy	Donor	1553	0.30

MaxEntScan identified 10 splice-site motifs newly created across the five variants, plus 4 pre-existing motifs that were strengthened (Table 2). A created site is one that is essentially absent in the wild-type sequence and scores below the detection threshold, but rises after optimisation, whereas a strengthened site is one that already resembled a weak splice motif in the wild-type sequence and was pushed to a higher, more splice-site-like score by codon optimisation. The created sites were spread unevenly between tools: IDT_legacy and OPTIMIZER produced 3 each, JCat 2, and GenSmart and IDT_pilot 1 each. SpliceAI, applied independently at $\Delta \geq 0.2$, flagged only a single created site, a donor at nucleotide 1554 in IDT_legacy ($\Delta \approx 0.30$; probability rising from 0.0004 to 0.298). This same site was also the strongest MaxEntScan motif in the entire dataset (10.88 bits), making it the one site both methods agreed on. Notably, the variant with the lowest GC content (IDT_legacy, 50.3%) created the strongest motif overall, showing that cryptic-site creation tracked specific local sequences rather than global GC content. GC content therefore cannot be used as a shortcut for splice safety (a single-gene observation).

Table 3

Predicted protein-level consequences of the 10 cryptic splice sites created by codon optimisation.

Tool	**Cryptic site**	**Partner site**	**Bases removed**	**Effect**	**Predicted protein outcome**	**Munc18-1 region affected**
IDT_legacy	Donor 1554	Acceptor 1728 (9.43 bits)	175	Frameshift	Out-of-frame C-terminus from residue 518; no stop codon	Domain 3b
IDT_legacy	Donor 956	Acceptor 1728 (9.43 bits)	773	Frameshift	Stop at residue 329; ~328-aa protein	Domains 3a + 3b
OPTIMIZER	Donor 1554	Acceptor 1728 (7.18 bits)	175	Frameshift	Stop at residue 524; ~523-aa protein	Domain 3b
GenSmart	Acceptor 322	Donor 64 (4.51 bits)	259	Frameshift	Stop at residue 44; ~43-aa protein	All domains
IDT_pilot	Acceptor 326	Donor 64 (5.73 bits)	263	Frameshift	Stop at residue 29; ~28-aa protein	All domains

JCat	Acceptor 326	Donor 64 (5.73 bits)	263	Frameshift	Out-of-frame from residue 22; no stop codon	All domains
OPTIMIZER	Acceptor 326	Donor 64 (5.73 bits)	263	Frameshift	Out-of-frame from residue 22; no stop codon	All domains
JCat	Donor 895	Acceptor 1728 (7.18 bits)	834	In-frame	~278-aa deletion (residues 299–576)	Domains 3a + 3b
OPTIMIZER	Donor 895	Acceptor 1728 (7.18 bits)	834	In-frame	~278-aa deletion (residues 299–576)	Domains 3a + 3b
IDT_legacy	Acceptor 321	No compatible donor	—	No event	Cannot splice; no sequence removed	None

Modelling of the 10 created sites predicted that 7 would cause a frameshift, 2 an inframe deletion, and one could not occur because it had no partner. Munc18-1 folds into a single arch-shaped unit whose domains act to clasp syntaxin-1 in its closed conformation. Thus, any frameshift or shortening of the protein that is most likely expected to prevent correct folding and binding of syntaxin-1 rather than a protein that is partially functional[3,4,36].

The highest-confidence event was the IDT_legacy donor at nucleotide 1553, which was paired with a strong downstream acceptor and shifted the reading frame and led to an aberrant the C-terminus from residue 518 onwards. A second IDT_legacy donor at nucleotide 956 removes 773 nt, which results in a premature stop codon at residue 329 and deletes domains 3a and 3b of the protein. The two in-frame events, (a donor at nucleotide 895 in both JCat and OPTIMIZER) would delete approximately 278 residues, removing most of the protein's C-terminal half.

Table 4

Rescue of the 10 created cryptic splice sites by synonymous codon substitution

Variant	**Site**	**Before (bits)**	**After (bits)**	**Codon swap**	**Number of nt changed**
IDT_legacy	donor 1554	10.88	2.69	AGG→AGA	1
OPTIMIZER	donor 1554	9.77	1.27	CGG→CGT	1
IDT_legacy	donor 956	8.72	0.54	GGT→GGA	1
GenSmart	acceptor 322	8.58	0.62	ACA→ACG	1
IDT_pilot	acceptor 326	9.67	−6.76	AGC→TCC	2
JCat / OPTIMIZER	acceptor 326	9.03	−7.40	AGC→TCC	2

JCat / OPTIMIZER	donor 895	8.70	−2.65	GTG→GTC; GAG→GAA	2
IDT_legacy	acceptor 321	9.32	0.96	ACA→ACT	1

Every one of the 10 created sites was eliminated by synonymous codon substitution, each falling below the 3.0 bit floor while leaving the protein identical to wild-type. Five of the ten needed only a singular nucleotide change. The most flagged site, IDT_legacy at nucleotide 1554, was silenced by one synonymous change in codon 518 (AGG→ AGA) which cut the MaxEntScan score from 10.88 to 2.69 bits and the SpliceAI probability from 0.298 to 0.002, essentially the wild-type baseline of 0.0006.

