Published on Jun 24, 2025 by Arcadia Science **DOI**: 10.57844/arcadia-ugg5-emyd

A framework for modeling human monogenic diseases by deploying organism selection

We designed a decision-making framework to find tractable genes from our organism selection dataset for pilot experiments. We focused on genes in two potential models of human monogenic disease, the choanoflagellate *Salpingoeca rosetta* and the tunicate *Ciona intestinalis*.

Contributors (A-Z)

Prachee Avasthi, Audrey Bell, Keith Cheveralls, Seemay Chou, Megan L. Hochstrasser, Austin H. Patton, Dennis A. Sun, Ryan York

Version 1 · Jun 23, 2025

Purpose

Drug development requires organismal models to evaluate the efficacy and safety of therapeutic candidates. Most pharmaceutical research uses rodents, assuming they're similar enough to humans to be useful; however, as others have noted **[1][2]**, such models can be expensive, slow, and even inaccurate. Can we unlock new opportunities by studying human diseases in different organisms?

We previously released a computational method to systematically identify similarities between proteins in humans and diverse research organisms by comparing protein secondary structural properties and correcting for phylogenetic relationships **[3]**. We found that phylogenetic distance doesn't always determine modeling utility; the best predicted organisms for a given gene could sometimes be very unexpected. We created the Zoogle interface, hoping this would make it easier for both basic science researchers and drug developers to use our dataset to create disease models. However, users struggled to leverage Zoogle for their own work.

In an effort to improve the usefulness of our predictions for external users, we tried to use Zoogle ourselves to identify actionable organism–gene pairings. We focused on developing a workflow for a particular user type, namely “organism experts” in biology. Such experts have critical, specialized knowledge about the life cycle, phenotypes, experimental tools, and relevant datasets for their organismal model of choice. They’re often part of larger organismal research communities, which helps with troubleshooting and collaborations. To test our workflow, we worked with experts on two organisms with unique biology that are suitable for genetic experiments — a unicellular protist that's closely related to animals, *Salpingoeca rosetta*, and a sea squirt that's closely related to vertebrates, *Ciona intestinalis*.

In this context, we aimed to identify which genes within a given organism might offer the greatest relevance to human biology and disease, helping experts prioritize their experimental efforts. Here, we present a heuristic decision-making approach that combines computational filtering with manual diligence to evaluate gene–disease pairs. We prioritized experimental feasibility and therapeutic impact by evaluating disease mechanisms, protein conservation, available genetic tools, and phenotypic assays. We ultimately identified seven actionable genes in *S. rosetta* and three in *C. intestinalis*. We’re funding two academic labs to pursue experimental testing of our predictions.

- **Data** from this pub is available on Zenodo.
- All associated **code** is available in this GitHub repository.
- Check out **companion pubs** documenting how we chose the most intriguing genes to pursue in *Salpingoeca rosetta* **[4]** and *Ciona intestinalis* **[5]**.

The problem

Through the Zoogle interface, we present a list of matches between the proteins in an organism's proteome and the proteins in the human proteome. Each match represents a hypothesis about the utility of modeling a human protein's function using the homologous protein in a non-human organism. These matches are ranked based on how unexpected their similarity is with respect to the phylogenetic distance between the human and non-human proteins.

What this means on a practical level is that Zoogle presents a ranked list of tens of thousands of predictions of genetic similarity to scientific users. For *Salpingoeca rosetta*, Zoogle catalogs 27,354 predictions; for *Ciona robusta*, there are 50,693. How can a scientist determine which, among these thousands of predictions, represents the most actionable set for experimental testing? We combined computational filtering with manual diligence into an overarching framework for winnowing these predictions to an actionable short list for downstream experiments.

Challenges of choosing a useful disease model

When considering what it means to model a human disease, there are many different strategies **[1]**. The most technically accurate model for human diseases would be humans. However, due to obvious ethical and safety considerations, this isn't the preferred starting point for drug development.

In practice, all drug development relies on disease models. The most common approaches to disease modeling are:

- Using *in vitro* cell culture of human cells, cell lines, tissues, or organoids
- Using non-human models, usually rodents, to approximate human disease pathology

Borrowing a common saying from statistics, we'd argue that *all models are wrong, but some are useful* – each strategy has its pitfalls. *In vitro* cell culture models often use immortalized cell lines with abnormal karyotypes **[6]**. Patient-derived primary cells have genetic and environmental variability and are expensive to acquire and maintain **[1]**. Organoids require long experimental timelines, while not fully capturing the

complexity of real human tissues **[7]**. Non-human models have fundamental differences at the molecular level — human and mouse proteins aren't identical and can have drastically different properties, which can lead to costly failures to translate **[2][8]**.

The inaccuracies and inefficiencies of existing models have been recognized by authorities such as the FDA, which recently announced a plan to phase out animal testing requirements for monoclonal antibody therapies.

Our strategy of using unconventional organisms doesn't overcome concerns with using non-human models. Sequence and structural differences between human and model proteins remain relevant. We account for some of these differences by identifying those pairs of non-human and human proteins with unusually high similarity **[3]**. But the ultimate goal of modeling diseases using such organisms isn't to eliminate the use of human cell or mouse models; rather, it's to complement them.

Our thinking

A phenotype-first experimental framework for modeling human disease

Our hypothesis is that experimentally tractable and more scalable model organisms, such as invertebrates and unicellular eukaryotes, are advantageous and underutilized tools at the earliest stages of therapeutic research and development. Some of these organisms may be more accurate biological models for a particular human disease than rodents are. Others might be comparable to existing models, and also have experimental advantages that complement rodents or in vitro studies, such as tissue-level testing opportunities or cheaper, higher-throughput ways to conduct early screens. Moreover, expanding the list of organisms with the potential for disease modeling provides more avenues for basic science to have translational impact.

