EasyChair Preprint

№ 15729

Positive-Unlabelled Learning to Identify New Genes Associated with Dietary Restriction

Jorge Paz-Ruza, Alex A. Freitas, Amparo Alonso-Betanzos and Bertha Guijarro-Berdiñas

January 18, 2025

Índice general

Aprendizaje Positivo sin Etiquetas para Identificar Nuevos Genes Relacionados con la Restricción Dietética

Jorge Paz-Ruza[1], Alex A. Freitas[2], Amparo Alonso-Betanzos[1], y Bertha Guijarro-Berdiñas[1]

[1] Universidade da Coruña - Grupo LIDIA, CITIC, 15071 A Coruña, España
[2] University of Kent - School of Computing, Canterbury CT2 7FS, Reino Unido
Correspondence: j.ruza@udc.es

DOI: *https://doi.org/543210/xxxxx1234567890*

Resumen:
La Restricción Dietética (DR) es un enfoque antienvejecimiento popular, y se han utilizado técnicas de aprendizaje automático para identificar genes relacionados con esta. Sin embargo, estas técnicas etiquetan los genes sin evidencia conocida como no relacionados con DR (es decir, como ejemplos negativos), lo que reduce su rendimiento. Este estudio presenta un nuevo método que utiliza un enfoque de Aprendizaje Positivo sin Etiquetas (PU) de dos pasos, lo que permite seleccionar ejemplos negativos confiables para entrenar un clasificador que distingue entre genes relacionados y no relacionados con DR. El método propuesto supera significativamente en rendimiento ($p<0.01$) a los enfoques no basados en PU. Como resultado, identificamos cuatro nuevos genes (PRKAB1, PRKAB2, IRS1, PRKAG1) con potencial relación con la DR.

1. Introducción

La Restricción Dietética (DR) es una de las intervenciones anti-envejecimiento más populares; involucra reducir la ingesta de nutrientes sin causar malnutrición (Most et al. 2017), y reduce el riesgo de patologías como las neurodegenerativas, cardiovasculares o cancerosas (de Carvalho 2022; López-Lluch and Navas 2016).

Recientemente, Vega Magdaleno et al. (2022) emplearon Aprendizaje Automático (AA) para identificar genes relacionados con el envejecimiento que serían candidatos a estar relacionados con la DR. Sin embargo, para proporcionar ejemplos con etiquetas binarias al clasificador a entrenar, se asumió que todos los genes sin evidencia experimental de relación con la DR se considerarían no relacionados con DR, etiquetándolos como muestras negativas. Como los autores reconocen, la ausencia de evidencia no equivale a evidencia de ausencia de relación del gen con la DR, implicando que los clasificadores se entrenaron con ejemplos negativos posiblemente erróneamente etiquetados, empeorando su aprendizaje, y por lo tanto la fiabilidad de las predicciones. En este tipo de datos, conocido como Datos Positivos y sin Etiquetar (PU), un subconjunto de ejemplos tiene etiqueta positiva conocida, mientras que el resto se conforma de ejemplos positivos y negativos pero con etiqueta desconocida (Elkan and Noto 2008).

En este trabajo proponemos el uso de Aprendizaje Positivo y sin Etiquetas, un paradigma de AA orientado a Datos PU, para mejorar la identificación de nuevos genes relacionados con la DR entre genes relacionados con el envejecimiento; en particular,

diseñamos una nueva técnica de Aprendizaje PU "en dos pasos" basada en distancias para tareas de priorización de genes. Nuestro método supera en rendimiento ($p < 0.01$) a la alternativa no-PU (Vega Magdaleno et al. 2022) en el conjunto real de datos de genes relacionados con la DR, y lo hemos utilizado para obtener un *ranking* más fiable de los nuevos genes potencialmente más relacionados con la DR.

2. Antecedentes

Esta sección formaliza la tarea de priorización genética, describe las aproximaciones existentes y sus limitaciones, e introduce nociones esenciales del Aprendizaje PU.