Discussion

This project aimed to investigate whether codon optimisation of an STXBP1 therapeutic transgene can introduce cryptic splice sites and whether such sites can be detected and subsequently corrected *in silico.* Each one of the five optimised variants introduced at least one new candidate splice site, with MaxEntScan identifying 10 motifs in total and SpliceAI identifying one. Modelling of the protein predicted that 7 out of the 10 sites would shift the reading frame and 2 would cause a large in-frame deletion, in each case removing certain domains of Munc18-1 that are predictably required for its function. All 10 sites were then eliminated by synonymous codon substitution without altering the encoded protein, while five of these sites required only a single nucleotide change and all ten cleared with at most two. The central finding of this work is twofold: codon optimisation with the aforementioned tools can potentially create cryptic splice-site motifs for the STXBP1 transgene, but these risks are detectable and correctable at the sequence design stage.

The above results support the idea that the optimisation step, which is intended only to improve expression for engineered transgenes and therapeutics[18], can have an unintended effect on splicing. Because all five variants were generated from the same wild-type sequence and encode the same protein, the differences between them arise from the choice of codons the tool utilised, attributing the creation of cryptic sites to the optimisation algorithm. The number of new splice sites varied between all five tools (from one to three), indicating that the tendency of splice signals depends on the specific algorithm and its settings. This potential effect on splicing adds to other documented unintended effects of codon optimisation, such as changes in protein folding, post-translational modification, and co-translational processing.

A major result was the only site that was flagged by both SpliceAI and MaxEntScan, the IDT_legacy donor site at nucleotide 1553. MaxEntScan gave it the highest score (10.88 bits), while SpliceAI increased its donor probability from 0.0004 in the wild-type sequence to 0.298. Given that both models identified it as a potential splice-site, it serves as a stronger candidate for a real cryptic splice site and makes it the most critical site for experimental follow up.

An unexpected result found that cryptic-site creation did not track overall GC content. The variant with the lowest GC content produced both the strongest motif and the most created sites, while the one of the two highest-GC variants (OPTIMIZER) created just as many sites as it did, the opposite of what would be expected if aggressive, high GC optimisation would drive splice-site risk. This suggests that splice risk is not simply caused by high-GC optimisation. Instead, splice sites depend on the exact bases at specific positions, meaning that global measures such as GC or GC3 content are poor indicators of splice safety and direct motif-level screening is needed. The optimisers also removed many more existing wild-type motifs than they created, suggesting that optimisation was more often protective than harmful.

If these predictions hold in both *in-vitro* and *in-vivo* experimental conditions for a variety of transgenes, the implications involve a potential wider use of computational splice-site screening prior to experimental work. Most created sites were eliminated by changing a single nucleotide, and every fix left the protein unchanged and created no new sites nearby. Because the method depends only on the sequence and is computational, it could be built into transgene engineering if the result holds for other genes.

However, these results have several limitations. First, every result is a computational prediction and none has been confirmed experimentally. The analysis identifies candidate sites and models their likely effect, yet only *in-vitro* and *in-vivo* work can indicate whether any of these sites are targeted by the spliceosome in cells itself, and at what efficiency. The analysis also uses score thresholds that were developed for other purposes; the MaxEntScan thresholds were adapted from studies of splice-site loss[34], while the SpliceAI thresholds were developed for genomic variants rather than intronless coding sequences[32]. Results could therefore change based on thresholds utilised. Furthermore, SpliceAI was applied in a new context and required the intronless coding sequence to be padded with unknown bases to fit the model, so its scores in this setting should be read critically as relative, not absolute, splice probabilities . Lastly, both tools mainly examine the core splice site sequence and do not incorporate the wider regulatory signals that control splicing in the cell. A high-scoring site may therefore never be identified and spliced by the spliceosome, while a rescued site may not be completely inactivated. Overall, significant experimental testing is needed to confirm that the changes found computationally lead to aberrant splicing, and that the rescue prevents splicing as well.

Future research should take the next step and experimentally test the strongest predicted site for STXBP1 by expressing the optimised and rescued sequence in a neuronal cell type, and examining the mRNA and translated protein for abnormal splicing or missing protein domains. The workflow should then be tested on more optimised transgenes and other therapeutic genes to investigate whether different optimisation tools consistently create cryptic-splice site risks, and whether these risks can be corrected. The rescued

sequences should also be checked against the original optimisation goals to make sure that fixing splice sites does not reduce the intended expression benefits. If these findings are confirmed, automated splice-site screening could be added to standard transgene design as a simple step before experimental work. Overall, this study suggests that codon optimisation can create potentially harmful splice sites in an STXBP1 transgene, but these sites can be identified and corrected computationally before experimental testing — a vital step in developing tangible therapeutics for paediatric patients.

