Adventures of Pop – the undruggable protein

Pop is a celebrity protein worth billions of dollars, but mad scientists are after it. Its crime? Causing a rare disease while being undruggable. In **Adventures of Pop – the undruggable protein**, you'll learn how scientists tackle target-based drug discovery to finally find a cure for the disease caused by Pop, a task that is both very difficult and costly. Indeed, developing a single drug-discovery program can cost a few hundred million dollars, and factoring all the failures, a total investment of 2B$ and 10 years are needed for every drug that hits the market. Fortunately, new technologies such as **machine learning** are being developed to help alleviate the costs of drug development and increase the success rates.

The current visual analogy presents Pop as an anthropomorphic protein and the banana as the ligand, i.e. the tentative drug that chemically binds to it. The goal to develop a good drug is analogous to having Pop bite onto the banana and stop causing any harm (inhibit the protein). The changes in the position of its arms and legs are analogous to the conformational changes of the protein, i.e., how the protein structure is altered based on its current function and random motion.

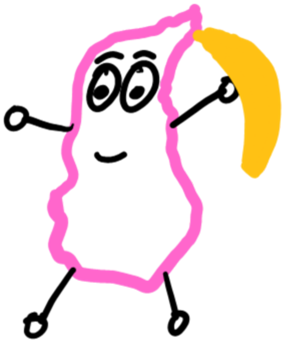
The mad scientists will do anything to get Pop to bite a banana. The problem is that Pop is very picky, he won't even hold on to most bananas, many bananas give him a nasty reaction, and the factors that drive Pop's reaction are very difficult for the mad scientists to pin down.

Part 1 - Experimental assays

In this first part, scientists try to find a drug for Pop using biochemical assays. Indeed, once scientists know they are targeting Pop, they usually turn to lab experiments. The questions they try to answer is **which** banana will Pop bite, **why** Pop likes it, and **how** Pop bites into it. These questions are difficult due to the extremely vast chemical space of 10^{60} possible ligands and the cost or speed limitations of running precise experiments. Over the past decades, scientists have developed many techniques to answer these questions, and they can all benefit from more recent machine learning algorithms.

Lab experiments are done by experts – don't try to replicate this at home.

1.1 Binding assays



Binding assay is a broad term for any of the various standard experiments that chemists perform to find a banana that Pop would like to eat. Scientists will generally opt for high-throughput / lower-precision assays in the early stages of drug discovery where they will try tens of thousands of bananas but without precisely knowing which bananas Pop loved the most. In later stages, scientists will opt for low-throughput / high-precision where only ~100 bananas are tested, and each experiment is repeated many times and more accurately find Pop's favorite bananas.

Such a high-throughput, lower precision approach may look like Surface Plasmon Resonance. In this assay it's like Pop is sitting on a tiny bench, waiting to take a bite out of a passing banana. Now, we introduce the banana into the system. Each time Pop takes a bite, it causes a signal change that we can measure. The rate at which Pop takes bites and how much of the banana it consumes tells us about how strongly Pop is binding to the banana (the "binding affinity") and how quickly it binds and releases ("binding kinetics").

There are a few informative metrics that can be extracted from these studies, notably **IC_{50}**, **EC_{50}**, and **K_d** .

IC_{50} looks at the inhibition curves to find the concentration of half-inhibition, i.e., the number of bananas Pop needs to eat to make it half-asleep. They are very relevant to drug discovery since we generally optimize for lower ligand concentrations. At high concentrations, many ligands are active but potentially dangerous. Indeed, if bananas surround Pop, it might be tempting to eat one, even if it's not its favorite food. However, perhaps one of Pop's friends is allergic to bananas ("off-target interaction", when a drug binds to proteins different from its original target) which can lead to major side effects, so we need to keep the concentration low to avoid undesired interactions (more about that later). **EC_{50}** is a similar assay that looks at the activity rather than the concentration.

$K_d = K_{off} / K_{on}$ is an equation governing the dissociation and association rate constants, respectively. The K_{on} is the rate at which Pop picks a banana; higher values are indicative of a better ligand that is more likely to bind to the protein. The K_{off} is the rate at which Pop lets go of the banana, with lower values indicative of a better ligand that stays bound for longer. Hence, by measuring the ratio $\mathbf{K}_d$, we can effectively capture the effectiveness of a ligand at binding and staying bound.

