Boundary–Crossing Probabilities for Stochastic Processes and Their Applications

Senren Tan

Supervisors: Prof. Vladimir K. Kaishev
Dr. Dimitrina S. Dimitrova

Faculty of Actuarial Science and Insurance
Cass Business School, City, University of London

A thesis submitted for the degree of
Doctor of Philosophy

December 2019

To my loving parents and my grandpa

Table of contents

City, University of London
Northampton Square
London
EC1V 0HB
United Kingdom

T +44 (0)20 7040 5060

THE FOLLOWING PARTS OF THIS THESIS HAS BEEN REDACTED FOR COPYRIGHT REASONS:

List of figures

List of tables

Acknowledgements

This PhD research project has been funded by way of a bursary from the Faculty of Actuarial Science and Insurance at Cass Business School, City, University of London.

I would like to first thank my supervisors, Prof. Vladimir K. Kaishev and Dr. Dimitrina S. Dimitrova for their constant support throughout every stage of my PhD. Their knowledgeability and insights in this topic, enthusiasms about academic research, scrutinized attitude toward science have driven myself to be a better researcher and led to this fruitful project. Also, I have learnt from them that doing academic research is not only about developing theories correctly, but also being able to communicate the theories to others in a clear, understandable way, and making theories applicable to a wider community.

I also owe my sincere gratitude to my grandpa. Without him I would not have started my PhD journey.

Finally, I am indebted to my parents and Jackie for their endless love and encouragement that support me through many difficult times during my PhD.

Declaration

I herby declare that I have produced this thesis without assistance of any third parties other than the coauthors of the papers. Except where specific reference is made to the work of others, the contents of this thesis are original and have not previously been submitted in identical or similar form to any other degree or qualification in this, or any other university.

This thesis work was conducted from October 2015 to June 2019 under the supervision of Prof. Vladimir K. Kaishev and Dr. Dimitrina S. Dimitrova at Cass Business School, City, University of London.

Senren Tan
December 2019

Declaration

Powers of discretion are hereby granted to the University Librarian to allow this thesis to be copied in whole or in part without further reference to the author. The permission covers only simple copies made for study purpose, subject to the normal conditions of acknowledgement.

Senren Tan
December 2019

Abstract

In this thesis, we focus on the problem that a stochastic process crossing (or not crossing) upper and/or lower deterministic boundaries and its application in statistics, inventory management, finance, risk and ruin theory and queueing. In Chapter 2, we provide a fast and accurate method based on fast Fourier transform (FFT), to compute the (complementary) cumulative distribution function (CDF) of the Kolmogorov-Smirnov (KS) statistic when the CDF under the null hypothesis, $F(x)$, is purely discrete, mixed or continuous, and thus obtain exact p values of the KS test. Secondly, we developed a C++ and an R implementation of the proposed method, which fills in the existing gap in statistical software. The numerical performance of the proposed FFT-based method, implemented both in C++ and in the R package **KSgeneral**, available from https://CRAN.R-project.org/package=KSgeneral, is illustrated when $F(x)$ is mixed, purely discrete, and continuous. In Chapter 3, we develop an efficient method based on FFT, for computing the probability that a non-decreasing, pure jump (compound) stochastic process stays between arbitrary upper and lower boundaries (i.e., deterministic functions, possibly discontinuous) within a finite time period. We further demonstrate that our FFT-based method is computationally efficient and can be successfully applied in the context of inventory management (to determine an optimal replenishment policy), ruin theory (to evaluate ruin probabilities and related quantities) and double-barrier option pricing or simply computing non-exit probabilities for Brownian motion with general boundaries. In Chapter 4, we give explicit formulas and a numerically efficient FFT-based method for computing the probability that a non-decreasing, pure jump stochastic process will first exit from above the strip between two deterministic, possibly discontinuous, time-dependent boundaries, within a finite-time interval with an overshoot (not) exceeding a positive value. The stochastic process is a compound process with events of interest arriving according to an arbitrary point process with conditional stationary independent increments (PPCSII), and event severities with any possibly dependent joint distribution. The class of PPCSII is rather rich covering point processes with independent increments (among which non-homogeneous Poisson processes and negative binomial processes), doubly stochastic Poisson (i.e., Cox processes) including mixed Poisson processes (among which processes with the order statistics property) and Markov modulated point processes. These assumptions make our framework and results generally applicable for a broad range of models arising in insurance, finance, queueing, economics, physics, astronomy and many other fields. We present examples of such applications in queueing, ruin and inventory management optimization, leading to new results in the latter fields, illustrated also numerically. In Chapter 5, we consider the large class of PPCSII and the family GD of random variables with arbitrary, possibly dependent joint distribution. These families are interchangeably used to model customers arrival and service times in the very general framework of GD/PPCSII/1 and its inverse PPCSII/GD/1 queueing models. The latter cover well known models, e.g. the G/M/1 and M/G/1 queues, but also models incorporating dependence in the arrival times, service times and across, either by directly stating their joint distribution, through a copula and appropriate marginals, or through the PPCSII class. We further

introduce a double–boundary crossing (DBC)–queueing duality that extends the known Cramér–Lundberg – G/M/1 duality. The DBC–queueing duality is used to establish new results with respect to the joint and marginal distributions of the busy period, idle time and the maximum waiting time, including bounds, approximations and closed form formulas. We present a FFT-based method for efficient computation of the latter distributions. We also formulate and solve novel profit optimization problems, e.g., of determining the optimal capacity of the server so as to maximize the worse-case profit margin jointly with its related probability. Results are illustrated numerically.