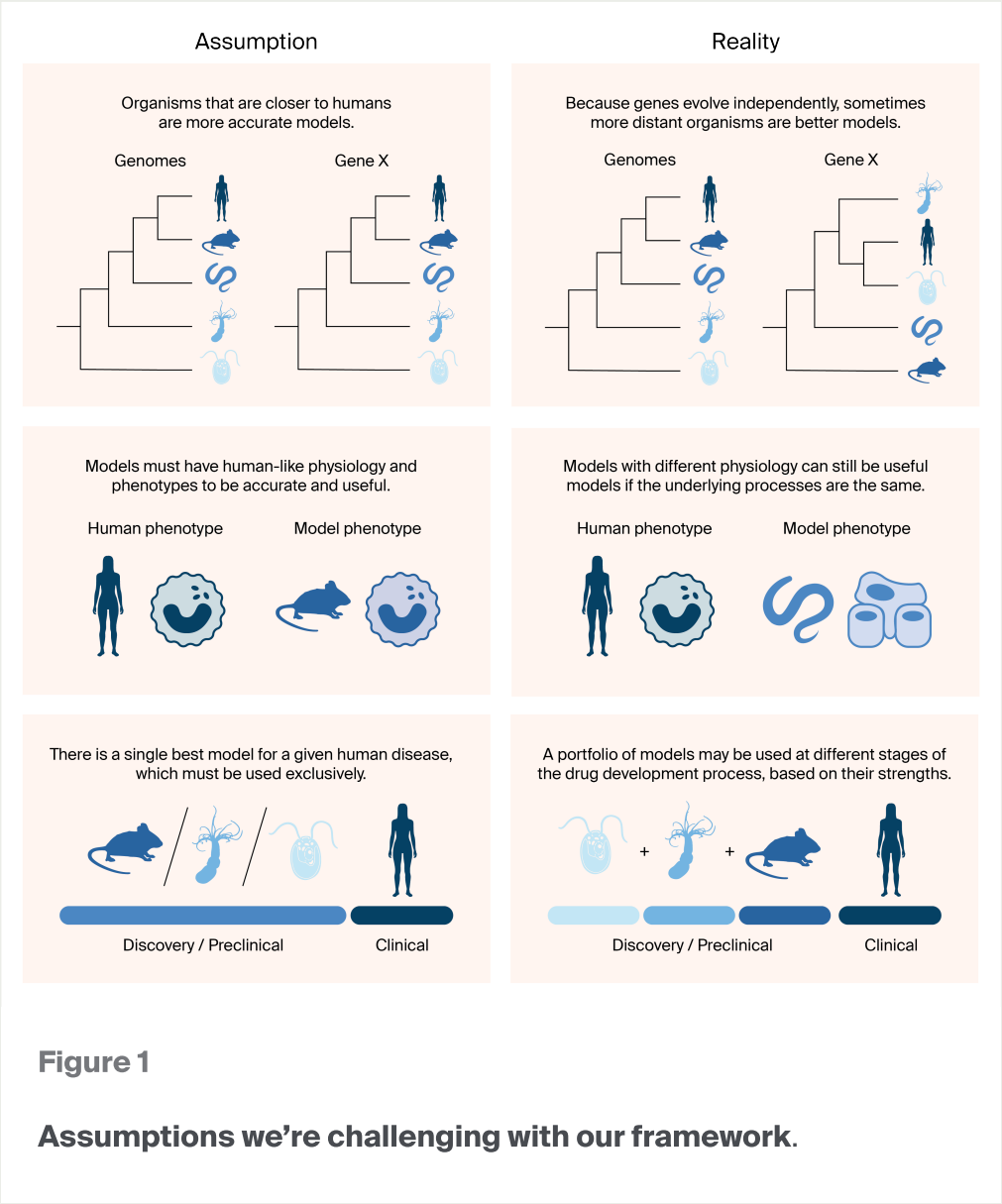

Figure 1

Assumptions we're challenging with our framework.

We aim to challenge three major assumptions about non-human models (Figure 1):

1. **Distant organisms can't be useful models**
   It's often assumed that the further an organism is from humans on the evolutionary tree, the less relevant it is as a model. But genes don't evolve in lockstep with species. They can evolve independently and sometimes even converge on similar functions in distantly related organisms.

   That means a gene in algae might actually behave more like the human version than the same gene in a mouse. Relying on evolutionary proximity alone — like defaulting to mammals — can lead to poor model choices. Our earlier work on organism selection **[3]** shows how this assumption can mislead and how gene-by-

gene thinking opens up better options.

2. **The model must mimic the human disease phenotype**
   Another widespread belief is that a good model must replicate the same physical symptoms or cell behaviors seen in human disease. For example, if a mutation disrupts blood cell migration in humans, researchers expect to see that same defect in the model organism.

   But biology doesn't always work that way. A mutation in a cytoskeletal gene might affect blood cells in humans but lead to a different, yet mechanistically related, problem in another species — like cell intercalation defects in a nematode. Even if the symptoms differ, the root cause (a cytoskeletal failure) may be the same. Studying and rescuing the phenotype in the model can still yield insights and therapeutic entry points for the human condition.

3. **There is a single "best" model for every disease**
   Researchers often try to identify one ideal organism that captures all aspects of a disease. But no model is perfect. Different organisms bring different strengths, and choosing a combination tailored to each stage of research might be more effective. For example, early discovery work might benefit from simple, fast-growing organisms, whereas later work might benefit from more human-like physiology.

   Rather than framing the problem as "which organism should I use instead of a mouse?" we ask scientists to consider "which organisms, used together, could increase progress towards curing human disease?"