2.1. Formulación de la tarea

Sea $\mathcal{G}_{\text{AGE}}$ el conjunto de genes relacionados al envejecimiento, y $\mathcal{G}_{\text{DR}\cap\text{AGE}^+}$ el subconjunto de esos genes que están implicados en la DR. Entre estos, otro subconjunto $\mathcal{G}_{\text{DR}\cap\text{AGE}}$ tiene evidencia experimental de su relación con DR; por lo tanto, se cumple que $\mathcal{G}_{\text{DR}\cap\text{AGE}} \subset \mathcal{G}_{\text{DR}\cap\text{AGE}^+} \subset \mathcal{G}_{\text{AGE}}$. El objetivo es encontrar con un modelo $\Phi : \mathcal{G}_{\text{AGE}} \rightarrow [0, 1]$ aquellos genes sin evidencia experimental de relación con la DR, pero que realmente están relacionados con ellas:

$$\begin{aligned} &\text{argmax}_{g\in\mathcal{G}_{\text{AGE}}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}}}\Phi(g) \\ &\approx \text{argmax}_{g\in\mathcal{G}_{\text{AGE}}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}}} Pr(g \in \mathcal{G}_{\text{DR}\cap\text{AGE}^+}) \\ &\approx \mathcal{G}_{\text{DR}\cap\text{AGE}^+}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}} \end{aligned} \tag{1.1}$$

donde la clase positiva ($Pr(g \in \mathcal{G}_{\text{DR}\cap\text{AGE}^+}) = 1$) son genes relacionados con la DR, y la clase negativa $Pr(g \in \mathcal{G}_{\text{DR}\cap\text{AGE}^+}) = 0$ son genes no relacionados con la DR.

2.2. Metodología existente para la identificación de genes relacionados con DR

Recientemente, Vega Magdaleno et al. (2022) utilizaron AA para predecir la potencial relación con la DR de genes relacionados con el envejecimiento, caracterizando cada gen mediante características biológicas y entrenando *ensembles* de árboles de decisión (Grinsztajn et al. 2022). Identificaron dos combinaciones de tipo de características y clasificador con mayor rendimiento ($\{\text{PathDIP}, \text{CatBoost}\}$ y $\{\text{GO}, \text{BRF}\}$), y las utilizaron para producir sendos *rankings* de genes con mayor probabilidad de relación con la DR.

Los autores asumieron que todos los genes sin relación conocida con la DR son ejemplos negativos en entrenamiento, lo que equivale a enseñar al modelo ϕ que:

$$\begin{aligned} &\Phi(g) = 0 \quad \forall g \in \mathcal{G}_{\text{AGE}}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}} \\ &Pr(g \in \mathcal{G}_{\text{DR}\cap\text{AGE}^+}) = 0 \quad \forall g \in \mathcal{G}_{\text{AGE}}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}} \\ &\mathcal{G}_{\text{DR}\cap\text{AGE}^+}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}} = \varnothing \end{aligned} \tag{1.2}$$

lo cual contradice a la tarea, consistente en encontrar los genes pertenecientes precisamente al subconjunto $\mathcal{G}_{\text{DR}\cap\text{AGE}^+}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}}$. En consecuencia, el modelo es entrenado con datos con ruido de etiquetado, reduciendo la fiabilidad de sus predicciones.

2.3. Nociones Esenciales de Aprendizaje Positivo y sin Etiquetas

El Aprendizaje Positivo y sin Etiquetas es un paradigma de AA diseñado para escenarios donde el conjunto está compuesto de un subconjunto de ejemplos positivos $\mathcal{P}$ y un subconjunto de ejemplos sin etiquetar $\mathcal{U}$, que contiene ejemplos tanto positivos como negativos (Elkan and Noto 2008). Ha sido explorado con frecuencia en tareas de bioinformática, donde los ejemplos positivos son prioritarios pero resultan difíciles y/o costosos de etiquetar (Li et al. 2022; Zhang et al. 2024; Zheng et al. 2019).

