ardigen

Artificial Intelligence & Bioinformatics
for Precision Medicine

# PRISM

# - a writing assistant for the language of proteins

*Figure 1.*
*Minoan Linear B script from the Phaistos disk.*

Language is an inseparable companion of our species. To study humans means to investigate the language we use, how it shapes us, and how we use it to discover the world. Languages hold the key to understanding our past, often posing challenges to translators. Celebrated examples are Ancient Egyptian Hieroglyphs and the Linear B script from Mycenaean Greece. The heroic decoding of these languages by Champollion, Ventris, and Chadwick, allowed us to recover lost languages and decipher the cultural heritage of these long-gone civilizations. There is, however, an even more ancient language which orchestrates what happens inside our bodies-the language of proteins.

Proteins and peptides, the indispensable building blocks of living organisms, are written using a secret language.[1] Its characters are 20 types of amino acid residues which bind together to form words and sentences which carry messages and instructions inside the body. In reality, these messages take the form of complex three-dimensional molecules, whose detailed dynamics remain poorly understood. For instance, understanding the process by which proteins fold remains one of science's most compelling mysteries. Moreover, the information required to crack this puzzle is hard to obtain and often missing-only about 50% of the human proteome structure is known.[2]

Just as decoding ancient languages provides a window into understanding long-gone cultures, cracking the code of proteins will open several avenues for progress in biology. With this knowledge, we could simply read the mechanisms behind countless diseases and refine our methods for drug discovery. For instance, important functions of a protein and its interactions with other macromolecules, are mediated by structural domains called binding pockets.[3] These pockets can be imagined as small indentations or cavities on the surface of the molecule. Their positioning and shape, encoded by the amino acid sequence, dictate the type of process they will participate in. Most importantly, some of these pockets are potentially druggable sites. A better understanding of binding pockets would inform and streamline the in silico stage of drug development.

Treating proteins as three-dimensional quantum mechanical systems has proven extremely time and resource consuming. Could it be that the linguistic picture goes beyond a superficial analogy and leads to actual breakthroughs? We hinted at the amino acid chain-or primary structure-as raw text written with a 20 amino-acid alphabet.

But the power of language lies in how characters connect to create meaning. Perhaps it would be possible to import techniques of linguistics to biology, for example, by borrowing the concept of corpus, which collects amino acid data and extracts information about correlations between them. These correlations can help explain rules governing the "protein language" codifying formation of the higher-order structures. And all this can be achieved without prior knowledge about 3D structure.