These metrics are generally expensive to obtain as they require multiple replicates over multiple worlconcentrations, and often lack reproducibility across labs, meaning that results from different publications are not directly comparable. For these reasons, there are efforts to create approaches to binding assays that increase the throughput and repeatability. Further, the need to synthesize all of the bananas (candidate molecules) that one wants to test makes it difficult to scale.

Machine learning models are often trained to predict properties such as $IC_{50}/EC_{50}/K_d$ given a molecular structure. Attempts have been made to build a generalizable model that can use both target (Pop) and ligand (Bananas) representations, a reasonable way of increasing the amount and richness of the available data. However, due to challenges in combining various data sources, limited success has been observed so far [MolTrans, MoDTI].

1.2 DNA-encoded libraries

Scientists have developed DNA-encoded libraries (DEL) to reduce the cost of experiments while also increasing the number of small molecule ligands (or bananas) that Pop will try.

DELs use recent advances and cost reductions in DNA sequencing technologies. They work by testing thousands of ligands simultaneously in a protein assay, where each banana has a DNA "flag" attached to it allowing its identification. Then, the ligands that bind to the protein are identified by washing out unbound ligands and amplifying the DNA signal of the bound ligands. This way, we can throw many bananas at Pop and still keep track of which bananas it prefers.

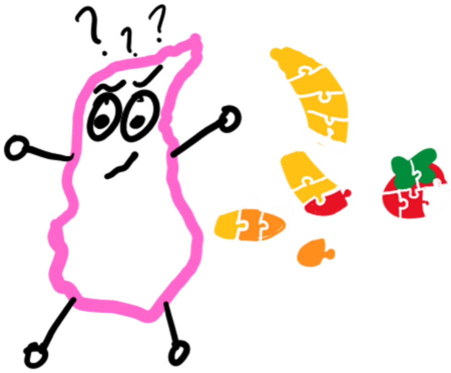
Issue 1: Pop doesn't like to bite on a flag –, the DNA identifiers are often much larger than the ligand and can alter how they interact (our drawings showing the little flags on the banana are not to scale).

Issue 2: Pop cannot eat more than a few bananas, so it will choose the most delicious one. This competition implies that only the strongest binders will be identified as they take the binding pocket for themselves.

Issue 3: It is difficult to customize the bananas in the DEL libraries, so they are mostly used for hit finding in the early stages of the drug discovery project.

Machine learning models can substantially benefit from DEL screens, since they provide much larger amounts of data than traditional binding assays. However, due to the two issues above, there is still a need to confirm the predictions with traditional experiments.

1.3 Fragment-based



Another approach to increase the throughput while reducing the experimental cost is to screen fragments, and then assemble the fragments into larger molecule ligands. Just like a sandwich, the order, orientation, and synergy with which the ingredients are assembled is crucial to Pop liking the resulting ligand.

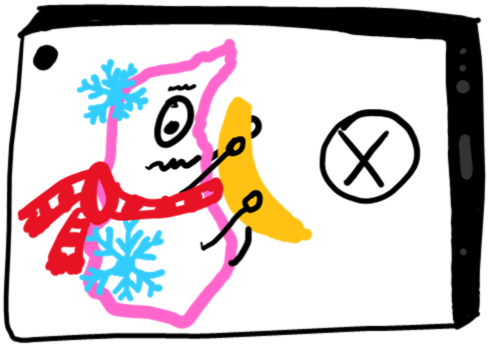
Indeed, the fragments must be complementary and act *synergistically*, meaning that Pop would be more likely to eat the “dish” than any of the individual ingredients. Pop might like tomatoes and bananas, but a tomato/banana puree doesn’t sound appetizing. However, a banana pancake does sound better than having bananas and pancakes separately: they have synergy.

To create this synergy, the fragments must bind to different parts of the same pocket. Also, when assembled, they must be aligned such that each fragment is optimally positioned and oriented in the pocket, again just like a sandwich cannot have vertical tomato slices on horizontal bread slices.

Co-crystallization (see next section) is often used to study the fragments and where they bind, which helps in determining the best way to assemble the fragments into a drug-like ligand.

Machine learning models are being developed both for drug synergy predictions [CongFu, RECOVER] and for de novo generation involving fragments [JT-VAE, SAFE, DiffLinker], i.e. creating novel molecular structures from scratch without relying on pre-existing templates.