Chapter 1

Introduction

This thesis focuses on the problem that a stochastic process crossing (or not crossing) upper and/or lower deterministic boundaries and its applications in statistics, risk and ruin theory, finance, queueing and inventory management.

In statistics, the two-sided Kolmogorov-Smirnov (KS) statistic is one of the most popular goodness-of-fit test statistics that is used to measure how well the distribution of a random sample (of size n) agrees with a pre-specified theoretical cumulative distribution function (CDF) under the null hypothesis. When the CDF under the null hypothesis is continuous, the distribution of the KS statistic is closely related to the probability that the order statistics of n $uniform(0,1)$ random variables all lie within an n-dimensional rectangle, also referred to as the *rectangle probability for uniform order statistics*. The latter probability can be expressed as the probability that the empirical process lies between two (appropriately defined) parallel straight lines, which can be re-expressed as a more easily computable probability that a homogeneous Poisson process stays within the corridor between two (appropriately defined) upper and lower boundaries. We refer to the latter probability as the *double-boundary non-crossing (DB(non-)C) probability* for a Poisson process.

On the other hand, there are many real-life applications, e.g., in biology, physics, engineering, finance, and insurance, in which fitting discrete or mixed distributions, i.e., with multiple jumps and continuous segments, to large samples of data is required.

However, due to inherent difficulties, the distribution of the KS statistic when the CDF under the null hypothesis has jump discontinuities has been studied to a much lesser extent and no exact and efficient computational methods have been proposed in the literature.

For this purpose, we develop a fast and accurate method to compute the (complementary) CDF of the KS statistic when the CDF under the null hypothesis is discontinuous, and thus obtain exact p values of the KS test. Our approach is to first express the complementary CDF through an appropriately defined rectangle probability for uniform order statistics, which is then re-expressed as the DB(non-)C probability for an empirical process, with modified non-linear boundaries. The latter probability can be obtained by considering an equivalent DB(non-)C probability for a homogeneous Poisson process and hence, an appropriate system of Chapman-Kolmogorov forward equations, which can then be efficiently computed, based on circular convolution theorem, using fast Fourier transform (FFT). We further implement the proposed method in C++ and in the R package **KSgeneral**, available from https://CRAN.R-project.org/package=KSgeneral, which fills in the existing gap in statistical software. In fact, the proposed method is also applicable for computing the distribution of other KS-type statistics that have higher statistical power when the CDF (possibly with jump discontinuities) under the alternative hypothesis behaves differently in the tails (e.g., the standardized Smirnov statistic, the Studentized Smirnov statistic, etc.).

We further generalize the proposed FFT-based method so as to compute the DB(non-)C probability for a very large class of models (processes and boundaries). Namely, we consider general boundaries (i.e., arbitrary deterministic functions with possible jump discontinuities) and assume that the underlying stochastic process may not necessarily be homogeneous Poisson. The latter can be any process from the wide class of compound processes in which the process modelling event arrivals belongs to the large family of *point processes with conditional stationary independent increments* (PPCSII). This rather general family includes not only (non-)homogeneous Poisson, binomial, negative binomial processes, but also processes that may not necessarily be stationary and have independent increments, such as the doubly stochastic (i.e., Cox)

and mixed Poisson processes, allowing for dependence/clustering of the event arrivals. We demonstrate that the proposed general boundary crossing model and FFT-based method can be very useful in the context of operations research, in formulating and solving inventory management optimization problems, in finance in pricing barrier options or computing non-exit probabilities for Brownian motion, in risk theory in computing ruin probabilities.

As one of the applications of the DB(non-)C problem, we consider a simple single-item (single-product) single warehouse periodic review inventory model in which batches of different sizes are shipped (i.e., replenished) from a supplier to the warehouse, over a fixed time horizon, with certain (fixed) lead times. To the best of our knowledge, we show for the first time that inventory management optimization problems can be elegantly formulated (and solved) by incorporating an appropriate DB(non-)C probability constraint. In the DB(non-)C problem, the demand arrival process is assumed to be from the family of PPCSII (i.e., cumulative demand over time modelled by a compound PPCSII process), and the fixed lower boundary is viewed as the minimum demand below which the firm will fail to reach its sales targets and ensure flow of revenue sufficient to cover its operating costs and sustain its business, whereas the upper boundary models the aggregate units of the item replenished throughout the finite-time period. By strategically selecting the upper boundary (i.e., the number of shipments, batch sizes and future shipment times), the total ordering and holding costs incurred to the warehouse are minimized, while at the same time the probability that within the finite-time interval, the demand does not exceed the cumulative amount of replenished items, and also does not fall below the minimum demand limit, is sufficiently large. In addition, by considering the above DBC problem involving the overshoot of the demand process (from the upper boundary), the stockout cost incurred to the warehouse is also directly taken into account.