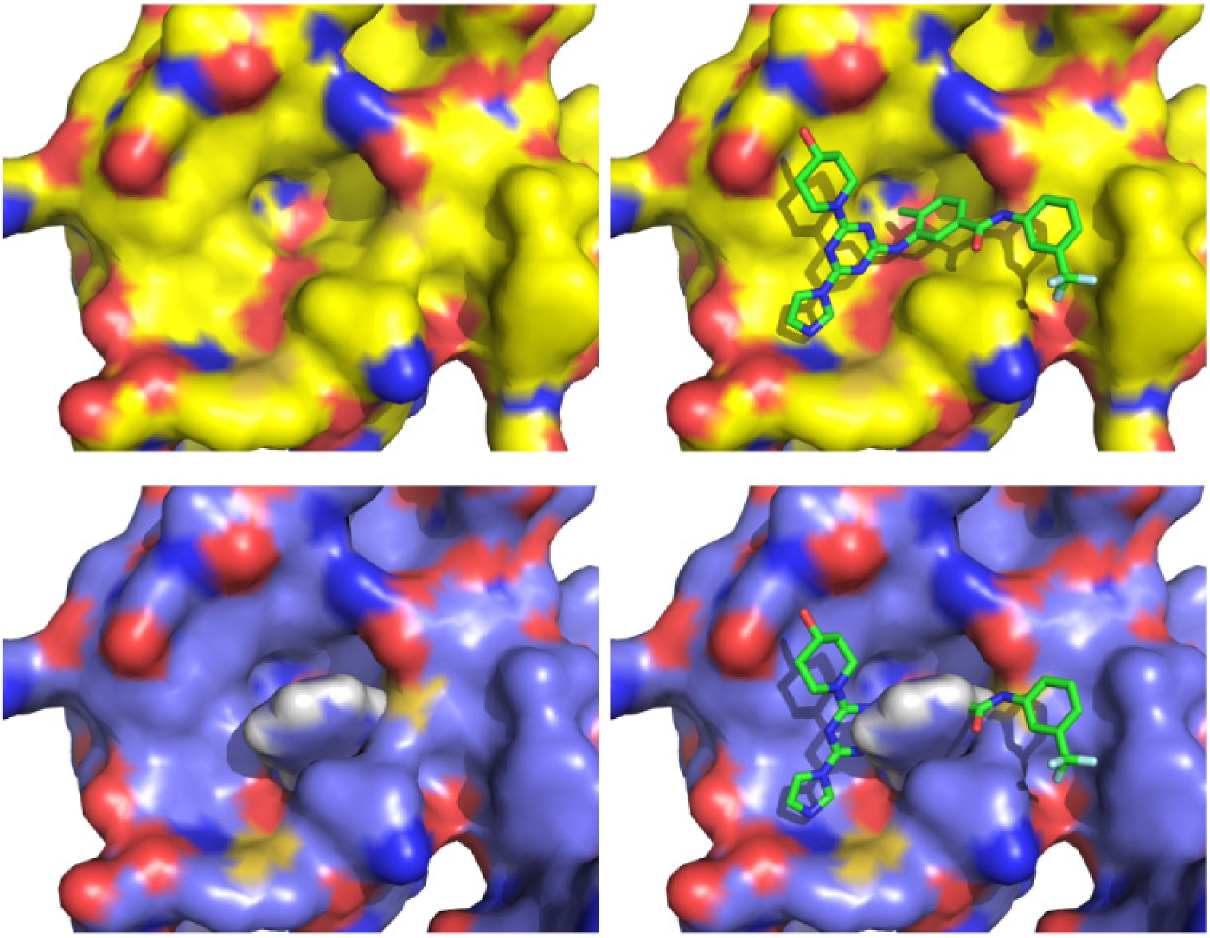


*Figure 2.*
*Representation of possible impact of changing a single amino acid within the binding pocket of a protein. Top left: a cross-section of a binding pocket of a kinase. Top right: The cross-section with a bound compound. Bottom left: The same protein with one amino acid changed to Tryptophan. Bottom right: imposed binding mode of the compound. This image clearly shows that changing a single amino acid can have a significant impact on the binding properties of a protein.*

Remarkably, this approach has proven quite promising. The last decade has seen momentous advances in ML-driven natural language processing (NLP) techniques. Methods for machine translation and text generation have reached astonishing levels of performance. Google's BERT[4] and OpenAI's GPT-3[5] stand as towering achievements in the field. The idea of using NLP methods such as these to decipher mysterious languages-whether ancient[6] or biological[7]-is gathering momentum.

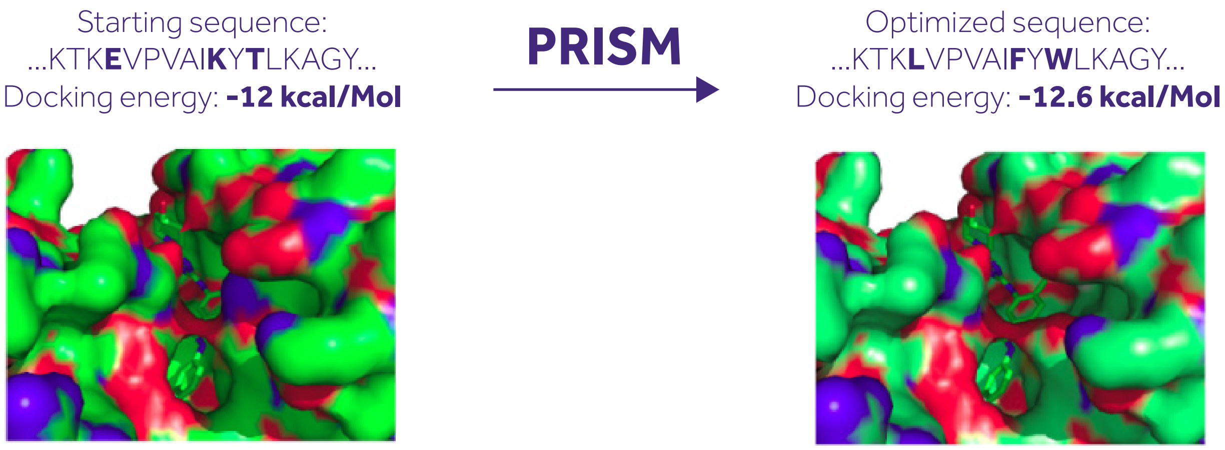

Starting sequence:
...KTK**E**VPVAI**K**Y**T**LKAGY...
Docking energy: **-12 kcal/Mol**
PRISM
Optimized sequence:
...KTK**L**VPVAI**F**Y**W**LKAGY...
Docking energy: **-12.6 kcal/Mol**