1.4 Co-crystallization



Scientists are not only interested in knowing which ligands inhibit a protein and by what amounts, they also want to know how they do it, which involves determining where the ligands bind and in which conformations. Of course, scientists can hypothesize “We think that Pop puts the banana in his mouth”, but this might not necessarily be the case. For example, the banana could be acting on an allosteric site (more on that later).

However, Pop moves constantly, with conformational changes and banana grabbing happening on a microsecond scale, way too quickly to be captured by any accurate 3D visualization technique. To slow down, Pop needs to be purified and crystallized in a solid lattice. Many methods accomplish this, and they require either reducing Pop’s temperature or increasing Pop’s concentration. Indeed, Pop tends to slow down when it is cold or cluttered in a large crowd, but this leads to undesired protein conformational changes that are not necessarily reflective of Pop’s natural state. Typical methods used to study protein conformation include X-ray crystallography, Cryo-Electron Microscopy (Cryo-EM), and Nuclear Magnetic Resonance (NMR) spectroscopy, but there are many more.

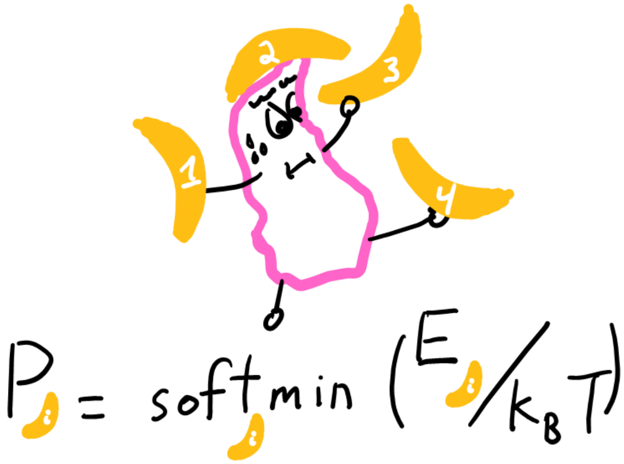
Machine learning methods used the crystallization data from PDB to drive the recent successes from AlphaFold, RosettaFold, ESMFold, etc. And with the CASP16 competition just around the corner, we expect a new breakthrough in co-crystallization, i.e. crystal structures involving multiple components such as 2 proteins or a protein and ligand.

Part 2 - Computational methods

Having found an initial set of ligands that could satisfy Pop, scientists now turn to computational methods to study them and learn how they can build the perfect banana. Or perhaps they’re pretending to work while playing a video game with Pop as the main character?

Why bother going to the lab at all? In this age of computers, we can simulate the ligand/protein interactions or just run an inference model to determine all their binding properties. Or can we? At least, that is the promise of machine learning.

2.1 Free energy



The most accurate way to estimate the binding of a ligand to a protein is to compare its free energy when it is bound to the protein to the free energy it has when hanging out on its own in a solvent like water. These free energies are determined by the energies E_i of each conformation (3D structure) that Pop and banana. Also, the probability P_i of a given state can be expressed as the *softmin* of the current energy over all possible states.

Hence, these computations requires us to know the energy of **every possible state of the protein-ligand pair**, an intractable problem that requires many approximations. Also, the binding affinity is highly correlated with the **difference in free energy** between Pop alone, and Pop with the banana.

Let's show some equations to compute the free energy F_e and the the probability of each state P_i (we promise, it's the only math in this blogpost). In the equations below, k_B is the Boltzmann constant, T is the temperature, and P_i the probability of each state. In both equations, one can see that there is a dependance on the exponential of the negative energy. This means that both the free energy and the probability distribution are dominated by lower energy conformations. Indeed, in machine learning terms, P_i is known as a softmin distribution.

$$F_e = -k_B T \ln \left(\sum_i \exp\left(-E_i / k_B T\right) \right)$$

$$P_i = \frac{\exp\left(-E_i / k_B T\right)}{\sum_i \exp\left(-E_i / k_B T\right)}$$

Pocket sampling is a method in which only the most promising protein pockets are explored. This means that we will place the banana relative to Pop only in regions that are likely to be local minima of energy, i.e., regions where the banana will naturally stick like hands do on top of the head, not on the nose or knees. Energy minima are higher-probability regions due to the *softmin*.