Este trabajo se centra en los métodos denominados "en dos pasos" (Bekker and Davis 2020): como ilustra la figura 1, en vez de considerar cualquier ejemplo sin etiquetar como negativo (e.g. como en el trabajo de Vega Magdaleno et al. (2022)), el clasificador es entrenado utilizando los ejemplos positivos $\mathcal{P}$ y un subconjunto de "negativos fiables" $\mathcal{RN} \subset \mathcal{U}$ extraídos de entre los ejemplos sin etiquetar. Así, el clasificador se entrena con un ruido de etiquetado mínimo, aumentando la calidad del entrenamiento y las predicciones.

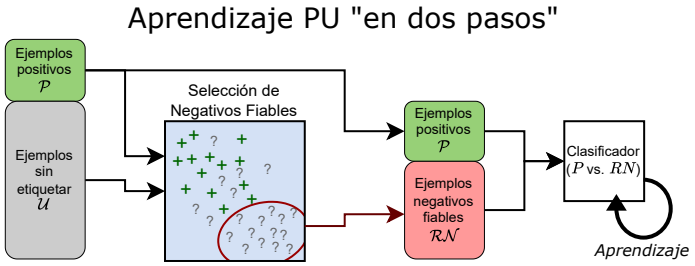

Figura 1: Estructura a alto nivel de una técnica de Aprendizaje Positivo y sin Etiquetas "en dos pasos".

3. Método de *PU Learning* propuesto

Nuestra metodología PU es un método basado en similitudes e inspirado por una clasificación kNN clásica (Cover and Hart 1967); dentro de la taxonomía de métodos Bekker and Davis (2020), constituye un método en dos pasos sin uso de probabilidad *a priori*. Dado un conjunto de datos PU con subconjuntos P (genes con relación conocida con la DR) y U (genes sin relación conocida), el conjunto de genes RN con improbable relación con la DR se obtiene como sigue:

1. Por cada par de genes (g_i, g_j), se computa su similitud mediante sus vectores de características $x_g = (f_1, ..., f_{|\mathcal{F}|})$, donde $\mathcal{F}$ es el conjunto de características usadas. Debido a la naturaleza binaria y de gran dimensionalidad de los conjuntos de características utilizados, optamos por el Índice de Jaccard $J(x_{g_i}, x_{g_j})$ (Jaccard 1901; Tanimoto 1958), tal que:

$$J(x_{g_i}, x_{g_j}) = \frac{\sum_{k=1}^{|\mathcal{F}|} x_{g_i}[k] * x_{g_j}[k]}{\sum_{k=1}^{|\mathcal{F}|} x_{g_i}[k] + \sum_{k=1}^{|\mathcal{F}|} x_{g_j}[k] - \sum_{k=1}^{|\mathcal{F}|} x_{g_i}[k] * x_{g_j}[k]} \tag{1.3}$$

2. El conjunto de negativos fiables $\mathcal{RN}$ se inicializa sin elementos.

3. Para cada gen sin etiquetar g_i en el conjunto de entrenamiento:
   a) Encontrar los k genes más similares a g_i según sus Índices de Jaccard.
   b) Si se cumplen dos condiciones, se añade el gen g_i a $\mathcal{RN}$:
      - Que el gen más similar a g_i no está etiquetado.
      - Que la proporción de genes sin etiquetar en los k genes más similares a g_i es mayor a un umbral t.