References

1. Xian, J. *et al.* Assessing the landscape of STXBP1-related disorders in 534 individuals. *Brain* **145**, 1668–1683 (2021).
2. Freibauer, A., Wohlleben, M. & Boelman, C. STXBP1-Related Disorders: Clinical Presentation, Molecular Function, Treatment, and Future Directions. *Genes* **14**, 2179 (2023).
3. Burkhardt, P., Hattendorf, D. A., Weis, W. I. & Fasshauer, D. Munc18a controls SNARE assembly through its interaction with the syntaxin N-peptide. *EMBO J* **27**, 923–33 (2008).
4. Rizo, J. & Xu, J. The Synaptic Vesicle Release Machinery. *Annual Review of Biophysics* **44**, 339–367 (2015).
5. STXBP1 syntaxin binding protein 1. *proteinatlas.org* (2026).
6. Saitsu, H. *et al.* De novo mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy. *Nature Genetics* **40**, 782–788 (2008).
7. Vinci, M. *et al.* STXBP6 Gene Mutation: A New Form of SNAREopathy Leads to Developmental Epileptic Encephalopathy. *International Journal of Molecular Sciences* **24**, 16436 (2023).
8. Chen, W. *et al.* Stxbp1/Munc18-1 haploinsufficiency impairs inhibition and mediates key neurological features of STXBP1 encephalopathy. *eLife* **9**, (2020).
9. Jang, H. J. *et al.* Distinct roles of parvalbumin and somatostatin interneurons in gating the synchronization of spike times in the neocortex. *Sci Adv* **6**, eaay5333 (2020).
10. Na, J.-H. *et al.* Targeted gene panel sequencing in early infantile onset developmental and epileptic encephalopathy. *Brain and Development* **42**, 438–448 (2020).

11. López-Rivera, J. A. *et al.* A catalogue of new incidence estimates of monogenic neurodevelopmental disorders caused by de novo variants. *Brain* **143**, 1099–1105 (2020).

12. Symonds, J. D. & McTague, A. Epilepsy and developmental disorders: Next generation sequencing in the clinic. *Epilepsy & Neurodevelopmental Disorders* **24**, 15–23 (2020).

13. Specchio, N. *et al.* International League Against Epilepsy classification and definition of epilepsy syndromes with onset in childhood: Position paper by the ILAE Task Force on Nosology and Definitions. *Epilepsia* **63**, (2022).

14. Ganesan, S. *et al.* A longitudinal footprint of genetic epilepsies using automated electronic medical record interpretation. *Genetics in Medicine* **22**, 2060–2070 (2020).

15. Wang, Q.-H. *et al.* Efficacy of levetiracetam in STXBP1 encephalopathy with different phenotypic and genetic spectra. *Seizure* **95**, 64–74 (2022).

16. A Clinical Trial of CAP-002 Gene Therapy in Pediatric Patients with Syntaxin-Binding Protein 1 (STXBP1) Encephalopathy. *Clinicaltrials.gov* (2026).

17. Gallo, N. B. *et al.* Identification of a therapeutic threshold for AAV-STXBP1 gene therapy in a rodent model of STXBP1 developmental and epileptic encephalopathy. *Molecular Therapy* **34**, 3763–3782 (2026).

18. Paremskaia, A. I. *et al.* Codon-optimization in gene therapy: promises, prospects and challenges. *Front Bioeng Biotechnol* **12**, 1371596 (2024).

19. Ertl, H. C. J. Immunogenicity and toxicity of AAV gene therapy. *Frontiers in Immunology* **13**, 975803 (2022).

20. Faust, S. M. *et al.* CpG-depleted adeno-associated virus vectors evade immune detection. *J Clin Invest* **123**, 2994–3001 (2013).

21. Sidorenko, L. V. *et al.* GC-rich coding sequences reduce transposon-like, small RNA-mediated transgene silencing. *Nature Plants* **3**, 875–884 (2017).

22. Sullivan, P. J. *et al.* Data-driven insights to inform splice-altering variant assessment. *Am J Hum Genet* **112**, 764–778 (2025).

23. Wang, H. *et al.* Case report: A novel STXBP1 splice variant and the landscape of splicing-involved STXBP1-related disorders. *Frontiers in Neurology* **14**, (2023).

24. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).

25. UniProt Consortium. UniProtKB: P61764 (STXBP1_HUMAN; Syntaxin-binding protein 1). *UniProt* https://www.uniprot.org/uniprotkb/P61764/entry (2026).

26. Grote, A. *et al.* JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Research* **33**, W526-W531 (2005).

27. Puigbò, P., Guzmán, E., Romeu, A. & Garcia-Vallvé, S. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res* **35**, W126–31 (2007).

28. GenSmart Codon Optimization. *Genscript.com* https://www.genscript.com/tools/gensmart-codon-optimization?src=pullmenu&utm (2020).