What we're ultimately interested in identifying are genes with the potential to be modeled advantageously in our organisms of interest, where we can identify a measurable phenotype to test therapeutic mechanisms of action. This leads to a simple overall experimental framework (Figure 2, right):

1. Identify tractable candidate genes for modeling in a non-human organism. The process of winnowing genes into a short list is detailed in the current pub.
2. Generate analogous mutants in non-human models.
3. Identify measurable phenotypic consequences of the mutation.

4. Use the models to screen for molecules that rescue the measurable phenotype.

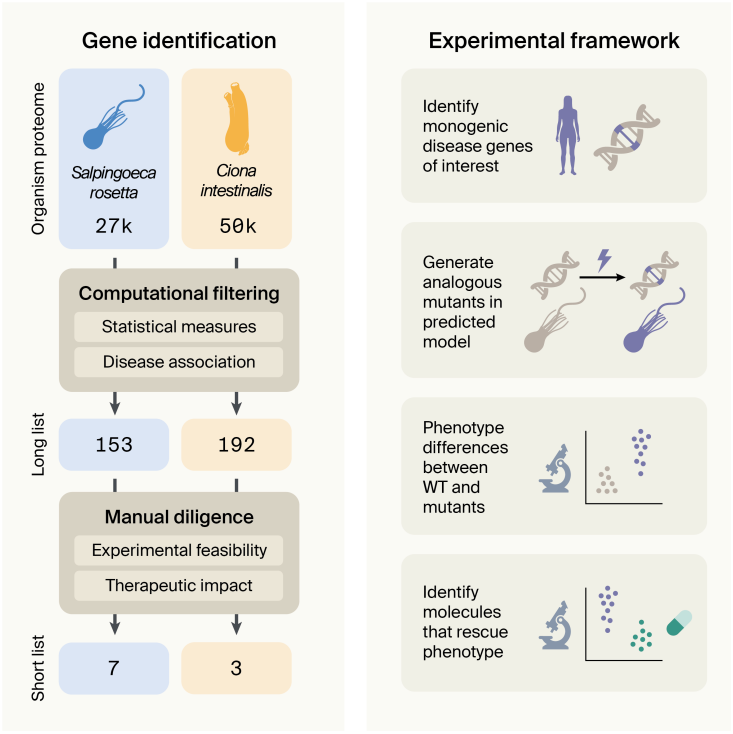

Figure 2

Graphical abstract illustrating the overall framework.

The 27k and 50k proteins originate from the number of proteins found in the final organism selection dataset for each organism.

While this experimental plan is fairly straightforward, choosing which of the thousands of candidate genes to pursue within a given organismal model is far more opaque. We spent a lot of time investigating tractable candidate genes for two example organisms, *Salpingoeca rosetta* and *Ciona intestinalis* (check out specific findings in individual pubs about each in **[4]** and **[5]**), and describe our overall approach (Figure 2, left) in the rest of this pub.

Choosing the right experts

To understand how organism experts might approach designing experiments using Zoogle, we needed feedback from external scientists. We interviewed a handful of experts in our personal networks whose organisms are included in Zoogle to understand:

- Whether they were aligned with our overall mission

- How willing they were to experiment with new approaches to scientific publishing, to be able to iterate in public without needing to cater to journal expectations
- Whether they could give us helpful feedback to improve organism selection

We decided to work with experts in two academic research laboratories: David Booth's laboratory at UCSF, which uses *S. rosetta*, and Alberto Stolfi's laboratory at the Georgia Institute of Technology, which uses *Ciona*. These experts provided invaluable feedback during our diligence process.

Our approach

Skip to "Methods" for nitty-gritty details, or read on to get a big-picture sense of how we tried to select the most useful and feasible disease-associated genes to study in two uncommon organismal models.

Computational filtering

Predictions within the organism selection dataset in Zoogle are currently ranked based on the percentile of the phylogenetically-corrected structural distance of proteins within gene families. This is essentially the relative ranking of each protein compared to others in the same gene family. While this metric was easy to implement into a web interface, it doesn't account for variability among gene families in their size and distribution of distances from human homologs.

To account for these differences, we included two new metrics aimed at quantifying whether each distance to human homologs was exceptionally similar or not, given the observed distribution of distances in each gene family. Specifically, we used a permutation test-based approach to calculate two *p*-values: one "within organism," and one "across organisms." The within-organism *p*-value indicates, for each gene family, whether the degree of similarity with respect to the human homolog is exceptional for a given species. In contrast, the across-organism *p*-value indicates, for each gene family, which species are exceptionally similar to humans.

For a detailed description of how we carried out these analyses, see "Methods."

Access our **code for generating these *p*-values** in an updated version of our organism selection GitHub repo (DOI: 10.5281/zenodo.15693939).

Access the updated organism selection **dataset with *p*-values** at this Zenodo deposition (DOI: 10.5281/zenodo.15685124).

We then filtered our genes for each organism using the following steps (as illustrated in Figure 3):

1. ***P*-value filtering**. We filtered genes across gene families based on their "across-organism" *p*-value.
2. **General disease association**. We used the list of human gene–disease associations found in the ClinVar database to quickly remove genes with no recorded disease association.
3. **Homolog count**. For our pilot experiments, we sought genes that are very likely to produce a phenotype through a single knockout. As such, we removed genes with multiple predicted copies in *Salpingoeca rosetta* or *Ciona intestinalis*.
4. **Single disease association**. To identify genes with simple mechanisms and further decrease the number of genes we needed to diligence, we removed genes associated with more than one ClinVar disease.