El Algoritmo 1 detalla la lógica arriba descrita, constituyendo el primer paso de la metodología. Así, el conjunto $\mathcal{RN}$ verificará:

$$\mathcal{RN} \subset \mathcal{G}_{\text{AGE}} \backslash \mathcal{G}_{\text{DR}\cap\text{AGE}^+} \quad , \quad \mathcal{RN} \cap \mathcal{G}_{\text{DR}\cap\text{AGE}^+} = \varnothing \tag{1.4}$$

Algorithm 1 Selección de Negativos Fiables basada en Vecinos Próximos

Require:
D: Ejemplos de entrenamiento
k: No. de vecinos
$t \in [0.5, 1]$: Umbral
$\mathcal{F}$: Conjunto de características
Ensure:
RN: Ejemplos negativos fiables para entrenamiento

function RELIABLENEGATIVES($P, U, k, t, \mathcal{F}$)
  $P, U \leftarrow P, U$ considerando las características $\mathcal{F}$
  $D \leftarrow P \cup U$
  Inicializar la matriz de similitudes $S \in \mathbb{R}^{|D|\times|D|}$
  for all (x_i, x_j); $x_i, x_j \in D$ **do** ▷ *Calcular Índice de Jaccard para todos los pares de ejemplo*
    $S_{ij} \leftarrow \text{J}(x_i, x_j)$ ▷ *Puede memoizarse S para ahorrar re-calcular* S_{ij}
  $RN \leftarrow \varnothing$
  for all $x_i \in U$ **do**
    $Top_k \leftarrow$ Top k ejemplos $x_j \in D \backslash x_i$ con mayor similitud S_{ij} a x_i
    $x_{max\_sim} \leftarrow$ El ejemplo $x_j \neq x_i$ con mayor similitud S_{ij} to x_i
    ▷ *Si el % de ejemplos sin etiquetar en los* Top_k *vecinos próximos de* x_i *excede una proporción dada y el vecino más próximo no está etiquetado, añadir* x_i *a los negativos fiables* ◁
    if $\frac{|T_k \cap U|}{|T_k|} \geqslant t$ & $x_{max\_sim} \in U$ **then**
      $RN \leftarrow RN \cup x_i$
  return RN

En el segundo y último paso, $\mathcal{P}$ y $\mathcal{RN}$ son utilizados para entrenar el clasificador final Φ con un ruido de etiquetado mínimo, de forma agnósitca al clasificador utilizado. Los hiperparámetros k y t pueden ser optimizados dentro de la validación cruzada anidada propuesta y utilizada por Vega Magdaleno et al. (2022) en su metodología.

4. Configuración experimental

Esta sección describe los conjuntos de características y clasificadores y la metodología de evaluación utilizada.

4.1. Características y clasificadores

Estando cada gen representado por un vector de características biológicas y una etiqueta indicando si hay evidencia experimental de su relación con DR, en este trabajo

utilizamos los dos conjuntos de características que reportaron mayor poder predictivo en el trabajo de Vega Magdaleno et al. (2022): indicadores PathDIP (Rahmati et al. 2017) (que indican si un gen pertenece o no a una determinada ruta metabólica), y términos biológicos GO (Ashburner et al. 2000) (*Gene Ontology*) (que indican si un gen está relacionado con un término de la ontología o cualquiera de sus descendientes. Para obtener el conjunto de genes relacionados con el envejecimiento, sus características y cualquier relación con la DR conocida, seguimos el mismo proceso detallado en el trabajo de Vega Magdaleno et al. (2022). El Cuadro 1 muestra estadísticas básicas de los dos conjuntos construidos a partir de cada conjunto de características.

Cuadro 1: Estadísticas de los conjuntos de datos construidos a partir de características PathDIP y GO.

Conjunto de características	Características	Genes relacionados al envejecimiento ($\|\mathcal{G}_{AGE}\|$)	Genes con relación a DR conocida ($\|\mathcal{G}_{DR \cap AGE}\|$)	Dispersión de características (%)
PathDIP	1,640	986	110	98.39 %
GO	8,640	1,124	114	98.46 %

Con respecto a los clasificadores, utilizamos los *ensembles* de árboles CatBoost (CAT) (Prokhorenkova et al. 2018) y *Balanced Random Forest* (BRF) (Chen et al. 2004); éstos fueron los clasificadores que obtuvieron mayor poder predictivo en el trabajo original de Vega Magdaleno et al. (2022).