Static proteins are inflexible forms of Pop, or ones with only a few poses. The static approximation means that we don't have to worry about Pop flailing its arms and legs and potentially blocking banana delivery.

Energy approximations will typically be used since quantum solvers have a complexity of at least O(N^3). Hence, methods such as force fields or semi-empirical approximations will not correctly estimate the energy of the Pop and banana interaction, which can yield to large errors in the free energy.

Machine learning can help by improving the sampling methods for both protein conformation and target/ligand binding conformations while also improving the energy estimates. Such methods are known as Boltzmann generators.

2.2 Docking

One of the most famous methods for estimating the binding affinity is Docking. Traditional docking methods focus only on the pocket of interest to rank the ligands by affinity, and most do not use explicit solvents. This has major drawbacks: It forces both the protein and the ligand to be static and uses low-accuracy energy predictions to optimize the conformations. It is equivalent to chaining up Pop and putting a yellow stick in its mouth – if Pop bites, then it means Pop likes bananas.

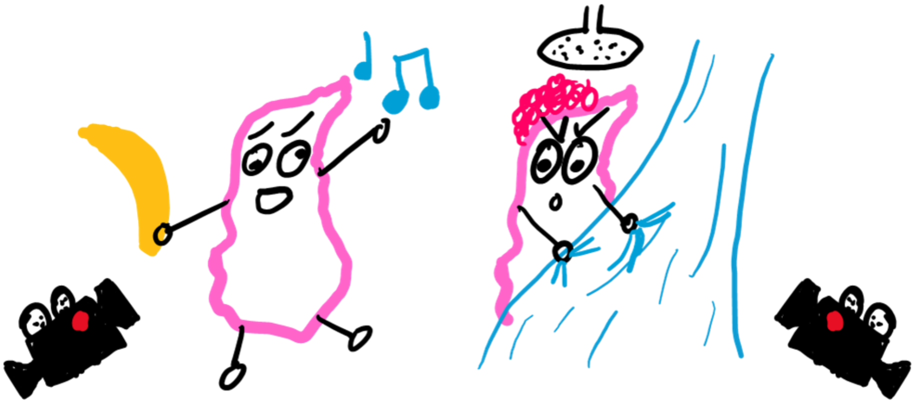
To compensate for their shortcomings, docking methods are often parametrized to experimental data, meaning that their score function and other key parameters are adjusted to match the results of experiments.

Inhibition or activation of the protein's activity is usually desired when designing drugs. Indeed, the drug could be binding in regions that don't affect the protein's activity. For example, when the banana is on Pop's head, Pop can still move freely and achieve its function, but when the banana is in its hands or feet, it hinders Pop's ability to move freely and achieve its function. When using docking, it is therefore important to make sure that the pocket is either an active site or an allosteric site of the protein, with more details in section "4.2 Allosteric modulation".

Docking can give us an interpretable and fast solution to our problems: binding a ligand within a desired pocket. It is usually useful for discarding very bad ligands whose geometries clearly don’t fit the pocket and for ranking ligands similar to those tested experimentally during lead optimization. However, it remains inherently limited since it does not consider the flexibility of the system by reducing the problem to a selected conformation and pocket and by not studying the effect of the binding on the protein's energy landscape.

Machine learning can help by improving the conformational sampling of both proteins and ligands to match experiments more closely [DiffDock], while also significantly boosting the accuracy of the binding affinity prediction [link], but there is still lots of work to be done before these methods become reliably enough to replace traditional docking.

2.3 Molecular Dynamics



A traditional method of computing the free energy is to do lengthy molecular dynamics (MD) simulations, watching the ligands and proteins dance together to study their interactions. This is used to sample as many protein/ligand conformations as possible, and is thus an (expensive) attempt at brute-forcing free energy.

Suppose you want to find out whether Pop likes bananas. One approach to answering this question is to film Pop 24/7 until it finally picks up a banana on the counter and eats it. But what if the banana is not on the counter and is hidden out of sight? It could be a long time before Pop finds and reaches for the banana, and Pop could be doing something completely unrelated in the meantime.

Having little control over the simulation is one of the biggest problems with MD. Although MD simulations make nice movies showing the protein/target interactions, they can diverge to completely undesired or unrealistic states and take days of simulations to show anything interesting.