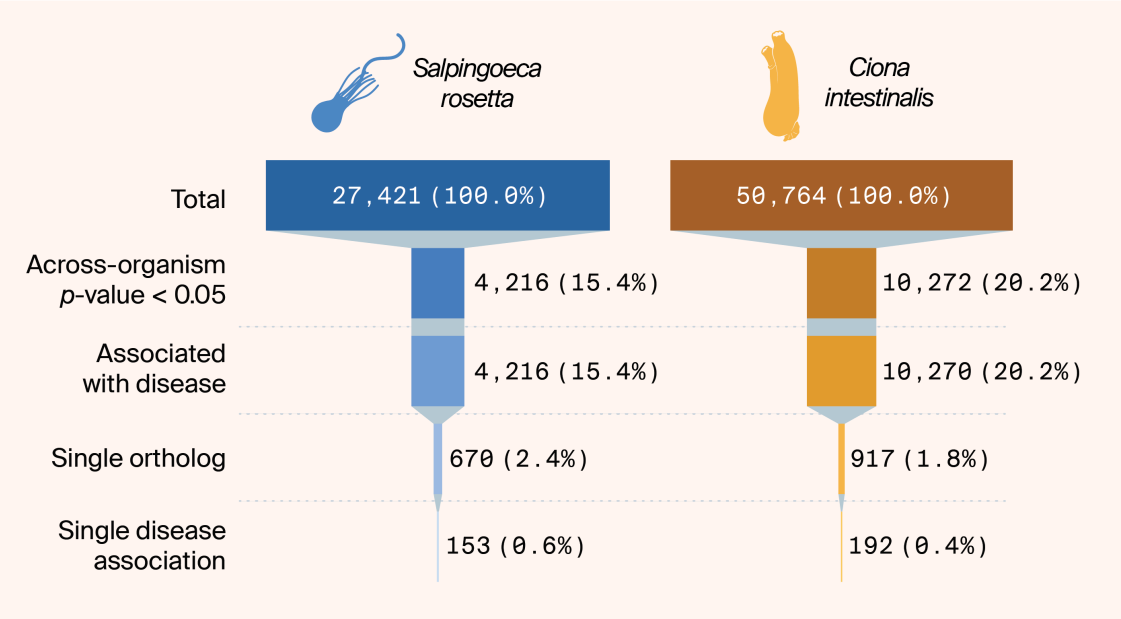

Figure 3

Funnel chart illustrating the stages of the computational filtering pipeline and the corresponding number of predictions remaining after each filter.

This set of crude filters is an initial prototype, and we recognize there are many ways to improve upon our approach. Our primary goal in the filtering process was to decrease the number of genes we needed to manually diligence. At the end of this filtering process, we were left with a "long list" of 153 genes in *Salpingoeca rosetta* and 192 genes in *Ciona intestinalis*.

> Access our filtering pipeline notebook on GitHub. This pipeline also generates hyperlinks to external resources for filtered genes, such as OMIM, OpenTargets, and MARRVEL. We added this functionality because we found it useful in our downstream manual diligence process.

Manual diligence

From our long list, we performed manual diligence to evaluate how actionable each possible gene would be for downstream experiments. We considered two high-level questions during our process:

- **Experimental feasibility**. Is it easy to make a genetic model of the disease in this organism?
- **Therapeutic impact**. Would making a model for the disease in this organism be useful? For example, could we more rapidly investigate therapeutic mechanisms of action in this system as opposed to standard models?

We didn't pursue comprehensive diligence for each hypothesis in our long list; rather, we assessed each individual gene until the first point of failure — that is, as soon as we determined that making a model wouldn't be easy or useful. We also didn't perform manual diligence on every single member of our long list, as this is time-consuming. Instead, we diligenced ~30–40 genes from each organism, starting with those with a low percentile score. We added a handful of others based on the research interests of two research groups we're funding to experimentally test our Zoogle predictions.

For each of our high-level questions, we cataloged a number of failure modes based on our technical analyses, listed below. The details of the technical analyses are described in the Methods section.

A schematic diagram of the desired qualities of a candidate gene and areas of consideration can be found in Figure 4.

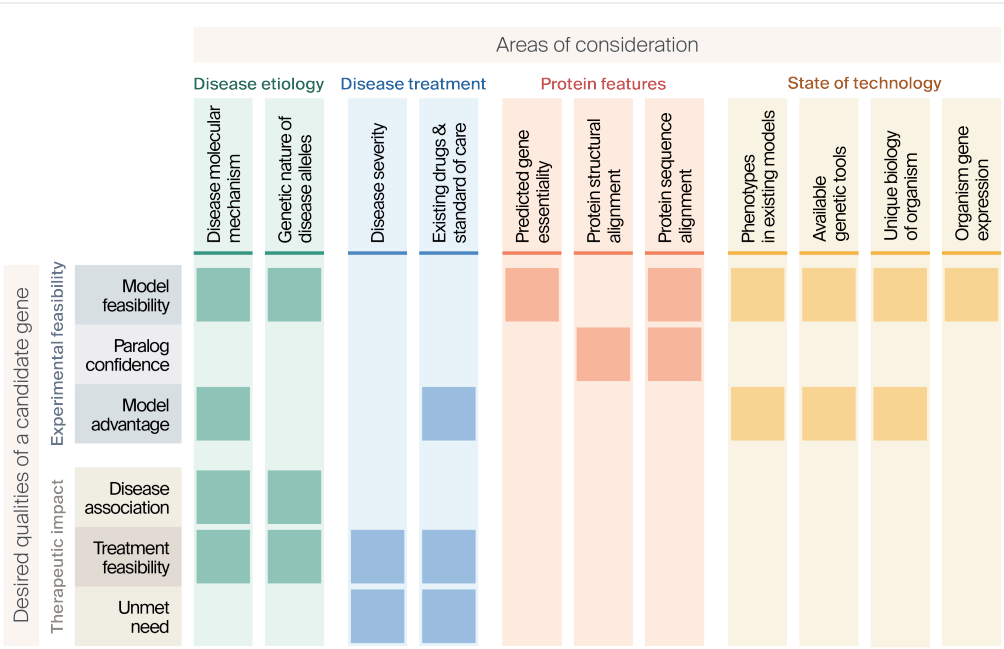

Figure 4

Guide showing areas to consider in diligencing essential qualities of a candidate gene.

