

AI as a partner in scientific exploration

Wong Limsoon

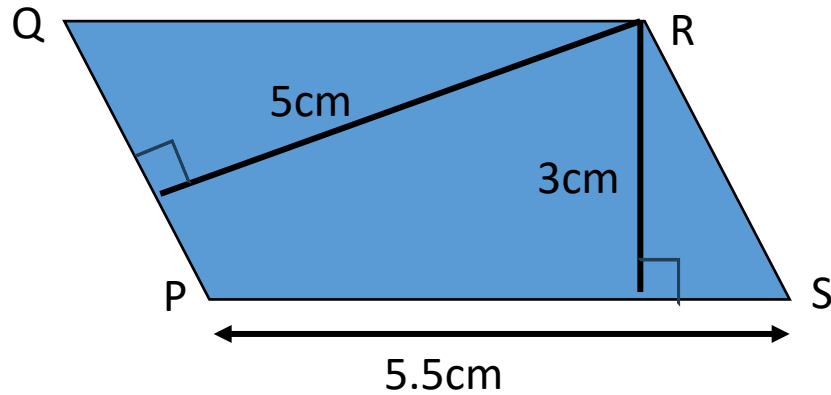


National University of Singapore

Story #1

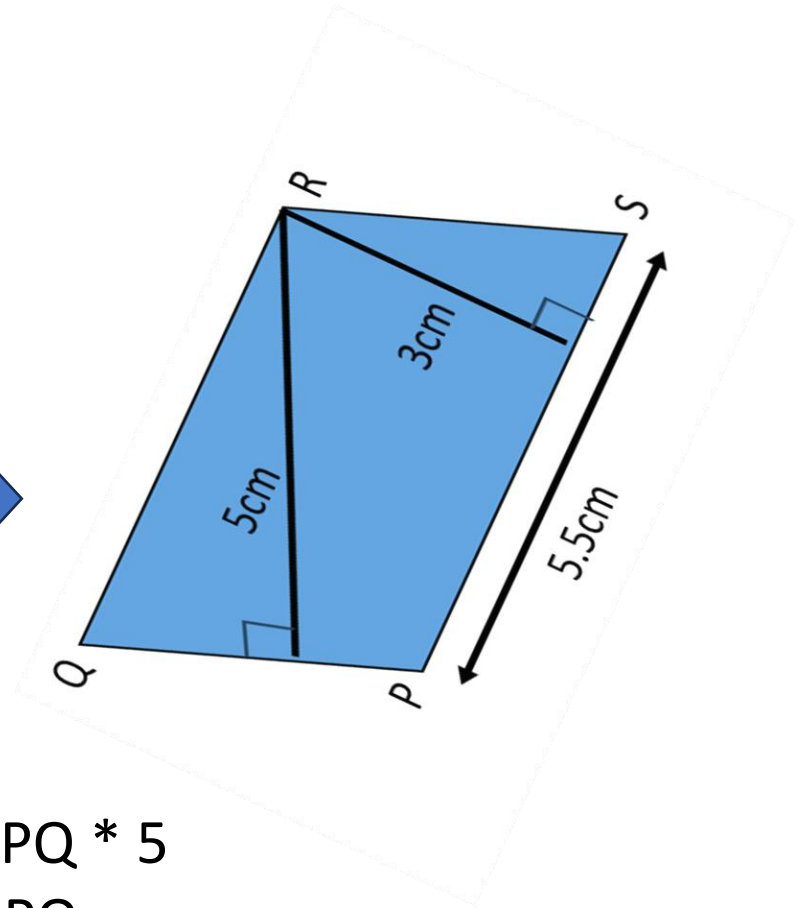
AI's aha moment: turn the shape, keep the area!

My son's homework



I am given a parallelogram PQRS. The base PS = 5.5cm. The height from PS to the opposite side QR is 3cm. A perpendicular line from PQ to the opposite corner R is 5cm. What is the perimeter of PQRS?

ChatGPT's
suggestion



$$\begin{aligned}PS * 3 &= PQ * 5 \\5.5 * 3/5 &= PQ \\3.3 &= PQ \\Perimeter &= 2 * 5.5 + 2 * 3.3 = 17.6\end{aligned}$$

| What the AI (ChatGPT) has demonstrated

Mathematical reasoning

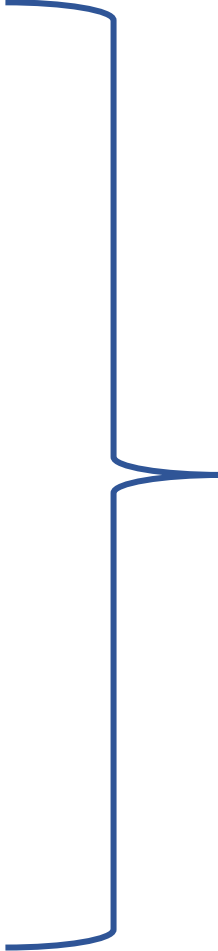
*Recognizing relationships
between area, base, and height*