*Figure 3.*
*PRISM running on a peptide sequence and suggesting modifications that would reduce the docking energy and lead to a more bindable molecule.*

At Ardigen, we decided to take part in this exploration and use the linguistic approach to tackle questions of clinical relevance; namely, that of optimizing the binding capabilities of peptides. Our model-Protein Refinement by Intelligent Sequence Modification (**PRISM**) -is capable of recognizing the binding pockets in a peptide starting just from the raw amino acid sequence. Further, it is capable of estimating the "stickiness" of the pocket-by measuring its docking energy-and suggesting modifications to increase it. **PRISM** can deliver new insights for predicting protein-protein interactions and protein-peptide interactions. It can also suggest atypical pocket-forming sequences, opening up new avenues in pharmacology research.

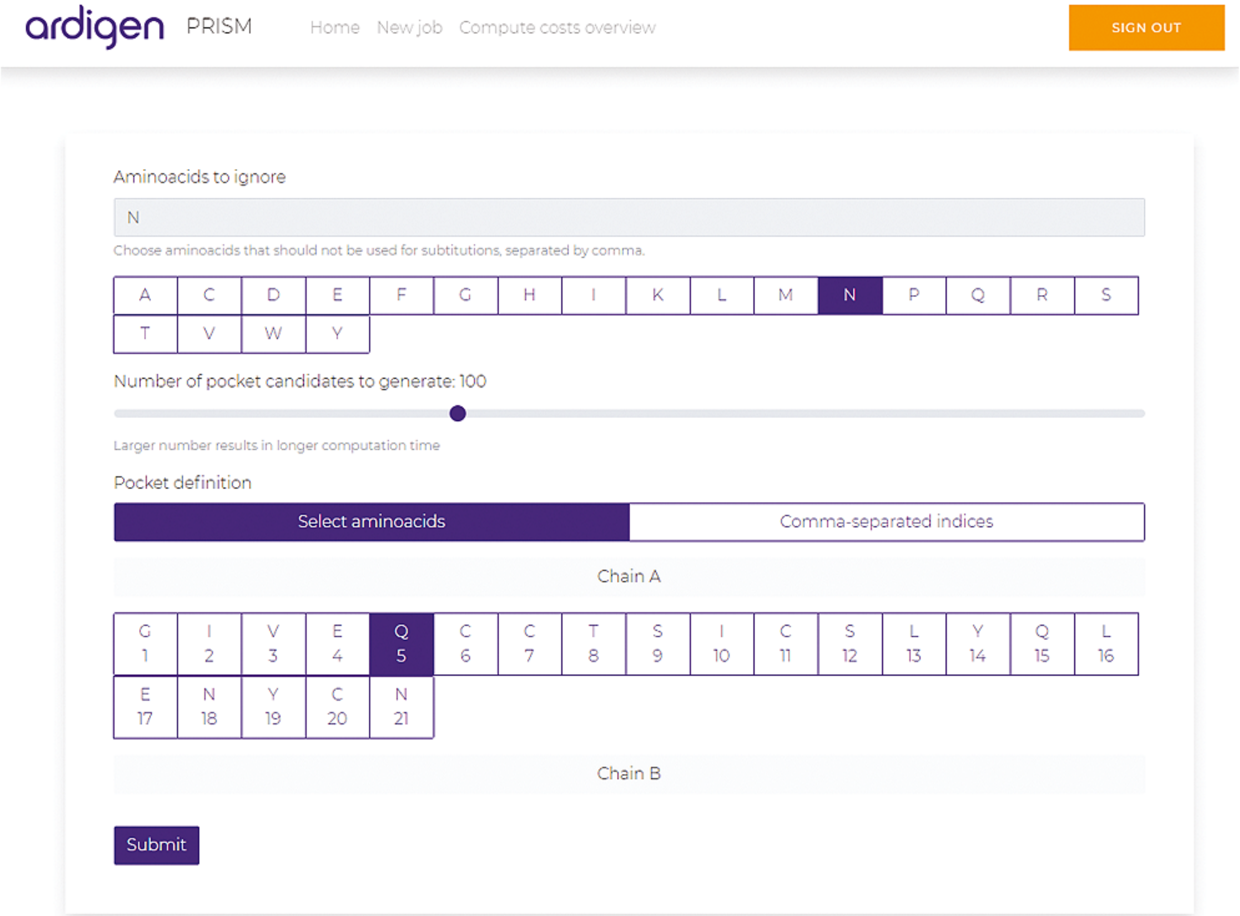
Step back and picture the enormous complexity of this problem. Finding the binding properties of such a complex molecule using three-dimensional modeling seems intractable. However, **PRISM** learns how to solve this problem in a surprisingly intuitive fashion. **PRISM** follows the concept of BERT, where the model internalizes the notion of sentence, including contextual information. First, **PRISM** looks at enormous corpora of protein sequences and learns to perform the task of sentence completion.