29. SciTools: Web Tools for Oligo Design & More | IDT. *IDT* https://www.idtdna.com/page/tools (2025).

30. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

31. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

32. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e24 (2019).

33. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377–94 (2004).

34. Houdayer, C. *et al.* Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum Mutat* **33**, 1228–38 (2012).

35. Jang, W., Park, J., Chae, H. & Kim, M. Comparison of In Silico Tools for Splice-Altering Variant Prediction Using Established Spliceogenic Variants: An End-User's Point of View. *International Journal of Genomics* **2022**, 1–6 (2022).

36. Dulubova, I. *et al.* A conformational switch in syntaxin during exocytosis: role of munc18. *EMBO J* **18**, 4372–82 (1999).

Appendices

Appendix A

Complete sequence composition of wild-type and codon optimised STXBP1 variants

Sequence	Length (bp)	GC (%)	GC3 (%)	CpG Count	CpG O/E	CAI	Identity to WT (%)	# of nucleotide changes to wt
Wildtype	1785	52.94	68.57	68	0.545	0.83	100.00	0
Optimizer	1785	63.75	99.83	124	0.685	1.00	83.08	302
JCAT	1785	63.75	99.83	133	0.739	0.99	83.14	301
IDT (legacy)	1785	50.25	63.19	88	0.782	0.79	78.60	382
IDT (pilot)	1785	55.13	76.81	31	0.229	0.90	82.07	320
GenSmart	1785	54.68	75.97	79	0.593	0.93	80.39	350

Appendix B

Complete SpliceAI-flagged site table

Variant	**Site type**	**Position (nt)**	**WT probability**	**Variant probability**	**Δ**	**Direction**
IDT_legacy	donor	1552	0.0004	0.2980	+0.2977	created
GenSmart	donor	1584	0.3126	0.0064	−0.3062	removed
IDT_legacy	donor	1584	0.3126	0.0002	−0.3124	removed
IDT_pilot	donor	1584	0.3126	0.0530	−0.2597	removed
JCat	donor	1584	0.3126	0.0213	−0.2914	removed
OPTIMIZER	donor	1584	0.3126	0.0158	−0.2969	removed

Appendix C

Complete MaxEntScan flagged site table

Variant	Site type	Position (nt)	Motif	WT score (bits)	Variant score (bits)	Δ (bits)	Relative change	Class	Distance to native junction (nt)
GenSmart	acceptor	322	CGCT CACG TGTT CTTC ACAG ACA	−0.78	8.58	9.36	n/a	created	3
IDT_legacy	donor	1553	CAGG TACG G	0.17	10.88	10.71	n/a	created	6
IDT_legacy	acceptor	322	CGCA CATG TCTT CTTC ACAG ATA	−0.78	9.32	10.11	n/a	created	3
IDT_legacy	donor	955	CCGG TGAG A	−0.65	8.72	9.37	n/a	created	8
IDT_legacy	acceptor	987	CAAT GCGC GATC TCTC ACAG ATG	3.64	6.66	3.02	0.831	strengthened	24
IDT_legacy	donor	1681	ATGG TAAA T	3.62	6.41	2.78	0.768	strengthened	21
IDT_legacy	donor	569	GCGG TACC G	4.51	5.34	0.84	0.186	strengthened	9
IDT_pilot	acceptor	326	CATG TGTT CTTC ACTG ACAG CTG	−7.74	9.67	17.41	n/a	created	1

JCat	acceptor	326	CACGTGTTCTTCACCGACAGCTG	−7.74	9.03	16.77	n/a	created	1
JCat	donor	894	GAGGTGAGC	1.98	8.70	6.73	n/a	created	8
OPTIMIZER	acceptor	326	CACGTGTTCTTCACCGACAGCTG	−7.74	9.03	16.77	n/a	created	1
OPTIMIZER	donor	1553	CCGGTACGG	0.17	9.77	9.60	n/a	created	6
OPTIMIZER	donor	894	GAGGTGAGC	1.98	8.70	6.73	n/a	created	8
OPTIMIZER	donor	569	GCGGTACCG	4.51	5.34	0.84	0.186	strengthened	9

Appendix D

Rescued FASTA Sequences per optimisation tool.