Algebraic manipulation

*Solving for unknown side lengths
using equations*

Geometric understanding

*Switching between ways of
expressing area*

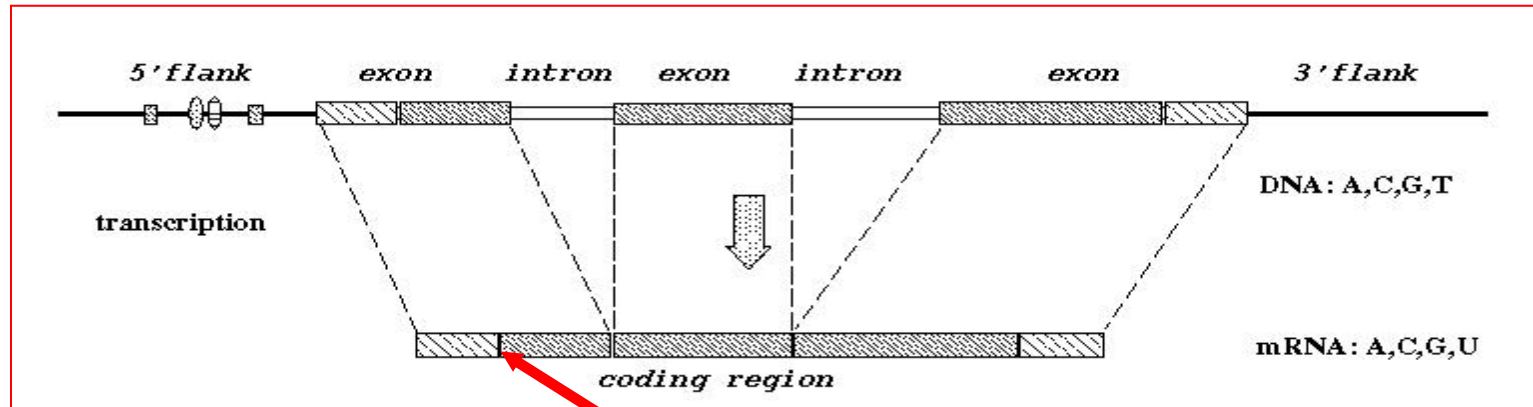


Connecting concepts
across subfields of
mathematics (geometry
and algebra) to solve
problems

Story #2

AI vs human insight: When AI knows biology but fumbles reasoning in new contexts

Translation initiation site

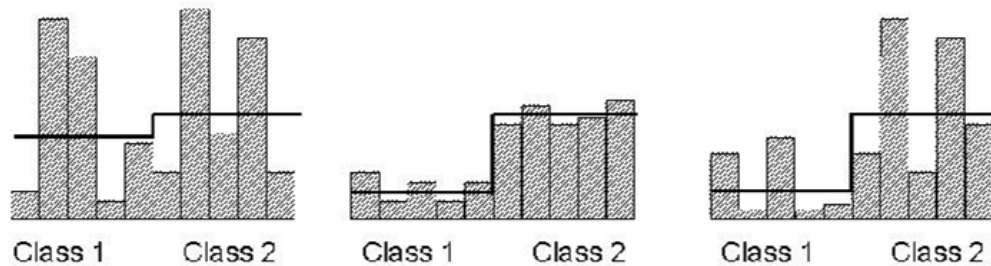


```

299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG      80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA      160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGGCCTGGTGCCGAGGA      240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACAAGGACTCCCCT
  
```

What makes the second ATG (AUG) the translation initiation site (TIS)?

Classical / statistical machine learning could recognize TIS well enough



CFS selected these features for recognizing TIS:

Position -3

in-frame upstream ATG

in-frame downstream

TAA, TAG, TGA,

CTG, GAC, GAG, and GCC



Apply NB or SVM left-to-right until first ATG predicted as positive. That's the TIS; skip the rest
NB & SVM models were trained using TIS vs. Upstream ATG

	Sensitivity	Specificity	Accuracy
NB+Scanning	87.3%	96.1%	93.9%
SVM+Scanning	88.5%	96.3%	94.4%

... but why these features?

| What ChatGPT said about position -3

WLS: Is the position -3 important for recognizing translation initiation sites in mRNA?

ChatGPT:

Yes, the **-3 position** is important for recognizing translation initiation sites. In eukaryotes, the sequence around the start codon (AUG) follows the **Kozak consensus**: (gcc)gccRccAUGG, where **R** (A or G) at **-3** is critical. A purine at **-3** enhances ribosome recognition and increases the likelihood that the AUG is used as the start site.

What ChatGPT said about in-frame up-stream ATG

WLS: *Is the number of in-frame up-stream ATG important for predicting translation initiation sites in mRNA?*

ChatGPT's initial response: *Ribosomes tend to initiate at the first ATG with a good context. So, a higher number of in-frame up-stream ATG makes it harder to predict TIS*

This part is wrong!

ChatGPT admits after a few rounds of interactions:

Yes, the number of **in-frame upstream ATGs** is important for predicting translation initiation sites. **Fewer upstream ATGs** increase the likelihood that a downstream ATG is the true start site, since ribosomes usually initiate at the first AUG in a favorable context. Multiple upstream ATGs may indicate **uORFs** or alternative starts, reducing the chance that a downstream ATG is used.