Rows are areas of evaluation and columns are technical analyses that we took into account to perform the evaluation.

Experimental feasibility

When considering experimental feasibility, we encountered the following failure modes. These criteria aren't necessarily dealbreakers for the overall utility of models — rather, they helped us eliminate options that weren't low-hanging fruit.

- **Model not feasible**. We evaluated the technical feasibility of making an analogous mutation to the causative disease allele in humans. We accounted for the precise molecular mechanism and genetic variation in the human disease, the state of technology in each organism, and the essentiality of the gene. In general, we strove to identify diseases where we could generate an analogous mutation to a human disease allele in the organism. For example, if a human gene had a nonsense or frameshift mutation leading to loss of function, we'd want to be able to introduce a nonsense or frameshift mutation in our organism. For some genes, we generated protein sequence alignments between the human and non-human proteins to look

for conservation of disease-causing residues. In some cases, we had evidence from literature that missense mutations could induce loss of protein expression; for those genes, induced nonsense mutations could be appropriate analogous mutations to generate.

Because the state of technology for generating genetic mutations differs between *Salpingoeca rosetta* and *Ciona intestinalis* (our two test organisms), our approach to evaluating feasibility also differed by organism. We describe the differences in the pubs linked in the "Organism narratives" section.

- **Lack of homolog confidence**. We assessed our level of confidence in precisely matching a non-human gene to its human paralog. We used structure-based Foldseek searches of the non-human protein across diverse proteomes as a crude measure of confidence. In one case, we also used structure-based clustering of proteins using ProteinCartography to understand more precise differences across a large gene family. In a small handful of cases, we weren't able to confidently determine which human paralog was the best match for the non-human protein. We avoided pursuing those ambiguous matches.

- **Lack of model advantage**. We considered whether building a model in each organism might be advantageous or whether an *in vitro* approach would be superior. We accounted for the molecular mechanism of the disease and the unique features of the organism's biology. For mutations in some metabolic enzymes, generating correctors for those proteins through an *in vitro* approach would likely be a superior approach to a non-human cellular or organismal model.

Therapeutic impact

When considering therapeutic impact, we encountered the following failure modes:

- **Disease association concerns**. We evaluated our level of confidence that the human gene causes its annotated disease. We relied on summaries from OMIM to make this assessment. For some diseases annotated in ClinVar **[9]**, there's a lack of clear genetic evidence that the disease is related to the given allele. We decided not to pursue those genes.

- **Treatment not possible**. We considered whether it would be possible to treat the disease in an actual patient. We used summaries from OMIM and a review of disease literature to make this assessment. For some diseases, the effects of a mutation manifest during embryonic development, resulting in morphological abnormalities that can't be easily corrected after birth. Given that our ultimate goal is to identify drug candidates for these diseases, we usually decided not to pursue these genes. In some cases, we were able to identify a compelling hypothesis for a possible phenotype, which we believed could be useful as a positive control — we didn't reject genes for this reason in those cases.

- **Lack of unmet need**. We evaluated whether there was a substantial unmet need for a given disease to be treated by considering both the severity of the disease and existing treatments. We used summaries from OMIM and other literature searches to make this assessment. In some cases, the disease in question didn't have a meaningful impact on patient lifespan. For example, Meier–Gorlin syndrome (associated with CDC45 mutations) results in patients of shorter stature but otherwise normal life expectancy and mostly normal health **[10]**. We decided not to pursue models for these genes. In some cases, a trivial mechanism is already used to treat the disease; for example, for congenital defect of folate absorption (e.g., caused by defects in SLC46A1), dietary supplementation of folate is sufficient to treat the disease **[11]**. We decided not to pursue models for such genes.

Other considerations

Notably, we didn't consider whether there were substantial market opportunities to treat a given disease (either due to market size, incidence, or degree of unmet need). A challenge for drug development in rare diseases is that economic forces make it difficult to justify investing the high capital cost of drug development for a small number of patients. For our proof-of-concept experiments, focusing on the financial upside would have been prohibitively limiting. Our hope is that our framework can help match academic researchers focused on specific model organisms with rare disease communities to spur transformative research without having to worry about turning a profit.

To get a sense of how it looks to do this process of elimination, check out copies of the working documents we used to catalog our thoughts for each organism:

- *Salpingoeca rosetta*
- *Ciona intestinalis*

Methods

Below are the technical analyses we performed as part of this work.

Conservation with humans

We originally quantified the degree of molecular conservation between non-human gene copies and their human homologs within each of the 9,260 gene families containing humans assessed in our recent pub **[3]** (see "The approach" for a detailed description of methods). Here, we extended this approach, statistically quantifying our confidence in asserting that measured distances were exceptional, whether looking within species and across gene families ("within organism"), or within gene family and across species ("across organism").

We calculated the within-organism *p*-value by permuting the distances from human homologs observed within *each species* and *across gene families* 10,000 times, determining the number of times a distance was smaller than observed. We calculated the across-species *p*-value by permuting the distances within *gene family* and across *species* 10,000 times, with the *p*-value corresponding to the probability of observing a distance smaller than observed. We carried out all analyses in R **[12]**, with the permutation tests implemented as custom Rcpp scripts (found here), and called by the `dist_permute_test` function implemented here.

Disease molecular mechanism

We manually reviewed the existing literature summaries on the known function of the wild-type protein and the consequences of mutation found in OMIM, MARRVEL **[13]**, and UniProt **[14]**. For genes we were interested in modeling, we dove deeper by reading the primary literature for each disease.