Moreover, computing DB(non-)C probabilities for Brownian motion has attracted considerable attention in the applied probability literature where approximation schemes have been developed for the case of (piece-wise) linear boundaries (Borovkov and Novikov, 2005, Wang and Pötzelberger, 2007, Ycart and Drouilhet, 2016), strictly

continuous boundaries (Fu and Wu, 2010) and a numerical approximation method for general boundaries based on direct convolution (Khmaladze and Shinjikashvili, 2001). We demonstrate that the proposed FFT-based method can be viewed as a significant enhancement of the approach taken by the latter authors, achieving much better efficiency in computing DB(non-)C probabilities for general, possibly discontinuous boundaries. Since the DB(non-)C probability for Brownian motion is closely related to the fair price of a barrier option in the Black-Scholes setting (see e.g., Borovkov and Novikov, 2005), we further illustrate the applicability of the proposed FFT-based method in pricing multi-step double-barrier options, with arbitrary number of jumps (i.e., steps) in the barriers. The latter options, as noted by Guillaume (2010), are popular in over-the-counter markets.

Furthermore, in insurance risk and ruin theory, computing ruin probability is important in modelling liquidity risk, estimating operational risk and assessing risk capital in insurance and banking, and also in other real-life risk analysis applications among which, flood risk, systems reliability risk and emerging disease spread risk (see Dimitrova et al., 2015). Ruin occurs when the compound process modelling aggregate claims exceeds for the first time the upper boundary (representing the aggregate insurance premium) within a finite time interval. Interpreting the latter as double-boundary crossing (DBC) probability (lower boundary equal to zero) allows us to employ the proposed FFT-based method to efficiently compute ruin probabilities for any claims arrival model from the PPCSII class and arbitrarily distributed claim sizes. In addition, the joint distribution of the time to ruin and the deficit at ruin for the very wide class of PPCSII can be obtained by considering the above DBC problem involving the overshoot of the aggregate claims process (from the upper boundary). To the best of our knowledge, no such alternative general method, or one specifically for Cox process arrivals has been considered in the actuarial literature.

Finally, it has for long been recognized that some important connections exist between single-server queues and inventory and insurance risk and ruin models (see e.g., Asmussen and Albrecher, 2010). Frostig (2004) has noted that the time to ruin, and the deficit at ruin in the classical Cramér-Lundberg (CL) insurance risk process are

correspondingly equivalent to the busy period and the idle time in the G/M/1 single-server queueing system, which we refer to as the CL-G/M/1 duality. We consider a very general single-server queueing model in which customer inter-arrival times may be dependent, with any joint distribution (which we code as GD), and service times form a point process from the very rich class of PPCSII, which we refer to as the GD/PPCSII/1 queue. We further introduce a new DBC-queueing duality which extends the known CL-G/M/1 duality by generalizing the G/M/1 model to the GD/PPCSII/1 one, by considering a finite-time horizon, and by introducing a second boundary. This has allowed us to consider for the first time the joint distribution of the busy period, idle time and the maximum waiting time in the very general GD/PPCSII/1 model. We also obtain lower bounds and approximations for the joint distribution of the busy period, idle time and the maximum waiting time in the general GD/PPCSII/1 and PPCSII/GD/1 models, as well as exact closed form expressions for the joint distribution of the busy period and idle time and its marginals, for the GD/OSPP/1 sub-model. Moreover, we extend the FFT-based method that, based on the DBC-queueing duality, can be used for fast and accurate computation of the joint distribution of the busy period, idle time and maximum waiting time. In addition, we formulate and solve, using the FFT-based method, a new profit optimization problem that focuses on the instantaneous maximization of the worst-case profit and its related probability. Maximization is carried out with respect to the parameter(s) of the service intensity (process), directly linked to the service capacity. As yet another contribution, we establish a novel duality between DBC problem and queueing and give new results and a closed form expression for the probability that the virtual waiting time process exceeds a fixed level. The latter process is central in queueing (see e.g., Cohen, 1982) and the related level crossing probability can be viewed as an important queue performance measure.

1.1 Chapter summaries

This thesis is organized as a series of papers, each of which is presented in a separate chapter. Chapter 2 has been accepted for publication by *Journal of Statistical Software*.

Chapter 3 has been accepted for publication by *European Journal of Operational Research*. Other chapters have been submitted to peer reviewed journals. It is worth pointing out that all the papers are based on joint work with my PhD supervisors. In what follows, we summarize the main results of each chapter, and provide a list of publications arising from this thesis.

In Chapter 2, we study the distribution of the one-sample Kolmogorov-Smirnov (KS) test statistic, which has been widely studied under the assumption that the underlying theoretical cumulative distribution function (CDF), $F(x)$, is continuous. However, there are many real-life applications in which fitting discrete or mixed distributions is required. Nevertheless, due to inherent difficulties, the distribution of the KS statistic when $F(x)$ has jump discontinuities has been studied to a much lesser extent and no exact and efficient computational methods have been proposed in the literature. In this chapter, we provide a fast and accurate method to compute the (complementary) CDF of the KS statistic when $F(x)$ is discontinuous, and thus obtain exact p values of the KS test. Our approach is to express the complementary CDF through the rectangle probability for uniform order statistics, and to compute it using fast Fourier transform (FFT). Secondly, we provide a C++ and an R implementation of the proposed method, which fills in the existing gap in statistical software. We give also a useful extension of the Schmid's asymptotic formula for the distribution of the KS statistic, relaxing his requirement for $F(x)$ to be increasing between jumps and thus allowing for any general mixed or purely discrete $F(x)$. The numerical performance of the proposed FFT-based method, implemented both in C++ and in the R package **KSgeneral**, available from https://CRAN.R-project.org/package=KSgeneral, is illustrated when $F(x)$ is mixed, purely discrete, and continuous. The performance of the general asymptotic formula is also studied.