4.2. Evaluación

Para evaluar el poder predictivo de nuestro método PU en comparación con el existente (no-PU) de Vega Magdaleno et al. (2022), utilizamos la medida F1 de la clase positiva, la media geométrica de la sensibilidad por clase, y el AUC-ROC (Japkowicz and Shah 2011). Cabe mencionar que, al estar utilizando datos PU genuinos, estas tres métricas son únicamente estimaciones (típicamente subestimaciones) de sus valores reales, y la F1 estimada siempre subestimará la F1 real en un factor constante para todos los modelos evaluados (Elkan and Noto 2008).

5. Resultados

Esta sección muestra y compara los resultados de nuestro modelo basado en Aprendizaje Positivo y sin Etiquetas (PU) con respecto al método (no-PU) existente en la identificación de genes relacionados con DR (Vega Magdaleno et al. 2022), comparando el rendimiento en experimentos computacionales y los *rankings* de nuevos genes candidatos con mayor probabilidad predicha de relación con la DR.

5.1. Resultados de experimentos computacionales

El Cuadro 2 muestra el rendimiento predictivo de nuestro método (PU) y el método existente (no-PU) de Vega Magdaleno et al. (2022), utilizando los dos conjuntos de características y los dos clasificadores con más poder predictivo reportado en su trabajo.

Los mejores resultados de cada métrica fueron obtenidos utilizando nuestro método PU, siendo éstos significativos ($p < 0.05$, con $p < 0.01$ en medida F1 y M. Geométrica) en todos los casos. En particular, es destacable el rendimiento de nuestro método PU utilizando {PathDIP, CAT}, puesto que maximiza la F1, la métrica más fiable en tareas PU, y es competitiva en AUC-ROC y M.Geométrica.

Cuadro 2: Rendimiento predictivo del método existente (no-PU) y nuestro método PU en la identificación de genes relacionados con DR (media de 10 ejecuciones de la validación cruzada anidada). El mejor resultado se resalta en negrita. Una daga (†) y dos dagas (††) representan significancia estadística contra el resto de resultados con $\alpha = 0.05$ y $\alpha = 0.01$ respectivamente.

Método	**Características**	**Clasificador**	**Rendimiento**		
			AUC-ROC	**M. Geométrica**	**Medida FU**
Método existente	PathDIP	CAT	0.829	0.717	0.522
		BRF	0.825	0.752	0.450
	GO	CAT	0.832	0.654	0.463
		BRF	0.827	0.755	0.377
Aprendizaje PU	PathDIP	CAT	0.829	0.750	**0.537††**
		BRF	0.815	0.728	0.381
	GO	CAT	**0.838†**	0.726	0.491
		BRF	0.829	**0.763††**	0.380

5.2. Análsisis de los nuevos genes con más probable relación predicha a la DR

Empleamos el método con mayor rendimiento predictivo (nuestro método PU utilizando características PathDIP y CatBoost como clasificador) para obtener el *ranking* de genes con mayor probabilidad de relación a la DR, y lo comparamos al *ranking* de 7 genes propuesto por Vega Magdaleno et al. (2022) en su trabajo con un método no-PU.

Cuadro 3: Top 7 genes con mayor probabilidad predicha de relación con la DR, calculados con el método existente (no-PU, izquierda) y nuestro método PU propuesto (derecha). La probabilidad predicha es la salida media sin binarizar dada por el modelo para el gen a lo largo de 10 ejecuciones de la validación cruzada anidada. Los genes comunes a ambos *rankings* se resaltan en cursiva.

Método existente (no-PU)		**Método con Aprendizaje PU**	
Gen	**Probabilidad de DR**	**Gen**	**Probabilidad de DR**
GOT2	0.86	*TSC1*	0.97
GOT1	0.85	*GCLM*	0.94
TSC1	0.85	IRS1	0.93
CTH	0.85	PRKAB1	0.92
GCLM	0.82	PRKAB2	0.90
IRS2	0.80	PRKAG1	0.90
SENS2	0.80	*IRS2*	0.90

Observamos que existen tres genes en común en los *ránkings* obtenidos con los dos métodos, con 4 genes que difieren. Puesto que nuestro método basado en Aprendizaje PU obtuvo un rendimiento predictivo significativamente superior, podemos concluir que los mejores candidatos propuestos por nuestro método tienen mayor fiabilidad, y por lo tanto probabilidad real de estar relacionados con la DR.