Nature of human variants

We manually reviewed the existing literature summaries on human genetic variation found in OMIM's case studies and in MARRVEL. We used the case studies highlighted in OMIM's "Allelic Variants" sections to understand the mechanistic underpinnings of mutations that can contribute to disease. We also reviewed the ClinVar **[9]**, Geno2MP, and gnomAD **[15]** data compiled in MARRVEL to understand the broader scope of human variation.

Predicted gene essentiality

We used the literature summaries from OMIM and the loss-of-function observed/expected (LoF o/e) score and lethality evaluation from MARRVEL's gnomAD module to evaluate whether a gene is likely to be lethal upon knockout.

Disease severity

We used the disease-specific literature summaries from OMIM, disease summaries from MedlinePlus, disease descriptions in Orphanet, and descriptions of patient phenotypes and experience from patient advocacy group websites, if available, to understand the severity of diseases.

Existing models

We used the literature summaries from OMIM focused on organismal models (usually found in the “Animal Model” section) to understand whether phenotypic information has been generated for organisms such as mice and zebrafish. In some cases, we also checked whether a mouse mutant exists in the Mouse Genome Informatics (MGI) database or used web searches to look for literature not cataloged in OMIM on existing mouse or zebrafish models.

Existing drugs

We used the “Approved Drugs and Active Ligands from PHAROS” info in MARRVEL and the protein-specific pages in OpenTargets, as well as treatment information from OMIM, Orphanet, and MedlinePlus to understand the current state of therapeutics for each disease.

Paralog matching

We used pre-folded structures from AlphaFold **[16]**, retrieved via UniProt ID, to search proteomes within the Foldseek Search server. We ran Foldseek searches in 3Di/AA mode against the AFDB-Proteome and AFDB-SwissProt databases with a taxonomic filter for human proteins. We evaluated whether the top human hit to the non-human protein matched the predictions in our organism selection pipeline. We also checked for differences in overall protein structure, such as new or differently sized domains. For proteins of very large size or with multiple domains, we didn’t consider a low TM-score to be disqualifying, as TM-score relies on static structural alignment, which doesn't account for the possibility of flexible domains.

Protein folding

For one pair of proteins — human VWA8 (UniProt ID A3KMH1) and its *Ciona* homolog (UniProt ID F6QXZ7), which were too large to be folded by AlphaFold (> 1,500 aa) and

therefore not included in publicly available datasets, we used ESMFold **[17]** to generate predicted structures. You can download these structures below.

pdb `A3KMH1.pdb` Download

Download an ESM-predicted structure for *Homo sapiens* VWA8.

pdb `F6QXZ7.pdb` Download

Download an ESM-predicted structure for *Ciona intestinalis* Vwa8.

Structure alignment

For visualization figures, we used the Pairwise Structural Alignment tool found in the RCSB databank **[18]**. We generated structural alignments using either TM-align **[19]** or, for large proteins where fixed-body alignment was likely to fail, using jFATCAT **[20]**.

Sequence alignment

We used ClustalOmega **[21]** provided by the UniProt web interface to perform pairwise sequence alignment, with default settings.

The zoogletools package

We created a lightweight, locally installable Python package called zoogletools to organize the code we used in our computational analyses, which is part of our GitHub repository. This package contains scripts for our filtering pipeline, as well as scripts for visualizing gene expression from *S. rosetta* and *Ciona* resources described in our companion pubs.

All of our **code** for this pub, including the zoogletools package, is in this GitHub repo (DOI: 10.5281/zenodo.15724881).

State of technology

As part of our evaluation process, we evaluated the state of genetic tools in each organism by reviewing existing literature and consulting with experts. This helped us evaluate what kinds of experiments would be easy to perform without substantial technical development. You can read more about the state of technology for each organism in the corresponding pubs.

AI tool usage

We used ChatGPT to help write code and comment our code. We used Claude to help write code, clean up code, comment our code, suggest wording ideas that informed our phrasing choices, write text that we edited, expand on summary text that we provided, and clarify and streamline our text. We also used Cursor to help generate and revise code.

Visualization

We used plotly (v5.17.0) **[22]** arcadia-pycolor (v0.6.3) **[23]** to generate figures before manual adjustment.

Organism narratives

You can read more about the specific takeaways from each organism in two companion pubs.

Modeling human monogenic diseases using...

- the choanoflagellate *Salpingoeca rosetta* **[4]**
- the tunicate *Ciona intestinalis* **[5]**

Key takeaways

In this pub, we present a framework for leveraging organism selection to identify which monogenic diseases you might effectively model with your favorite research organism. This framework isn't a computational pipeline; rather, it's a recorded example of the types of scientific reasoning that scientists do almost every day.

As we developed our framework, we realized that this work was more challenging and time-consuming than we expected, particularly as novices in working with these organisms. The predictions presented in Zoogle were a useful starting point, but we needed to gather a lot of additional information about diseases and the technologies in each organism to develop an actionable experimental plan. In some cases, we had to onboard to community resources or integrate expert opinions on the most practical ways to test our predictions. Access to such implicit and explicit knowledge was essential.

We've sometimes described this work as a miniature version of a qualifying exam, and hope that sharing our framework will help others identify new ways to deploy their favorite research organisms for broader impact on human health. Below, we summarize some of the lessons we learned from this exercise.

Importance of expert knowledge

When we performed our initial reasoning, we relied on existing literature and publicly available resources to help us understand the state of technology in each organism. Relying on existing literature sometimes failed to give us a clear picture of what experiments were trusted in the field; in other cases, it completely misled us. For example, a single report of targeted genetic engineering in the literature may not reflect the likelihood of success, applicability to other examples, or represent a

dependable protocol. Speaking with experts cleared this up quickly. It also helped us understand what data resources were most useful to integrate into our analyses and how to navigate the bespoke datasets that are common in emerging research organisms. When attempting to design experiments in unfamiliar organisms, human experts remain irreplaceable.