What ChatGPT said about in-frame down-stream TAA, TAG, TGA

WLS: Is the number of in-frame down-stream TAA, TAG, and TGA important for recognizing translation initiation sites?

ChatGPT:

- A real TIS typically leads to a **long open reading frame (ORF)** before hitting a stop codon.
- If an AUG is followed by **many in-frame stop codons close by**, it likely produces **short, nonfunctional peptides**, making it **less likely** to be a true TIS.
- Fewer early in-frame stops suggest a **longer, viable ORF**, supporting the AUG as a **more likely start site**.

What ChatGPT said about in-frame down-stream CTG, GAC, GAG, GCC

WLS: Is the number of in-frame down-stream CTG, GAC, GAG, and GCC important for recognizing translation initiation sites?

ChatGPT's initial response:

No, the number of in-frame downstream CTG, GAC, GAG, and GCC codons doesn't help recognize translation initiation sites because they are regular amino acid codons and don't affect start site selection.

This part is wrong!

Chat GPT admits after a few rounds of interactions:

These codons themselves don't signal starts or stops, but their **frequency patterns differ** between coding regions and UTRs due to **codon bias**. Since true translation start sites lead into coding regions, the presence and pattern of these codons downstream help identify whether an ATG is a real start site or just a UTR sequence.

| What ChatGPT has demonstrated

Knowledge of translation (e.g. scanning, Kozak, codon bias), but inadequate reasoning in explaining their impact on prediction making

ChatGPT summarizes its reasoning mistakes as “modeling bias”:

The initial oversight came from focusing on **mechanistic signals**—like AUG and stop codons—while overlooking how **statistical patterns** in codon usage can inform TIS prediction. Codons like CTG, GAC, GAG, and GCC don’t directly signal translation initiation, but their distribution downstream of an ATG can reflect **coding potential**. Your insight reframed the question from a purely biological mechanism to one involving **sequence-based statistical cues**, highlighting how such patterns help distinguish coding regions from UTRs—key for identifying true TIS.

Knowledge in silos; cannot synthesize across domains

Story #3

From rote to responsive: AI gets a little more context-smart

Many students taking classes on machine learning or AI cannot answer this question

On a given test set with 20% negative and 80% positive instances, a classifier C achieves 90% sensitivity and 90% precision

What is the likely sensitivity and precision if C is deployed in the real world where 80% of the population is negative and 20% is positive?

But ChatGPT also could not answer it a year ago

ChatGPT could answer this now

Given:

Test set: 80% positive, 20% negative.

Sensitivity = 90%, Precision = 90%.

Assume 1000 samples:

Positives = 800, Negatives = 200.

$TP = 0.9 \times 800 = 720$, $FN = 80$.

Precision: $720 / (720 + FP) = 0.9 \rightarrow FP = 80$.

$TN = 200 - 80 = 120$.

False positive rate = $80 / (80 + 120) = 0.4$ (40%).



Sensitivity
specificity
are both
preserved

Real world:

20% positive, 80% negative.

Assume 1000 samples: 200 positive, 800 negative.

$TP = 0.9 \times 200 = 180$, $FN = 20$.

$FP = 0.4 \times 800 = 320$, $TN = 480$.

New metrics:

Sensitivity = $180 / (180 + 20) = 90\%$ (same).

Precision = $180 / (180 + 320) = 36\%$.

Conclusion:

Sensitivity stays 90%.

Precision drops to 36% due to class shift.

| What ChatGPT has demonstrated

Quantitative problem solving

Applied performance metrics

Used sensitivity, precision, false positive rate, and specificity to analyze effects of class distribution changes

Assumed stable classifier behavior

Sensitivity and false positive rate stay constant from test set to real world

But does it know why?

Justifications provided by ChatGPT

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

ChatGPT assumes sensitivity and specificity are preserved because, by definition, they reflect the classifier's performance conditional on the true class, not the overall prevalence. This is a standard and reasonable assumption in the absence of additional data shifts—enabling valid recalculation of precision under new class balances.

An understated crucial assumption!

Perhaps ChatGPT is still unable to prioritize its reasoning well

Concluding observations

| AI as a research partner who ...

Earlier AI	Recent AI
Missed which assumptions or metrics are stable across scenarios	Recognizes and applies appropriate stable assumptions (e.g., metric preservation)
Struggled to adapt known knowledge to new or shifted contexts	Adapts reasoning to new contexts and data distributions
Often required correction or hints from user	Works through reasoning steps more independently and transparently

Misconception	What Actually Happens
AI has mechanistic understanding	AI recognizes patterns/statistics, not true mechanisms
AI reasons well in new/unfamiliar contexts	AI may misapply knowledge or need correction
AI gives answers based on causal reasoning	AI often relies on surface associations or memorized information
AI can generate new biological insights	AI is best at summarizing learned knowledge, not synthesizing novel insights
AI errors mean it lacks information	Errors often reflect reasoning limitations, not just knowledge gaps

Has decent math skills

Has broad background knowledge

Struggles to integrate across fields

Does not prioritize reasoning well

But it keeps learning