>STXBP1_GenSmart_rescued
ATGGCCCCTATCGGCCTGAAGGCCGTGGTGGGCGAAAAAATCATGCACGACGTGATCAAG
AAGGTGAAAAAAAAGGGCGAGTGGAAGGTGCTCGTGGTGGACCAGCTTAGCATGAGAATG
CTGTCCTCTTGCTGCAAGATGACAGATATCATGACCGAGGGCATTACCATCGTGGAAGAC
ATCAACAAGCGGCGGGAACCCCTGCCTTCACTGGAAGCCGTGTACCTGATCACACCAAGC
GAGAAGAGCGTGCACAGCCTGATCAGCGACTTCAAGGACCCTCCTACAGCCAAATATAGA
GCCGCTCACGTGTTCTTCACGGACAGCTGTCCTGACGCCCTGTTCAACGAACTGGTCAAG
TCCAGAGCAGCTAAGGTCATCAAAACCCTGACAGAGATCAACATCGCCTTCCTGCCATAC
GAGAGCCAAGTGTACTCTCTGGACAGCGCCGACTCTTTTCAGAGCTTTTACAGCCCCCAC
AAGGCCCAGATGAAAAACCCCATCCTGGAACGGCTGGCCGAGCAGATCGCCACCCTGTGC
GCCACACTGAAGGAATACCCTGCCGTGAGATACCGGGGCGAGTATAAGGACAACGCCCTG
CTGGCTCAGCTGATCCAGGACAAACTGGACGCTTACAAAGCTGACGATCCCACCATGGGC

GAGGGCCCTGATAAGGCCAGAAGCCAGCTGCTAATCCTGGATAGAGGATTTGACCCCAGC
AGCCCTGTGCTGCACGAGCTGACCTTCCAGGCCATGAGCTACGACCTGCTGCCAATCGAG
AACGACGTGTACAAGTACGAAACGAGCGGCATCGGCGAGGCCAGAGTGAAGGAAGTGCTG
TTGGATGAGGACGACGACCTGTGGATCGCCCTGCGGCACAAACATATCGCTGAAGTGTCC
CAGGAGGTGACAAGATCCCTGAAAGATTTCTCATCTTCTAAGAGGATGAACACCGGCGAG
AAAACAACCATGCGGGACCTGAGCCAGATGCTGAAGAAGATGCCTCAGTACCAAAAGGAG
CTGTCCAAGTACAGCACCCACCTCCATCTGGCCGAGGACTGCATGAAGCACTACCAGGGC
ACAGTGGACAAGCTGTGTAGGGTTGAACAGGATCTGGCAATGGGAACAGACGCCGAGGGA
GAAAAGATCAAGGACCCCATGAGAGCCATCGTGCCCATCCTGCTGGATGCCAACGTGTCT
ACCTATGATAAGATCCGGATCATTTTGCTGTACATCTTCCTGAAGAATGGCATCACCGAA
GAGAACCTGAACAAGCTGATTCAGCACGCCCAGATCCCTCCAGAGGATTCTGAGATAATC
ACCAACATGGCCCACCTGGGCGTGCCTATCGTCACAGATAGCACCCTGAGACGGCGCTCC
AAGCCTGAGAGAAAGGAAAGAATTTCTGAACAGACCTACCAGCTGAGCAGATGGACCCCT
ATCATCAAGGATATCATGGAAGATACCATCGAGGACAAGCTGGACACCAAGCACTACCCT
TACATCTCTACTAGATCCTCCGCCAGCTTCAGCACCACCGCCGTTAGCGCCAGATACGGC
CACTGGCACAAGAACAAGGCTCCTGGCGAGTACAGAAGCGGACCTAGACTGATCATCTTC
ATCCTGGGCGGAGTGAGCCTGAATGAGATGCGCTGCGCCTACGAGGTGACACAGGCCAAT
GGCAAGTGGGAGGTGCTGATCGGCAGCACCCACATCCTGACCCCCCAAAAGCTGCTGGAC
ACCCTGAAGAAACTGAACAAGACCGACGAGGAAATCAGCAGCTGA