We recommend that organism experts who follow our framework carefully consider the unique strengths and limitations of their organism of interest when evaluating which diseases to model. What opportunities are uniquely unlocked by the biology of your organism? And how does your organism provide an advantage over the status quo?

Some choices were counterintuitive

Our final gene lists contained some intuitive examples — for example, modeling the function of a stereocilia gene in a *S. rosetta*, an organism with a stereocilium, might appear sensible and even obvious. In other cases, our reasoning arrived at highly counterintuitive results. For example, all three of the genes we chose in *Ciona* are implicated in immunodeficiencies, yet all three might be well-modeled through a completely different cellular process: notochord lumen morphogenesis.

It's important to note that we didn't set out to search for either intuitive or counterintuitive examples when performing our reasoning exercise. Starting with data, we reasoned through the options through a practical lens to arrive at our final short list. It was heartening to see that taking a data-driven approach to choosing scientific questions can lead to surprising and exciting new research directions.

Next steps

Our most important next step is to evaluate whether our predictions or decision-making steps led to actionable outcomes. We'll pursue this by funding organism experts to perform experiments based on our predictions. Results from this work will be published openly through modular units (see below). Assuming that our framework proves useful, there are a variety of ways we could imagine improving it.

Potential improvements

The framework presented in this pub is a prototype with many possible areas for improvement, including:

- **Systematic approaches for understanding the state of technology**. We manually determined the state of technology in each organism through literature review and discussions with experts. Building a centralized database of this information across organisms could substantially accelerate diligence and make it easier for researchers to design experiments across systems they aren't familiar with.

- **Improved filtering**. Our current filtering framework uses a variety of simple heuristics and statistical tests from the organism selection pipeline. To improve our filtering approach, it could be useful to curate a positive and negative control dataset — such as by collating data on the accuracy of existing disease models in mice and zebrafish — to help us determine filtering cutoffs and approaches more empirically.

- **Automating manual technical analyses**. We ran a variety of technical analyses, such as structure-based searches, sequence alignment, and review of results from technical analyses in OMIM and MARRVEL. Many of these steps could be automated by building snakemake or Nextflow workflows and accessing the APIs of such resources to retrieve data, rather than reviewing data manually.

- **Developing numerical heuristics**. The current framework relies on human judgment and reasoning to determine which disease–gene pairs are most actionable. To increase efficiency and consistency in our evaluations, we could develop numerical heuristics — for example, developing a "feasibility score" on a scale from 1 to 5, where different aspects of the state of technology in an organism are given a numerical value. This could allow for more systematic and uniform evaluation of disease–gene pairs.

- **Leveraging machine intelligence**. State-of-the-art language models appear to possess powerful evaluation capabilities. Presenting our framework for evaluation alongside underlying data to such systems could allow for flexible and efficient reasoning through possible disease–gene pairs at scale.

Stay tuned

This pub and its organism-specific companion pieces **[4][5]** are just the start of a longer series of experiments. Stay tuned to learn more about the results of our testing, which the two labs we're funding for this work will publish through an open, journal-independent approach.

Acknowledgments

We'd like to thank Alberto Stolfi and David Booth for their feedback on our diligence process and the final list of candidates, which they'll pursue experimentally with funding from Arcadia.

References

1. Loewa A, Feng JJ, Hedtrich S. (2023). Human disease models in drug development. https://doi.org/10.1038/s44222-023-00063-3
2. Frangogiannis NG. (2022). Why animal model studies are lost in translation. https://doi.org/10.20517/jca.2022.10
3. Avasthi P, McGeever E, Patton AH, York R. (2024). Leveraging evolution to identify novel organismal models of human biology. https://doi.org/10.57844/ARCADIA-33B4-4DC5
4. 10.57844/arcadia-bp0f-v1xx
5. 10.57844/arcadia-084m-a3v2
6. Voloshin N, Tyurin-Kuzmin P, Karagyaur M, Akopyan Z, Kulebyakin K. (2023). Practical Use of Immortalized Cells in Medicine: Current Advances and Future Perspectives. https://doi.org/10.3390/ijms241612716

7 Doss MX, Sachinidis A. (2019). Current Challenges of iPSC-Based Disease Modeling and Therapeutic Implications. https://doi.org/10.3390/cells8050403

8 Ferreira GS, Veening-Griffioen DH, Boon WPC, Moors EHM, van Meer PJK. (2020). Levelling the Translational Gap for Animal to Human Efficacy Data. https://doi.org/10.3390/ani10071199

9 Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. (2013). ClinVar: public archive of relationships among sequence variation and human phenotype. https://doi.org/10.1093/nar/gkt1113

10 de Munnik SA, Hoefsloot EH, Roukema J, Schoots J, Knoers NV, Brunner HG, Jackson AP, Bongers EM. (2015). Meier-Gorlin syndrome. https://doi.org/10.1186/s13023-015-0322-x