Además, mientras que el método no-PU de Vega Magdaleno et al. (2022) propuso genes para los cuales los autores no encontraron posibles relaciones con mecanismos asociados a la DR en la literatura científica relevante, en nuestro caso observamos que la literatura existente respalda la experimentación en laboratorio para todos los genes propuestos por nuestro método PU. Esto es cierto para *PRKAB1* y *PRKAB2* (Katwan et al. 2019), *PRKAG1* (Ripa et al. 2023), e *IRS1* (Dean and Cartee 2000).

6. Conclusiones

Este trabajo explora el uso de técnicas de Aprendizaje PU para mejorar las tareas de priorización de genes en el ámbito de la Restricción Dietética. Comparado con los métodos existentes no-PU, que introducen ruido de etiquetado en el entrenamiento, nuestra técnica PU proporciona un entrenamiento con datos refinados de mayor calidad, mejorando el rendimiento predictivo ($p < 0.01$ en Medida F1) en experimentos computacionales. Usando nuestro mejor modelo para obtener un ránking más fiable de

nuevos genes prometedores con mayor probabilidad de relación con la DR, observamos que la literatura existente apoyaría la experimentación en laboratorio de la potencial relación con la DR de los genes propuestos por nuestro modelo.

Respecto a la investigación futura, identificamos: 1) la validación del rol dentro de la DR de los genes identificados en experimentos de laboratorio, 2) la combinación de distintos conjuntos de características biológicas para mejorar la calidad de los predicciones, y 3) la exploración de otros clasificadores para evaluar nuestro método y potencialmente aumentar su rendimiento.

Agradecimientos

Este trabajo está financiado por MICIU/AEI/10.13039/501100011033/ y FEDER Una manera de hacer Europa (PID2019-109238GB-C22, PID2023-147404OB-I00), y el FSE+ (FPU21/05783), así como la Xunta de Galicia (ED431C 2022/44) mediante fondos ERDF de la Unión Europea, y el Ministerio de Transformación Digital y de la Función Pública (TSI-100925-2023-1). CITIC, como Centro de Investigación acreditado por el Sistema Universitario Galego, está financiado por la Consellería de Cultura, Educación e Universidades de la Xunta de Galicia, apoyado un 80 % por el Programa Operacional ERDF Galicia 2021-2027, y un 20 % por la Secretaría Xeral de Universidades de la Xunta de Galicia (ED431G 2023/01).

Bibliografía

Ashburner, Michael, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al., 2000. "Gene ontology: tool for the unification of biology." *Nature genetics* 25(1), pages 25–29.

Bekker, Jessa and Jesse Davis, 2020. "Learning from positive and unlabeled data: A survey." *Machine Learning* 109(4), pages 719–760.

de Carvalho, Tayana Silva, 2022. "Calorie restriction or dietary restriction: how far they can protect the brain against neurodegenerative diseases?" *Neural Regeneration Research* 17(8), pages 1640–1644.

Chen, Chao, Andy Liaw, and Leo Breiman, 2004. "Using random forest to learn imbalanced data." Technical Report 666, Department of Statistics, UC Berkley. Available from: *http://xtf.lib.berkeley.edu/reports/SDTRWebData/accessPages/666.html*.

Cover, T. and P. Hart, 1967. "Nearest neighbor pattern classification." *IEEE Transactions on Information Theory* 13(1), pages 21–27. doi:10.1109/TIT.1967.1053964.

Dean, D.J. and G.D. Cartee, 2000. "Calorie restriction increases insulin-stimulated tyrosine phosphorylation of insulin receptor and insulin receptor substrate-1 in rat skeletal muscle." *Acta Psyshiol. Scand.* 169, pages 133–139.