Biasing the MD simulation is a possible way to overcome the time it takes to observe the desired states. One strategy called metadynamics discourages Pop from using the same dance move twice. Another strategy is replica exchange, which runs the same simulation at different temperatures simultaneously to encourage Pop to try new dance moves. However, these methods require a knowledge of the task and the system to be implemented correctly.

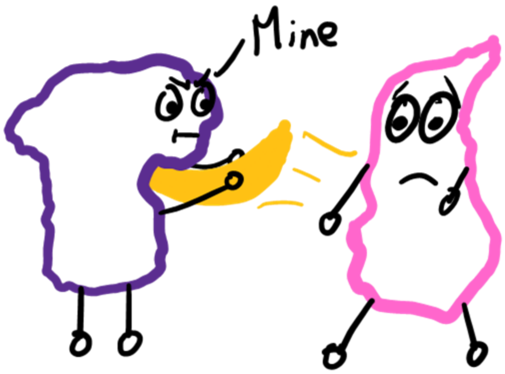
The low precision energy predictions used by molecular dynamics are of course another major issue that was discussed in the “free energy” section. Indeed, in MD simulations, Newtonian mechanics is used to approximate quantum dynamics. Despite Pop’s celebrity status, its movies are made with a cheap and blurry camera – who wants to watch that?

Machine learning methods can help reduce the computation time by taking larger trajectory steps [TimeWarp], compressing the conformations in a latent space [LSS], or increasing the diversity of the initial states. ML can also help improve the accuracy of force fields to be close to the quantum mechanics simulations accuracy, but still within a reasonable compute time [MACE, TorchMDNet, Allegro]. However, these force-fields are trained on systems of <100 atoms, it is unclear how well they generalize to protein/solvent systems of ~10,000 atoms or to unstable conformations.

Part 3: Optimizing the drug

In this third part, Pop’s adventures continue to get more exciting. Scientists have found some good bananas. Now it’s time to make them so delicious that Pop stops causing diseases and spends its whole day eating bananas instead.

3.1 Selectivity & Specificity

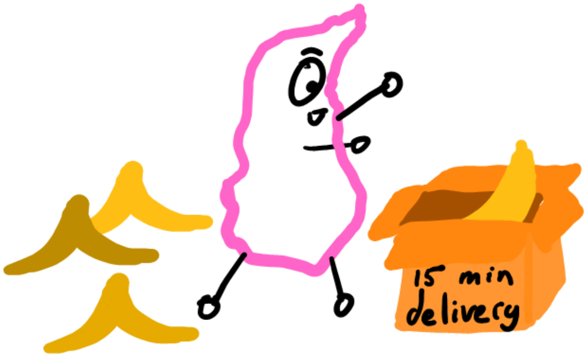


It’s all great if Pop loves bananas, but what if other proteins also love them, or if the presence of the banana disrupts other processes, then they can lead to undesired side-effects.

An ideal banana is developed **specifically** for Pop’s taste and for no one else’s. But this is quite difficult: Pop has many siblings and cousins with similar tastes (structures) but different goals (functions). On top of that, we also need the banana to **select** Pop and not get lost on its way, possibly disrupting other important processes unrelated to the disease. It should be a mutual love story between Pop and the banana.

Machine learning predictions can be used to assess the selectivity and specificity of ligands by virtually screening thousands of potential proteins, although sometimes the data needed to accomplish this is scarce. Furthermore, small changes in protein structures can be challenging for machine learning to handle, potentially reducing their effectiveness at tackling specificity for evolutionarily related proteins.

3.2 DMPK and ADME optimization



Another important aspect of drug development is drug metabolism and pharmacokinetics (DMPK). Roughly, DMPK measures how well a ligand is delivered to a protein, how long it stays in the body, and how it is excreted from the body after it has served its purpose.

Drug Metabolism examines how a drug is biochemically transformed in the body and eliminated, impacting its efficacy and potential toxicity. Maybe the banana gets crushed during delivery, and Pop doesn't like the smushed result. Maybe there's no garbage service that accepts banana peels and Pop's house becomes contaminated with peels as a result?