11 Goldman DI. (2024). Hereditary folate malabsorption. https://www.ncbi.nlm.nih.gov/books/NBK1673/

12 R Core Team. (2022). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

13 Wang J, Al-Ouran R, Hu Y, Kim S-Y, Wan Y-W, Wangler MF, Yamamoto S, Chao H-T, Comjean A, Mohr SE, Perrimon N, Liu Z, Bellen HJ, Adams CJ, Adams DR, Alejandro ME, Allard P, Ashley EA, Azamian MS, Bacino CA, Balasubramanyam A, Barseghyan H, Beggs AH, Bellen HJ, Bernstein JA, Bican A, Bick DP, Birch CL, Boone BE, Briere LC, Brown DM, Brush M, Burke EA, Burrage LC, Chao KR, Clark GD, Cogan JD, Cooper CM, Craigen WJ, Davids M, Dayal JG, Dell'Angelica EC, Dhar SU, Dipple KM, Donnell-Fink LA, Dorrani N, Dorset DC, Draper DD, Dries AM, Eckstein DJ, Emrick LT, Eng CM, Esteves C, Estwick T, Fisher PG, Frisby TS, Frost K, Gahl WA, Gartner V, Godfrey RA, Goheen M, Golas GA, Goldstein DB, Gordon MG, Gould SE, Gourdine J-PF, Graham BH, Groden CA, Gropman AL, Hackbarth ME, Haendel M, Hamid R, Hanchard NA, Handley LH, Hardee I, Herzog MR, Holm IA, Howerton EM, Jacob HJ, Jain M, Jiang Y, Johnston JM, Jones AL, Koehler AE, Koeller DM, Kohane IS, Kohler JN, Krasnewich DM, Krieg EL, Krier JB, Kyle JE, Lalani SR, Latham L, Latour YL, Lau CC, Lazar J, Lee BH, Lee H, Lee PR, Levy SE, Levy DJ, Lewis RA, Liebendorfer AP, Lincoln SA, Loomis CR, Loscalzo J, Maas RL, Macnamara EF, MacRae CA, Maduro VV, Malicdan MCV, Mamounas LA, Manolio TA, Markello TC, Mazur P, McCarty AJ, McConkie-Rosell A, McCray AT, Metz TO, Might M, Moretti PM, Mulvihill JJ, Murphy JL, Muzny DM, Nehrebecky ME, Nelson SF, Newberry JS, Newman JH, Nicholas SK, Novacic D, Orange JS, Pallais JC, Palmer CGS, Papp JC, Pena LDM, Phillips JA III, Posey JE, Postlethwait JH, Potocki L, Pusey BN, Ramoni RB, Robertson AK, Rodan LH, Rosenfeld JA, Sadozai S, Schaffer KE, Schoch K, Schroeder MC, Scott DA, Sharma P, Shashi V, Silverman EK, Sinsheimer JS, Soldatos AG,

Spillmann RC, Splinter K, Stoler JM, Stong N, Strong KA, Sullivan JA, Sweetser DA, Thomas SP, Tifft CJ, Tolman NJ, Toro C, Tran AA, Valivullah ZM, Vilain E, Waggott DM, Wahl CE, Walley NM, Walsh CA, Wangler MF, Warburton M, Ward PA, Waters KM, Webb-Robertson B-JM, Weech AA, Westerfield M, Wheeler MT, Wise AL, Wolfe LA, Worthey EA, Yamamoto S, Yang Y, Yu G, Zornio PA. (2017). MARRVEL: Integration of Human and Model Organism Genetic Resources to Facilitate Functional Annotation of the Human Genome. https://doi.org/10.1016/j.ajhg.2017.04.010

14 Consortium TU. (n.d.). UniProt: the Universal Protein Knowledgebase in 2025. https://doi.org/10.1093/nar/gkae1010

15 Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Aguilar Salinas CA, Ahmad T, Albert CM, Ardissino D, Atzmon G, Barnard J, Beaugerie L, Benjamin EJ, Boehnke M, Bonnycastle LL, Bottinger EP, Bowden DW, Bown MJ, Chambers JC, Chan JC, Chasman D, Cho J, Chung MK, Cohen B, Correa A, Dabelea D, Daly MJ, Darbar D, Duggirala R, Dupuis J, Ellinor PT, Elosua R, Erdmann J, Esko T, Färkkilä M, Florez J, Franke A, Getz G, Glaser B, Glatt SJ, Goldstein D, Gonzalez C, Groop L, Haiman C, Hanis C, Harms M, Hiltunen M, Holi MM, Hultman CM, Kallela M, Kaprio J, Kathiresan S, Kim B-J, Kim YJ, Kirov G, Kooner J, Koskinen S, Krumholz HM, Kugathasan S, Kwak SH, Laakso M, Lehtimäki T, Loos RJF, Lubitz SA, Ma RCW, MacArthur DG, Marrugat J, Mattila KM, McCarroll S, McCarthy MI, McGovern D, McPherson R, Meigs JB, Melander O, Metspalu A, Neale BM, Nilsson PM, O'Donovan MC, Ongur D, Orozco L, Owen MJ, Palmer CNA, Palotie A, Park KS, Pato C, Pulver AE, Rahman N, Remes AM, Rioux JD, Ripatti S, Roden DM, Saleheen D, Salomaa V, Samani NJ, Scharf J, Schunkert H, Shoemaker MB, Sklar P, Soininen H, Sokol H, Spector T, Sullivan PF, Suvisaari J, Tai ES, Teo YY, Tiinamaija T, Tsuang M, Turner D, Tusie-Luna T, Vartiainen E, Vawter MP, Ware JS, Watkins H, Weersma RK, Wessman M, Wilson JG, Xavier RJ, Neale BM, Daly MJ, MacArthur DG. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. https://doi.org/10.1038/s41586-020-2308-7

16 Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S,

Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. (2021). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. https://doi.org/10.1093/nar/gkab1061

17 Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. https://doi.org/10.1126/science.ade2574

18 Bittrich S, Segura J, Duarte JM, Burley SK, Rose Y. (2024). RCSB protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. https://doi.org/10.1093/bioinformatics/btae370

19 Zhang Y. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. https://doi.org/10.1093/nar/gki524

20 Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A. (2020). FATCAT 2.0: towards a better understanding of the structural diversity of proteins. https://doi.org/10.1093/nar/gkaa443

21 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. https://doi.org/10.1038/msb.2011.75

22 Plotly Technologies Inc. (2015). Collaborative data science. https://plot.ly

23 arcadia-pycolor. (2025). https://github.com/Arcadia-Science/arcadia-pycolor