>STXBP1_IDT_legacy_rescued
ATGGCGCCCATCGGATTGAAGGCGGTTGTAGGCGAGAAGATCATGCACGACGTGATTAAG
AAAGTCAAGAAGAAGGGCGAGTGGAAGGTCTTGGTGGTTGATCAGCTCTCCATGCGGATG
CTTTCATCATGCTGCAAAATGACCGATATCATGACCGAGGGTATAACGATCGTTGAGGAC
ATCAATAAAAGAAGGGAACCACTGCCTTCCCTGGAGGCAGTCTATCTGATAACACCCTCC
GAGAAGTCCGTGCACTCCCTGATAAGCGACTTCAAAGATCCGCCAACTGCCAAATATCGG
GCCGCACATGTCTTCTTCACTGATAGCTGTCCCGATGCTCTCTTTAATGAGCTGGTTAAA
TCAAGGGCCGCCAAGGTTATTAAAACGCTGACTGAAATCAACATCGCATTCTTGCCATAC
GAATCCCAGGTCTACAGTCTGGATTCTGCGGATTCCTTCCAGTCCTTTTACTCACCGCAC
AAAGCACAGATGAAAAACCCAATTCTGGAGAGACTGGCCGAGCAAATCGCCACCCTTTGT
GCCACTCTTAAAGAATATCCAGCCGTGCGGTACCGCGGAGAGTACAAGGACAATGCCCTG
CTCGCTCAGTTGATCCAGGACAAGCTGGACGCCTACAAGGCTGACGACCCGACCATGGGC
GAGGGTCCAGATAAGGCTCGCAGCCAACTGCTCATCTTGGACCGAGGATTTGACCCCTCC
TCCCCAGTGCTGCATGAACTGACTTTCCAAGCCATGAGTTACGACCTCTTGCCCATCGAA
AATGACGTCTACAAATACGAGACCTCAGGAATCGGGGAAGCCCGCGTCAAAGAGGTTCTT
CTGGATGAGGATGATGATCTGTGGATCGCCTTGCGACACAAGCACATCGCGGAGGTCAGC
CAAGAGGTCACACGGTCTTTGAAGGACTTTTCCTCCTCTAAAAGAATGAATACCGGAGAG
AAAACGACAATGCGCGATCTCTCACAGATGCTTAAAAAGATGCCTCAATACCAGAAGGAG
CTCAGCAAATACTCTACCCACCTTCACCTGGCTGAAGATTGCATGAAACACTACCAGGGG
ACCGTAGACAAACTCTGCAGGGTGGAACAGGATTTGGCTATGGGCACCGATGCTGAAGGG
GAGAAGATCAAGGACCCAATGAGGGCCATCGTTCCAATCCTCCTCGACGCCAACGTCAGC
ACATACGATAAGATAAGAATAATTCTGCTGTATATCTTTCTCAAGAACGGTATCACCGAG
GAAAACCTGAACAAGCTCATCCAGCACGCTCAAATCCCCCCAGAGGATTCTGAGATTATT
ACGAATATGGCGCACCTCGGGGTTCCAATAGTGACGGATTCTACCCTGCGCAGAAGGTCT
AAGCCGGAGCGGAAGGAACGGATTAGCGAGCAGACCTATCAGCTCAGCAGATGGACGCCT
ATTATTAAAGACATCATGGAGGACACTATCGAAGATAAACTGGACACGAAGCACTATCCT
TACATCTCAACAAGGAGCTCCGCAAGCTTTAGCACAACCGCAGTTTCCGCCAGATACGGT
CACTGGCATAAGAACAAGGCTCCAGGAGAGTATAGGTCCGGACCACGACTGATCATATTC
ATCCTGGGGGGCGTCTCTCTGAACGAGATGCGGTGCGCTTATGAAGTGACGCAGGCCAAT
GGTAAATGGGAGGTCCTGATTGGAAGCACCCACATCCTTACTCCACAGAAACTTCTCGAC
ACCCTCAAAAAGCTCAATAAAACAGATGAAGAGATTTCTTCCTAA

>STXBP1_IDT_pilot_rescued
ATGGCGCCCATTGGGTTAAAAGCCGTGGTGGGAGAGAAGATCATGCATGATGTGATCAAG

AAGGTGAAGAAGAAAGGAGAGTGGAAAGTCCTGGTGGTGGACCAGCTGAGCATGCGGATG
CTGTCCAGCTGCTGCAAGATGACAGACATCATGACTGAAGGCATCACCATTGTGGAGGAC
ATCAACAAGCGGCGGGAGCCCTTGCCCAGCCTGGAGGCTGTGTACCTGATCACACCCTCT
GAGAAGTCTGTGCACAGCCTGATCTCCGACTTTAAGGATCCACCCACAGCCAAGTACCGG
GCTGCCCATGTGTTCTTCACTGACTCCTGCCCAGATGCCTTGTTCAATGAGCTGGTCAAG
TCCAGAGCTGCCAAAGTGATCAAGACCCTCACGGAGATCAACATTGCCTTCCTGCCCTAT
GAGAGCCAGGTCTATTCCCTGGATTCAGCCGACTCCTTCCAGAGCTTCTACTCACCCCAC
AAGGCTCAGATGAAGAACCCCATCCTGGAGAGGCTGGCTGAGCAAATTGCCACACTGTGT
GCCACCCTGAAGGAATATCCTGCTGTGCGCTACCGAGGGGAGTACAAGGACAACGCTCTG
CTGGCCCAGCTCATCCAGGACAAGCTGGATGCCTACAAGGCTGATGACCCCACCATGGGA
GAGGGGCCAGACAAAGCTAGAAGCCAGCTGCTGATCCTGGACAGAGGCTTTGACCCATCC
AGCCCTGTGCTGCACGAGCTGACCTTCCAGGCCATGTCCTATGACCTGCTGCCCATTGAG
AATGATGTGTACAAGTATGAGACATCTGGCATTGGGGAGGCCAGGGTGAAGGAGGTGCTG
CTGGACGAGGATGATGACTTGTGGATTGCCCTGAGACACAAGCACATTGCTGAGGTCTCC
CAGGAGGTGACCCGCAGCCTGAAGGACTTCAGCAGCTCCAAGAGGATGAACACTGGAGAG
AAGACCACCATGCGGGACCTGTCCCAGATGCTGAAGAAGATGCCCCAGTACCAGAAGGAG
CTCTCCAAATACTCCACACACCTGCACCTGGCTGAGGACTGCATGAAGCACTACCAGGGC
ACTGTGGACAAGCTGTGCCGCGTGGAGCAGGATCTGGCCATGGGCACCGATGCTGAAGGA
GAGAAGATCAAGGACCCCATGAGAGCCATTGTGCCCATCCTGTTGGATGCCAATGTGTCC
ACCTATGACAAAATTCGGATTATTCTTCTGTACATCTTCCTGAAGAATGGCATCACTGAG
GAGAACCTGAACAAGCTGATTCAACATGCCCAGATTCCTCCTGAAGACTCTGAAATCATC
ACCAACATGGCCCACCTGGGTGTGCCCATTGTGACTGACTCCACCCTGCGGAGGAGGAGC
AAGCCTGAGAGGAAGGAGCGGATCAGCGAGCAAACCTACCAGCTGAGCCGGTGGACACCC
ATCATCAAGGACATTATGGAGGACACCATTGAGGACAAGCTGGACACCAAGCACTACCCC
TACATCTCCACCCGCTCTTCAGCCTCCTTCTCCACCACAGCTGTGTCTGCCCGCTATGGG
CATTGGCACAAGAACAAGGCACCTGGCGAGTACAGGAGCGGGCCCCGCCTGATCATCTTC
ATTCTGGGGGGCGTCTCCCTGAATGAGATGCGCTGTGCCTATGAAGTGACCCAGGCCAAT
GGCAAGTGGGAGGTGCTGATTGGCAGCACACACATTCTGACCCCCCAGAAGCTGCTGGAC
ACATTGAAGAAGCTCAATAAAACTGATGAGGAGATCTCCTCCTGA