Pharmacokinetics is the study of a drug's absorption, distribution, and excretion rates, shaping dosing regimens and overall therapeutic effectiveness. It asks the question, "How easy is it to get the banana from the tree all the way to Pop?" For an oral pill, absorption is about how the banana goes through the digestive system and how fast it makes it to the bloodstream. Distribution is the efficiency and speed of the delivery system to do in-house shipping of the banana to Pop into the right organs and cells. Excretion is about the body's ability to eliminate the banana peels after the bananas have performed their function, usually through the kidney or liver.

ADME (absorption, distribution, metabolism, and excretion) is another widely used acronym that is often synonymous with DMPK. It specifically focuses on understanding the journey of the banana through the human body, from the moment it is taken until it is excreted, to understand the drug's behavior within biological systems.

Machine learning multi-objective optimization is a difficult problem, especially in the late discovery stages where we want to optimize tens of properties. Fortunately, GFlowNets are very well-suited for this purpose as they allow one to efficiently explore the Pareto front of optimal characteristics [link]. The main challenges remain in the score function, i.e. a predictive model trained on property predictions. However, if the models that predict DMPK properties are not accurate, the ligands that are suggested by the algorithm might not work in real life. Also, making sure that the proposed ligands can actually be synthesized in a lab is a difficult problem [link].

3.3 Scaffold hopping & decoration

Once we've found that Pop likes bananas, it's time to optimize them, either for binding, drug metabolism and pharmacokinetics (DMPK), or toxicity. However, we don't want to start from scratch; we want to preserve the banana's binding properties.

Scaffold decoration involves keeping the core of the banana and changing only the extremities to make it more appealing to Pop or to make it have a better pharmacokinetic profile. Scaffold decoration is usually intended for property optimizations, for fixing some major issues like solutbility/toxicity, or to study the structure-activity relationship (SAR) of the banana. In our example, scaffold decoration is represented as the bow to the banana which makes Pop more likely to pick it up and bite it without changing the backbone.

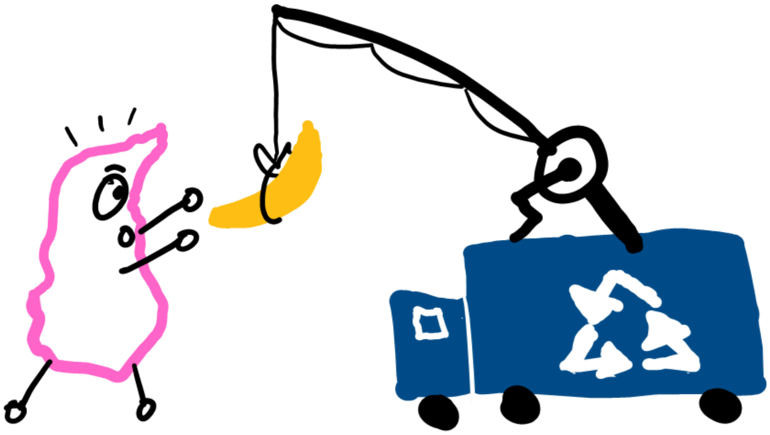
Scaffold hopping is a more radical solution that aims at solving inherent problems with the scaffold or backbone of the ligand, either for physicochemical reasons or for making a patentable ligand. For example, let's take the case of molecular resistance where the cell or the protein learns to evade the drug. If Pop grows suspicious of bananas, we can trick it with a yellow apple which preserve some of the reasons Pop likes bananas (yellow color, sweetness, and fruitiness), but completely changes the nature of the fruit.

Machine learning methods are often difficult to control and constrain, making it hard for ML models to do scaffold hopping and decoration. However, recent models such as SAFE make it much easier for LLMs to think in terms of these operations.

Part 4: Alternative approaches

In most cases, scientists fail to find a banana with a strong enough binding affinity to the active protein pocket, so Pop remains undruggable for years. But scientists are unwilling to give up; they have a few alternative tricks to cure Pop's induced diseases.

4.1 Targeted degradation



Another strategy to improve a drug's efficacy is targeted degradation, with the most popular methods being molecular glues and Proteolysis Targeting Chimeras (PROTACs). While regular drugs work by inhibiting the function of a protein during binding, targeted degradation aims to degrade the protein by increasing the number of encounters between the target protein and the ubiquitin-proteasome, or recycling, system.