In Chapter 3, we develop an efficient method for computing the probability that a non-decreasing, pure jump (compound) stochastic process stays between arbitrary upper and lower boundaries (i.e., deterministic functions, possibly discontinuous) within a finite time period. The compound process is composed of a process modelling the arrivals of certain events (e.g., demands for a product in inventory systems, customers

in queueing, or claims/capital gains in insurance/dual risk models), and a sequence of independent and identically distributed random variables modelling the sizes of the events. The events arrival process is assumed to belong to the wide class of point processes with conditional stationary independent increments (PPCSII) which includes (non-)homogeneous Poisson, binomial, negative binomial, mixed Poisson and doubly stochastic Poisson (i.e., Cox) processes as special cases. The proposed method is based on expressing the non-exit probability through Chapman-Kolmogorov equations, re-expressing them in terms of a circular convolution of two vectors which is then computed applying FFT. We further demonstrate that our FFT-based method is computationally efficient and can be successfully applied in the context of inventory management (to determine an optimal replenishment policy), ruin theory (to evaluate ruin probabilities and related quantities) and double-barrier option pricing or simply computing non-exit probabilities for Brownian motion with general boundaries.

In Chapter 4, we give explicit formulas and a numerically efficient FFT-based method for computing the probability that a non-decreasing, pure jump stochastic process will first exit from above the strip between two deterministic, possibly discontinuous, time-dependent boundaries, within a finite-time interval with an overshoot (not) exceeding a positive value. The stochastic process is a compound process with events of interest arriving according to an arbitrary member of the family of PPCSII, and event severities with any possibly dependent joint distribution. The class of PPCSII is rather rich covering point processes with independent increments (among which non-homogeneous Poisson processes and negative binomial processes), doubly stochastic Poisson (i.e., Cox processes) including mixed Poisson processes (among which processes with the order statistics property) and Markov modulated point processes. These assumptions make our framework and results generally applicable for a broad range of models arising in insurance, finance, queueing, economics, physics, astronomy and many other fields. We present examples of such applications in queueing, ruin and inventory management optimization, leading to new results in the latter fields, illustrated also numerically.

In Chapter 5, we consider the large class of PPCSII and the family GD of random variables with arbitrary, possibly dependent joint distribution. These families are interchangeably used to model customers arrival and service times in the very general framework of GD/PPCSII/1 and its inverse PPCSII/GD/1 queueing models. The latter cover well known models, e.g. the G/M/1 and M/G/1 queues, but also models incorporating dependence in the arrival times, service times and across, either by directly stating their joint distribution, through a copula and appropriate marginals, or through the PPCSII class. We further introduce a double–boundary crossing (DBC)–queueing duality that extends the known Cramér–Lundberg – G/M/1 duality. The DBC–queueing duality is used to establish new results with respect to the joint and marginal distributions of the busy period, idle time and the maximum waiting time, including bounds, approximations and closed form formulas. We present a FFT-based method for efficient computation of the latter distributions. We also formulate and solve novel profit optimization problems, e.g., of determining the optimal capacity of the server so as to maximize the worse-case profit margin jointly with its related probability. Results are illustrated numerically.

1.2 Publications arising from this thesis

Chapter 2: *Computing the Kolmogorov-Smirnov Distribution when the Underlying CDF is Purely Discrete, Mixed or Continuous.*

This chapter is based on the paper:

Dimitrova, D.S., Kaishev, V.K., Tan, S. 2019. Computing the Kolmogorov-Smirnov Distribution when the Underlying CDF is Purely Discrete, Mixed or Continuous. *Journal of Statistical Software*, forthcoming.

Chapter 3: *On Double-Boundary Non-Crossing Probability for a Class of Compound Processes with Applications.*

This chapter is based on the paper:

Dimitrova, D.S., Ignatov, Z.G., Kaishev, V.K., Tan, S. 2019. On Double-Boundary Non-Crossing Probability for a Class of Compound Processes with Applications, *European Journal of Operational Research*, forthcoming.

Chapter 4: *On Double Boundary Crossing and the Overshoot: Applications in Queueing, Ruin and Inventory.*

This chapter is based on the paper:

Dimitrova, D.S., Ignatov, Z.G., Kaishev, V.K., Tan, S. 2019. On Double Boundary Crossing and the Overshoot: Applications in Queueing, Ruin and Inventory, submitted to a peer reviewed journal.

Chapter 5: *On a Single Server Queueing Model and Its Double Boundary Crossing Duality.*

This chapter is based on the paper:

Dimitrova, D.S., Kaishev, V.K., Tan, S. 2019. On a Single Server Queueing Model and Its Double Boundary Crossing Duality, submitted to a peer reviewed journal.