>STXBP1_JCAT_rescued
ATGGCCCCCATCGGCCTGAAGGCCGTGGTGGGCGAGAAGATCATGCACGACGTGATCAAG
AAGGTGAAGAAGAAGGGCGAGTGGAAGGTGCTGGTGGTGGACCAGCTGAGCATGCGCATG
CTGAGCAGCTGCTGCAAGATGACCGACATCATGACCGAGGGCATCACCATCGTGGAGGAC
ATCAACAAGCGCCGCGAGCCCCTGCCCAGCCTGGAGGCCGTGTACCTGATCACCCCCAGC
GAGAAGAGCGTGCACAGCCTGATCAGCGACTTCAAGGACCCCCCCACCGCCAAGTACCGC
GCCGCCCACGTGTTCTTCACCGACTCCTGCCCCGACGCCCTGTTCAACGAGCTGGTGAAG
AGCCGCGCCGCCAAGGTGATCAAGACCCTGACCGAGATCAACATCGCCTTCCTGCCCTAC
GAGAGCCAGGTGTACAGCCTGGACAGCGCCGACAGCTTCCAGAGCTTCTACAGCCCCCAC
AAGGCCCAGATGAAGAACCCCATCCTGGAGCGCCTGGCCGAGCAGATCGCCACCCTGTGC
GCCACCCTGAAGGAGTACCCCGCCGTGCGCTACCGCGGCGAGTACAAGGACAACGCCCTG
CTGGCCCAGCTGATCCAGGACAAGCTGGACGCCTACAAGGCCGACGACCCCACCATGGGC
GAGGGCCCCGACAAGGCCCGCAGCCAGCTGCTGATCCTGGACCGCGGCTTCGACCCCAGC
AGCCCCGTGCTGCACGAGCTGACCTTCCAGGCCATGAGCTACGACCTGCTGCCCATCGAG
AACGACGTGTACAAGTACGAGACCAGCGGCATCGGCGAGGCCCGCGTGAAGGAGGTGCTG
CTGGACGAGGACGACGACCTGTGGATCGCCCTGCGCCACAAGCACATCGCCGAAGTCAGC
CAGGAGGTGACCCGCAGCCTGAAGGACTTCAGCAGCAGCAAGCGCATGAACACCGGCGAG
AAGACCACCATGCGCGACCTGAGCCAGATGCTGAAGAAGATGCCCCAGTACCAGAAGGAG
CTGAGCAAGTACAGCACCCACCTGCACCTGGCCGAGGACTGCATGAAGCACTACCAGGGC
ACCGTGGACAAGCTGTGCCGCGTGGAGCAGGACCTGGCCATGGGCACCGACGCCGAGGGC
GAGAAGATCAAGGACCCCATGCGCGCCATCGTGCCCATCCTGCTGGACGCCAACGTGAGC
ACCTACGACAAGATCCGCATCATCCTGCTGTACATCTTCCTGAAGAACGGCATCACCGAG
GAGAACCTGAACAAGCTGATCCAGCACGCCCAGATCCCCCCCCGAGGACAGCGAGATCATC
ACCAACATGGCCCACCTGGGCGTGCCCATCGTGACCGACAGCACCCTGCGCCGCCGCAGC