That is, if we present it with a partially covered sequence of amino acids, **PRISM** is capable of predicting the missing sequence with high accuracy. Later, **PRISM** is trained to recognize secondary structures, i.e. local geometry out of raw sequences. Finally, **PRISM** spends some training cycles looking at binding pockets until it internalizes the concept of "pocketedness".

With this knowledge, starting from raw sequences **PRISM** is capable of identifying binding pockets and exploring the space of real proteins-i.e. *grammatically* correct sequences-in search of more *bindable* alternatives.

In the linguistic analogy, we advocate, drug development consists of writing the right amino acid sentences to confront disease. **PRISM** acts like a writing assistant helping biomedical researchers streamline their efforts. We have fashioned this tool with a user-friendly GUI enabling remote access with cloud computing and fast processing advantages. Users just need to introduce the raw amino acid sequence and select their preferences. Equipped with **PRISM**, our customers can streamline their research efforts by starting the discovery process with highly optimised candidates. **PRISM** can help users improve the catalytic properties of enzymes, and introduce substrate selectivity.

> *The **PRISM** pipeline has dramatically improved our discovery efforts and uncovered new unexpected lead candidates. In addition, the deployment on cloud computing resources has allowed our discovery team to explore possibilities that would be untenable under traditional settings.*
>
> *Pawel Fludzinski, PhD. CEO, AmideBio*



Aminoacids to ignore

N

Choose aminoacids that should not be used for subtitutions, separated by comma.

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | C | D | E | F | G | H | I | K | L | M | [✓ N] | P | Q | R | S |
| T | V | W | Y |   |   |   |   |   |   |   |   |   |   |   |   |

Number of pocket candidates to generate: 100

Larger number results in longer computation time

Pocket definition

[✓ Select aminoacids] Comma-separated indices

Chain A

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G<br>1 | I<br>2 | V<br>3 | E<br>4 | [✓ Q<br>5] | C<br>6 | C<br>7 | T<br>8 | S<br>9 | I<br>10 | C<br>11 | S<br>12 | L<br>13 | Y<br>14 | Q<br>15 | L<br>16 |
| E<br>17 | N<br>18 | Y<br>19 | C<br>20 | N<br>21 |   |   |   |   |   |   |   |   |   |   |   |

Chain B

*Figure 4.*
*Fragment of the user interface of PRISM where user refine search for new binding pockets*

While a complete understanding of protein dynamics remains elusive, the linguistic picture has proven a viable approach. Just as cracking an ancient script, deciphering the language of proteins will bring an entirely new level of understanding. The code hasn't been broken yet, but today's Champollions are ML-enhanced.

## Bibliography:

1. Mohammed AlQuraishi, "End-to-End Differentiable Learning of Protein Structure" (2019) ***https://www.biorxiv.org/content/10.1101/265231v1***

2. Sirawit Ittisoponpisan, Suhail A. Islam, Turan Khanna, Eman Alhuzimi, Alessia David, Michael J.E. Stenberg, "Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated?" (2019) ***https://www.sciencedirect.com/science/article/pii/S0022283619302037***

3. Ryan G. Coleman, Kim A. Sharp, "Protein Pockets: Inventory, Shape, and Comparison" (2010) ***https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2859996/***

4. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention Is All You Need" (2017) ***https://arxiv.org/pdf/1706.03762.pdf***

5. Greg Brockman, Mira Murati, Peter Welinder, OpenAI, "OpenAI API" (2020) ***https://openai.com/blog/openai-api/***

6. Jiaming Luo, Yuan Cao, Regina Barzilay, "Neural Decipherment via Minimum-Cost Flow: from Ugaritic to Linear B" (2019) ***https://arxiv.org/pdf/1906.06718.pdf***

7. Neil Thomas, Nicholas Bhattacharya, Roshan Rao, "Can We Learn the Language of Proteins?" (2019) ***https://bair.berkeley.edu/blog/2019/11/04/proteins/***

ardigen

Artificial Intelligence & Bioinformatics
for Precision Medicine

EU: Podole 76, 30-394 Kraków,
Poland

US: 611 Gateway Boulevard
South San Francisco, CA 94080

EU: +48 12 340 94 94

US: +1 628 200 09 14

ardigen.com

michal.warchol@ardigen.com