Chapter 2

Computing the Kolmogorov-Smirnov Distribution when the Underlying CDF is Purely Discrete, Mixed or Continuous

This chapter is based on the paper:

Dimitrova, D.S., Kaishev, V.K., Tan, S. 2019. Computing the Kolmogorov-Smirnov Distribution when the Underlying CDF is Purely Discrete, Mixed or Continuous. *Journal of Statistical Software*, forthcoming.

Abstract

The distribution of the Kolmogorov-Smirnov (KS) test statistic has been widely studied under the assumption that the underlying theoretical cumulative distribution function (CDF), $F(x)$, is continuous. However, there are many real-life applications in which fitting discrete or mixed distributions is required. Nevertheless, due to inherent difficulties, the distribution of the KS statistic when $F(x)$ has jump discontinuities has been

studied to a much lesser extent and no exact and efficient computational methods have been proposed in the literature.

In this paper, we provide a fast and accurate method to compute the (complementary) CDF of the KS statistic when $F(x)$ is discontinuous, and thus obtain exact p values of the KS test. Our approach is to express the complementary CDF through the rectangle probability for uniform order statistics, and to compute it using fast Fourier transform (FFT). Secondly, we provide a C++ and an R implementation of the proposed method, which fills in the existing gap in statistical software. We give also a useful extension of the Schmid's asymptotic formula for the distribution of the KS statistic, relaxing his requirement for $F(x)$ to be increasing between jumps and thus allowing for any general mixed or purely discrete $F(x)$. The numerical performance of the proposed FFT-based method, implemented both in C++ and in the R package **KSgeneral**, available from https://CRAN.R-project.org/package=KSgeneral, is illustrated when $F(x)$ is mixed, purely discrete, and continuous. The performance of the general asymptotic formula is also studied.

2.1 Introduction

The two-sided Kolmogorov-Smirnov (KS) statistic is one of the most popular goodness-of-fit test statistics that is used to measure how well the distribution of a random sample $\{X_1,...,X_n\}$ agrees with a theoretical distribution. It is defined as

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{2.1}$$

where n is the sample size, $F_n(x)$ denotes the empirical (cumulative) distribution function (EDF) of $\{X_1,...,X_n\}$, and $F(x)$ denotes the cumulative distribution function (CDF) of a pre-specified theoretical distribution under the null hypothesis (H_0) that the sample $\{X_1,...,X_n\}$ comes from $F(x)$.

Many authors have studied the distribution of D_n, i.e., its CDF $\mathsf{P}(D_n \leq q|H_0), q \in [0,1]$ under the assumption that $F(x)$ is continuous. Kolmogorov (1933), Smirnov

(1939), Feller (1948), Doob (1949), and Smirnov (1948) considered the limiting distribution of D_n. Massey (1951) showed that the exact distribution of D_n is independent of $F(x)$ if $F(x)$ is continuous, and provided a table for exact critical levels of the KS test corresponding to certain significance levels for sample sizes $n \leq 35$. Durbin (1968) studied the probability that the EDF of an ordered sample of n independent observations from the uniform $(0,1)$ distribution lies between two parallel straight lines. He also obtained the exact distribution of D_n for $F(x)$ continuous, when the two parallel straight lines are $ny = a + nx$ and $ny = -a + nx$. Durbin (1968) also noted the important link between this probability and the double-boundary non-crossing probability for a Poisson process that is easier to compute. Epanechnikov (1968), Steck (1971), Noé (1972), Niederhausen (1981) obtained the exact distribution of D_n when $F(x)$ is continuous, by studying the probability that the order statistics of n uniform $[0,1]$ random variables all lie within an n-dimensional rectangle. For brevity, we will further refer to this probability as the rectangle probability for uniform order statistics. Numerically computing the distribution of D_n when $F(x)$ is continuous is not easy and has been recently considered by Marsaglia et al. (2003), Simard and L'Ecuyer (2011), Carvalho (2015), among others. Details related to these works and further references are provided in Section 2.3.3.

While performing KS tests when $F(x)$ is continuous is widely applicable, there are many real-life applications, e.g., in biology, physics, engineering, finance, and insurance, in which fitting discrete or mixed distributions, i.e., with multiple jumps and continuous segments, to large samples of data is required. For example, Calabrese and Zenga (2010) modeled the bank loan recovery rates using mixed random variables, since empirical data suggest that loans are either not repaid at all (recovery rate = 0), partially repaid (recovery rate between 0 and 1), or fully repaid (recovery rate = 1). This leads to considering a mixed CDF $F(x)$ with jumps at 0 and 1 and a continuous segment in between. It is important to accurately model bank loan recovery rates, because this is required by the Basel II solvency framework. Mixed distributions with multiple jumps arise also in reinsurance, in relation to fitting claim amount data in multi-layer excess-of-loss treaties. We consider such an example in Section 2.3.1. Furthermore, numerous risk modeling applications in (general) insurance, e.g., car insurance and