AAGCCCGAGCGCAAGGAGCGCATCAGCGAGCAGACCTACCAGCTGAGCCGCTGGACCCCC
ATCATCAAGGACATCATGGAGGACACCATCGAGGACAAGCTGGACACCAAGCACTACCCC
TACATCAGCACCCGCAGCAGCGCCAGCTTCAGCACCACCGCCGTGAGCGCCCGCTACGGC
CACTGGCACAAGAACAAGGCCCCCGGCGAGTACCGCAGCGGCCCCCGCCTGATCATCTTC
ATCCTGGGCGGCGTGAGCCTGAACGAGATGCGCTGCGCCTACGAGGTGACCCAGGCCAAC
GGCAAGTGGGAGGTGCTGATCGGCAGCACCCACATCCTGACCCCCCAGAAGCTGCTGGAC
ACCCTGAAGAAGCTGAACAAGACCGACGAGGAGATCAGCAGCTAA

>STXBP1_OPTIMIZER_rescued
ATGGCCCCCATCGGCCTGAAGGCCGTGGTGGGCGAGAAGATCATGCACGACGTGATCAAG
AAGGTGAAGAAGAAGGGCGAGTGGAAGGTGCTGGTGGTGGACCAGCTGAGCATGCGGATG
CTGAGCAGCTGCTGCAAGATGACCGACATCATGACCGAGGGCATCACCATCGTGGAGGAC
ATCAACAAGCGGCGGGAGCCCCTGCCCAGCCTGGAGGCCGTGTACCTGATCACCCCCAGC
GAGAAGAGCGTGCACAGCCTGATCAGCGACTTCAAGGACCCCCCCACCGCCAAGTACCGG
GCCGCCCACGTGTTCTTCACCGACTCCTGCCCCGACGCCCTGTTCAACGAGCTGGTGAAG
AGCCGGGCCGCCAAGGTGATCAAGACCCTGACCGAGATCAACATCGCCTTCCTGCCCTAC
GAGAGCCAGGTGTACAGCCTGGACAGCGCCGACAGCTTCCAGAGCTTCTACAGCCCCCAC
AAGGCCCAGATGAAGAACCCCATCCTGGAGCGGCTGGCCGAGCAGATCGCCACCCTGTGC
GCCACCCTGAAGGAGTACCCCGCCGTGCGGTACCGGGGCGAGTACAAGGACAACGCCCTG
CTGGCCCAGCTGATCCAGGACAAGCTGGACGCCTACAAGGCCGACGACCCCACCATGGGC
GAGGGCCCCGACAAGGCCCGGAGCCAGCTGCTGATCCTGGACCGGGGCTTCGACCCCAGC
AGCCCCGTGCTGCACGAGCTGACCTTCCAGGCCATGAGCTACGACCTGCTGCCCATCGAG
AACGACGTGTACAAGTACGAGACCAGCGGCATCGGCGAGGCCCGGGTGAAGGAGGTGCTG
CTGGACGAGGACGACGACCTGTGGATCGCCCTGCGGCACAAGCACATCGCCGAAGTCAGC
CAGGAGGTGACCCGGAGCCTGAAGGACTTCAGCAGCAGCAAGCGGATGAACACCGGCGAG
AAGACCACCATGCGGGACCTGAGCCAGATGCTGAAGAAGATGCCCCAGTACCAGAAGGAG
CTGAGCAAGTACAGCACCCACCTGCACCTGGCCGAGGACTGCATGAAGCACTACCAGGGC
ACCGTGGACAAGCTGTGCCGGGTGGAGCAGGACCTGGCCATGGGCACCGACGCCGAGGGC
GAGAAGATCAAGGACCCCATGCGGGCCATCGTGCCCATCCTGCTGGACGCCAACGTGAGC
ACCTACGACAAGATCCGGATCATCCTGCTGTACATCTTCCTGAAGAACGGCATCACCGAG
GAGAACCTGAACAAGCTGATCCAGCACGCCCAGATCCCCCCCGAGGACAGCGAGATCATC
ACCAACATGGCCCACCTGGGCGTGCCCATCGTGACCGACAGCACCCTGCGGCGGCGGAGC
AAGCCCGAGCGGAAGGAGCGGATCAGCGAGCAGACCTACCAGCTGAGCCGGTGGACCCCC
ATCATCAAGGACATCATGGAGGACACCATCGAGGACAAGCTGGACACCAAGCACTACCCC
TACATCAGCACCCGGAGCAGCGCCAGCTTCAGCACCACCGCCGTGAGCGCCCGTTACGGC
CACTGGCACAAGAACAAGGCCCCCGGCGAGTACCGGAGCGGCCCCCGGCTGATCATCTTC
ATCCTGGGCGGCGTGAGCCTGAACGAGATGCGGTGCGCCTACGAGGTGACCCAGGCCAAC
GGCAAGTGGGAGGTGCTGATCGGCAGCACCCACATCCTGACCCCCCAGAAGCTGCTGGAC
ACCCTGAAGAAGCTGAACAAGACCGACGAGGAGATCAGCAGCTAA