Elkan, Charles and Keith Noto, 2008. "Learning classifiers from only positive and unlabeled data." In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. pages 213–220.

GRINSZTAJN, LÉO, EDOUARD OYALLON, and GAËL VAROQUAUX, 2022. "Why do tree-based models still outperform deep learning on typical tabular data?" *Advances in neural information processing systems* 35, pages 507–520.

JACCARD, PAUL, 1901. "Étude comparative de la distribution florale dans une portion des alpes et des jura." *Bull Soc Vaudoise Sci Nat* 37, pages 547–579.

JAPKOWICZ, NATHALIE and MOHAK SHAH, 2011. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

KATWAN, O.J., F. ALGHAMDI, T.A. ALMABROUK, S.J. MANCINI, S. KENNEDY, J.S. OAKHILL, J.W. SCOTT, and I.P. SALT, 2019. "Amp-activated protein kinase complexes containing the beta 2 regulatory subunit are up-regulated during and contribute to adipogenesis." *Biochemical Journal* 476, pages 1725–1740.

LI, FUYI, SHUANGYU DONG, ANDRÉ LEIER, MEIYA HAN, XUDONG GUO, JING XU, XIAOYU WANG, SHIRUI PAN, CANGZHI JIA, YANG ZHANG, ET AL., 2022. "Positive-unlabeled learning in bioinformatics and computational biology: a brief review." *Briefings in bioinformatics* 23(1), page bbab461.

LÓPEZ-LLUCH, GUILLERMO and PLÁCIDO NAVAS, 2016. "Calorie restriction as an intervention in ageing." *The Journal of physiology* 594(8), pages 2043–2060.

MOST, JASPER, VALERIA TOSTI, LEANNE M REDMAN, and LUIGI FONTANA, 2017. "Calorie restriction in humans: an update." *Ageing research reviews* 39, pages 36–45.

PROKHORENKOVA, LIUDMILA, GLEB GUSEV, ALEKSANDR VOROBEV, ANNA VERONIKA DOROGUSH, and ANDREY GULIN, 2018. "Catboost: unbiased boosting with categorical features." *Advances in neural information processing systems* 31.

RAHMATI, SARA, MARK ABOVSKY, CHIARA PASTRELLO, and IGOR JURISICA, 2017. "pathdip: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis." *Nucleic acids research* 45(D1), pages D419–D426.

RIPA, R., E. BALLHYSA, J.D. STEINER, R. LABOY, A. ANNIBAL, N. HOCHHARD, C. LATZA, L. DOLFI, C. CALABRESE, A.M. MEYER, M.C. POLIDORI, R.U. MULLER, and A. ANTEBI, 2023. "Refeeding-associated ampkγ1 complex activity is a hallmark of health and longevity." *Nature Aging* 3, pages 1544–1560.

TANIMOTO, TAFFEE T, 1958. "Elementary mathematical theory of classification and prediction." .

VEGA MAGDALENO, GUSTAVO DANIEL, VLADISLAV BESPALOV, YALIN ZHENG, ALEX A FREITAS, and JOAO PEDRO DE MAGALHAES, 2022. "Machine learning-based predictions of dietary restriction associations across ageing-related genes." *BMC bioinformatics* 23, pages 1–28.

ZHANG, DACHUAN, HUADONG XING, DONGLIANG LIU, MENGYING HAN, PENGLI CAI, HUIKANG LIN, YU TIAN, YINGHAO GUO, BIN SUN, YINGYING LE, ET AL., 2024. "Discovery of toxin-degrading enzymes with positive unlabeled deep learning." *ACS Catalysis* 14, pages 3336–3348.

ZHENG, YI, HUI PENG, XIAOCAI ZHANG, ZHIXUN ZHAO, XIAOYING GAO, and JINYAN LI, 2019. "Ddi-pulearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions." *BMC bioinformatics* 20, pages 1–12.