catastrophe insurance, require fitting appropriate discrete distributions to claim numbers data. The need to fit discrete distributions to data naturally arises also in almost any field of research in science and economics. In all such cases, the underlying CDF $F(x)$ has discontinuities at some points and it is important to be able to perform goodness-of-fit tests, such as the chi-squared test and the KS test. As demonstrated by Pettitt and Stephens (1977), the KS test for discrete distributions can have greater power than the chi-squared test. On the other hand, Noether (1963), Slakter (1965), and Walsh (1963) showed that conducting a discontinuous KS test is more conservative than conducting a continuous KS test in terms of accepting/rejecting the null hypothesis. Thus, as we illustrate in Section 2.3.1, a null hypothesis that a sample comes from a discontinuous distribution will be accepted more often if one uses the continuous KS test, as opposed to using the discontinuous KS test. It should also be noted that the sample size in many applications can be substantial. Therefore, it is important to accurately and efficiently perform KS tests for $F(x)$ with discontinuities, when sample sizes are large. For the purpose, one needs to be able to efficiently and accurately compute probabilities of the type, $\mathsf{P}(D_n \geq q)$, known as the complementary CDF, for any values of n and q, $q \in [0,1]$. Addressing this problem is the main objective of this paper.

The distribution of the KS test statistic D_n in this more general case, when $F(x)$ may have jump discontinuities (including purely discrete $F(x)$), has been studied to a much lesser extent. In an early paper, Schmid (1958) found the limiting distribution of D_n when $F(x)$ has countable number of jumps and is increasing between the jumps. Carnal (1962) has generalized Schmid (1958)'s formula by allowing constant segments between jumps. Conover (1972) provided an approach to finding the exact critical level for the one-sided KS test statistics $D_n^- = \sup_x(F(x) - F_n(x))$ and $D_n^+ = \sup_x(F_n(x) - F(x))$ for discontinuous $F(x)$. Approximated critical levels for the two-sided KS test statistic D_n were also provided. Gleser (1985) studied the exact power of two-sided KS tests. He showed that existing algorithms designed for KS tests with continuous $F(x)$ could be used (after some necessary adjustments) for KS tests when $F(x)$ is discontinuous. Specifically, Gleser (1985) showed that the power of the KS test when $F(x)$ has jump discontinuities could still be expressed as a rectangle probability with

respect to uniform order statistics, but with modified non-linear boundaries. Therefore, the determinantal and recurrence formulae for the latter rectangle probability due to Steck (1971), Noé (1972) and Niederhausen (1981) could be applied in order to obtain the exact distribution of D_n when $F(x)$ is discontinuous. However, implementing these results is computationally expensive, especially when the sample size is large, and may lead to numerical instabilities, as noted by some authors and also illustrated in Section 2.3.2.

In summary, computing the distribution of D_n when $F(x)$ is discontinuous is even harder and much less explored than in the continuous case. To the best of our knowledge, no methods have been proposed in the literature to compute the exact distribution of D_n when $F(x)$ is mixed. Looking at the statistical software literature, all major packages implement the KS test only when $F(x)$ is continuous, see for example, the `ks.test` function of the package **stats** (R Core Team, 2018) and `ks.test.imp` function of the package **kolmim** (Carvalho, 2015) in R (R Core Team, 2018), SPSS (IBM Corp., 2017), `ksmirnov` function in Stata (StataCorp., 2017), the `kstest` function in MATLAB (The MathWorks Inc., 2018), the `KolmogorovSmirnovTest` function in Mathematica (Wolfram Research, Inc., 2018).

There is one exception, Arnold and Emerson (2011) provide the R function `ks.test` as part of the package **dgof** that calculates exact p values of the KS test assuming $F(x)$ is purely discrete. In `ks.test` function, a one-sided KS p value is calculated by combining the approaches of Conover (1972) and Niederhausen (1981), while two-sided KS p values are calculated by combining the approaches of Gleser (1985) and Niederhausen (1981). However, the `ks.test` function due to Arnold and Emerson (2011) only provides exact p values for sample sizes less than or equal to 30, since as noted by the authors, when the sample size is large, numerical instabilities may occur. In the latter case, Arnold and Emerson (2011) suggest using simulation to approximate p values, which as we show in Section 2.3.2, is rather slow and inaccurate.

Our aim in this paper is two-fold. The first goal is to provide a fast and accurate method to compute $\mathsf{P}(D_n \geq q)$ when $F(x)$ is discontinuous (i.e., mixed or purely discrete), and thus obtain exact p values of the KS test for any (small or large) sample

size n, and any $q \in [0,1]$, possibly close to 1. Our second goal is to give the C++ code and an R package **KSgeneral**, based on the C++ code that implements this fast and accurate method, which we believe fills in the gap in the existing statistical software. As we will see, the proposed method is also applicable and highly competitive when $F(x)$ is continuous. The approach we take, described in Section 2.2.1, is to express $\mathsf{P}(D_n \geq q)$ as an appropriate rectangle probability for uniform order statistics, as noted by Gleser (1985), and to compute the latter probability using the fast Fourier transform (FFT) method. FFT has been recently utilized by Moscovich and Nadler (2017) to calculate this rectangle probability when $F(x)$ is continuous. Furthermore, in Section 2.2.2, we provide a useful extension (cf., (2.15) and (2.20)) of Schmid (1958)'s asymptotic formula, relaxing his requirement for $F(x)$ to be increasing between jumps and thus allowing for any general mixed or purely discrete $F(x)$. Similar formula has been obtained by Carnal (1962), but the embedded implicit index structure makes its numerical implementation prohibitive. In Section 2.3, we illustrate the C++ and the R implementation as the package **KSgeneral** of the proposed FFT-based method. In particular, in Section 2.3.1, we study its numerical properties based on some mixed (inflated) distributions and also illustrate the performance of the general asymptotic formula (2.15). We show in Section 2.3.2 that when $F(x)$ is purely discrete, our approach to computing $\mathsf{P}(D_n \geq q)$, based on FFT and the asymptotic formula (2.22), outperforms in terms of speed and accuracy the R function of Arnold and Emerson (2011), especially for large sample sizes. Finally, in Section 2.3.3, we consider the case of continuous $F(x)$ and compare with the state-of-the-art procedures of Simard and L'Ecuyer (2011) and Carvalho (2015).