In the illustration, we can see how PROTACs work by leveraging the E3 ligase, a protein responsible for degrading proteins or "recycling truck". The PROTAC is equivalent to a small ligand "fishing rod" attached to a long carbon chain or "cord", on which another small ligand "banana" is attached, designed specifically to bring Pop closer to the recycling truck. The same fishing rod can be used to attach different binders that appeal specifically to the protein of interest. Once Pop bites into the banana, it is dragged into the truck and sent for recycling.

Although both PROTACs and molecular glues are designed to recruit an E3 ligase recycling truck, the latter do it in a more indirect manner. Instead, molecular glues such as Thalidomide tag the banana with a beacon so that when Pop grabs it, it signals the recycling truck to come and do the cleaning.

The main advantage of PROTACs is that they don't merely inhibit the protein but, in fact, completely knock out its function. This means that the required concentration of a ligand is lower since binding only needs to occur for a short period of time and that the ligand's effect lasts longer. Even once the ligand is no longer in the system, Pop cannot do any harm until it is re-synthesized at a large enough concentration by the cell.

The main disadvantage is the large size of the ligands, which can negatively impact DMPK properties and make them harder to synthesize.

Machine learning models can be used to predict the protein degradation [DeepPROTAC] or to design the linkers [PROTAC-RL], while there are also some general fragment-based models that can work both for small molecules and larger molecular glues [DiffLinker].

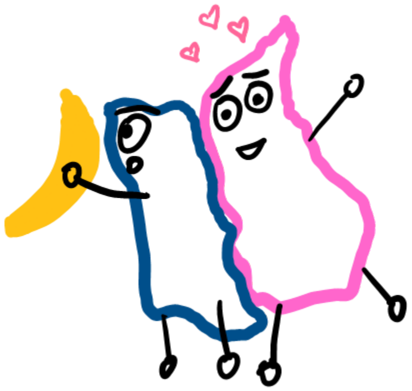
4.2 Allosteric modulation

Most target-based drug discovery focuses on binding a ligand directly in the protein's functional pocket, i.e. the pocket that is directly responsible for the protein's activity (also known as orthosteric binding site). However, an alternative is to bind the ligand on an allosteric site.

Allosteric sites are functional pockets on proteins that, when bound to a ligand, induce a conformational change in the protein's structure, thereby modulating the activity of the protein's active site, either enhancing or inhibiting its function. For example, if Pop is sleeping on a banana couch, it won't be bothering us with disease.

Machine learning models can aid in detecting allosteric sites [link], a typically tedious task, while also improving the targeting of such sites via improved docking and binding affinity prediction.

4.3 Pathway targeting



If you can't tackle Pop, go for its loved ones! Go for Pete! Dirty move you say? Keep the loved one out of this? We say, in the war against Pop's disease, every move is permitted.

When a protein remains undruggable for many years, or even decades, either due to the difficulty of targeting it specifically or delivering the drug, a typical strategy is to target its pathway.

Pathways define the processes in which Pop is involved, one of which leads to the disease that we want to cure. Typical pathways contain many protein-protein interactions, which we referred to earlier as the "loved ones". However, pathway discovery is a complex problem that deserves its own blog post.

Pathway targeting typically involves developing a banana that targets Pete, another protein in the pathway such that it has the same effect as targeting Pop. This is not always possible since proteins can be involved in many pathways, in which case disrupting Pop's loved one could cause major side effects.

Synthetic lethality is another approach in genetic and cancer therapy that exploits the relationship between two genes, where the first is associated to Pop, and the second to Pete, wherein the impairment of either Pop-Pete alone is survivable for the cell, but the simultaneous impairment of both leads to cell death. *It's a microscopic Shakespearean story, for never was a story of more woe than this of Pop, Pete, and their banana …*🥲.

Machine learning can help us drug another target on a given, known pathway. However, most pathways are not known, or are only partially known, making it harder for current ML methods to succeed, although there have been some efforts in inferring edges in pathway graphs [review, PDGrapher].

The End!

Pop has been through numerous adventures, but finally, thanks to relentless efforts and the help of machine learning, we have a banana that Pop likes! We can finally cure a disease previously thought to be undruggable. **And everyone lived happily ever after ♥**

We hope you enjoyed Pop's metaphorical adventures! And that it helped you grasp the most important techniques and challenges in machine learning for drug discovery.