2.2 Distribution of D_n when $F(x)$ is discontinuous

It is well known that the distribution of D_n does not depend on $F(x)$ when the latter CDF is continuous. To see this, note that

$$\begin{aligned} D_n &= \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = \sup_{0\le t\le 1} \left|F_n(F^{-1}(t)) - F(F^{-1}(t))\right|, \\ &= \sup_{0\le t\le 1} \left|F_n(F^{-1}(t)) - t\right| = \sup_{0\le t\le 1} |U_n(t) - t|, \end{aligned} \tag{2.2}$$

where $F^{-1}(t) \equiv \inf\{x : F(x) \ge t\}, t \in [0,1]$, and $U_n(t)$ is the empirical CDF of the uniform random sample $\{U_i = F(X_i), i = 1,...,n\}$. In this section, we relax the assumption of continuity of $F(x)$ and assume that $F(x)$ is non-decreasing and right-continuous, with countable (possibly infinite) number of jumps. From the right-continuity of $F(x)$, it follows that $F(F^{-1}(t)) \ge t$ and $F^{-1}(F(x)) \le x$ and hence, the distribution-free property, illustrated by (2.2) is no longer valid. Therefore, it becomes difficult to compute the exact and asymptotic distributions of D_n. This problem is addressed in the next two sections.

2.2.1 The exact distribution of D_n

Our approach to computing the exact distribution of D_n is based on the following four major steps:

Step 1. It is not difficult to show (see Appendix A.1) that the complementary CDF $\mathsf{P}(D_n \ge q), q \in [0,1]$, can be expressed in terms of a rectangle probability for the vector of n uniform order statistics as

$$\mathsf{P}(D_n \ge q) = 1 - \mathsf{P}\left(A_i \le U_{(i)} \le B_i,\ 1 \le i \le n\right), \tag{2.3}$$

where

$$\begin{aligned} A_i &= \lim_{\varepsilon \downarrow 0} F\Big(\Big(F^{-1}\big(\frac{i}{n} - q + \varepsilon\big)\Big) - \Big), \\ F(x-) &= \lim_{z \uparrow x} F(z) = \mathsf{P}(X < x), \\ B_i &= \lim_{\varepsilon \downarrow 0} F\Big(F^{-1}\big(\frac{i-1}{n} + q - \varepsilon\big)\Big),\ i = 1,2,...,n, \end{aligned} \tag{2.4}$$

and where $U_{(i)}$, $i = 1,...,n$, are the order statistics of n independent and identically distributed uniform $(0,1)$ random variables $U_i, i = 1,2,...,n$.

Step 2. Express the rectangle probability on the right hand side of (2.3) in terms of the double-boundary non-crossing probability with respect to the empirical process $\eta_n(t) = nU_n(t) = \sum_{i=1}^n \mathbb{1}(U_i \le t), 0 \le t \le 1$, where $U_n(t)$ is the EDF of the sample $\{U_1,...,U_n\}$. In particular, it can be directly verified that (2.3) can be rewritten as

$$\begin{aligned} \mathsf{P}(D_n \ge q) &= 1 - \mathsf{P}\left(A_i \le U_{(i)} \le B_i,\ 1 \le i \le n\right), \\ &= 1 - \mathsf{P}\left(g(t) \le \eta_n(t) \le h(t), \forall\ 0 \le t \le 1\right), \end{aligned} \tag{2.5}$$

where the upper and lower boundary functions $h(t)$, $g(t)$ are defined as

$$h(t) = \sum_{i=1}^{n} \mathbb{1}_{(A_i < t)},\ \ g(t) = \sum_{i=1}^{n} \mathbb{1}_{(B_i \le t)}. \tag{2.6}$$

Let us note that $h(t)$ and $g(t)$ are correspondingly left and right continuous functions which equivalently satisfy the following conditions

$$\sup\{t \in [0,1] : h(t) < i\} = A_i, \text{ and } \inf\{t \in [0,1] : g(t) > i-1\} = B_i, \tag{2.7}$$

with A_i, B_i defined in (2.4)[1]. The last equality in (2.5) is illustrated in Figure 2.1, where one can see that considering the rectangle probability with respect to the uniform order statistics, $\mathsf{P}\left(A_i \le U_{(i)} \le B_i,\ 1 \le i \le n\right)$ is equivalent to considering the non-exit probability, $\mathsf{P}\left(g(t) \le \eta_n(t) \le h(t), \forall\ 0 \le t \le 1\right)$.

[1] An expression similar to (2.5) for the case of $\mathsf{P}(D_n > q)$ has been obtained by Gleser (1985) (cf., Theorem 2 therein).

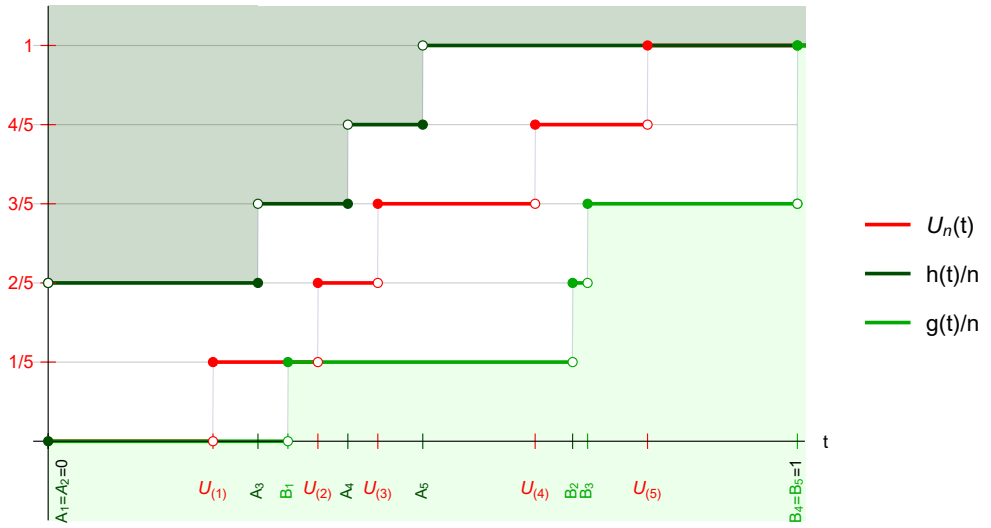


Fig. 2.1 Illustration of the fact that the non-exit probability, $\mathsf{P}\left(A_i \le U_{(i)} \le B_i,\ 1 \le i \le n\right)$ is equivalent to the non-exit probability, $\mathsf{P}\left(g(t) \le \eta_n(t) \le h(t), \forall\, 0 \le t \le 1\right)$, where $g(t)$ and $h(t)$ are defined as in (2.6) using $F(x)$ given in (2.23) (cf., Example 2.2.8), with $n = 5$.

Step 3. Use the fact that the process $\eta_n(t)$, $t \in [0,1]$, has the same distribution as the conditional distribution of a Poisson process with intensity n, denoted by $\xi_n(t) : [0,1] \mapsto \{0,1,2...\}$, given $\xi_n(1) = n$, (see e.g., Shorack and Wellner, 1986, Chapter 8, Proposition 2.2). Therefore, the non-crossing probability in (2.5) can be re-expressed as

$$\begin{aligned}
&\mathsf{P}(g(t) \le \eta_n(t) \le h(t), \forall\, 0 \le t \le 1) \\
&= \mathsf{P}(g(t) \le \xi_n(t) \le h(t) | \xi_n(1) = n, \forall\, 0 \le t \le 1) \\
&= \frac{\mathsf{P}(g(t) \le \xi_n(t) \le h(t) \text{ and } \xi_n(1) = n, \forall\, 0 \le t \le 1)}{\mathsf{P}(\xi_n(1) = n)} = \frac{Q(1,n)}{e^{-n} n^n / n!},
\end{aligned} \tag{2.8}$$

where $\xi_n(1)$ follows a Poisson(n) distribution and $Q(1,n)$ is defined as in (2.9). It is not difficult to see that in order to compute the non-crossing probability $\mathsf{P}(g(t) \le \xi_n(t) \le h(t)$ and $\xi_n(1) = n, \forall\, 0 \le t \le 1)$ on the right-hand-side of (2.8), defined on a continuum of times $t \in [0,1]$, it suffices to consider the events of non-crossing only over some fixed times, $0 = t_0 < t_1 < t_2 < \ldots < t_N = 1$, which are the ordered set of all distinct points in $\{1, A_i, B_i, i = 1, \ldots, n\}$, where A_i and B_i

are specified in (2.4) (and (2.7)). Based on this discretization, similarly as done by Khmaladze and Shinjikashvili (2001) and Moscovich and Nadler (2017) in the continuous case, the non-crossing probability in (2.8) can be calculated by solving recursively an appropriate system of Chapman-Kolmogorov forward equations[2]. In order to introduce these equations, for any $s \in [0,1]$ and $m \in \{0,1,2,...,\}$, let

$$Q(s,m) = \mathsf{P}(g(t) \leq \xi_n(t) \leq h(t),\ \forall t \in [0,s] \text{ and } \xi_n(s) = m), \tag{2.9}$$

where $g(s) \leq m \leq h(s)$ and $Q(0,0) = \mathsf{P}(g(0) \leq 0 \leq h(0)) = 1$ by assumption. For any $j \in \{0,1,...,N-1\}$ and any $m \in \{0,1,2,...\}$, the Chapman-Kolmogorov equations are

$$Q(t_{j+1},m) = \begin{cases} \sum\limits_{g(t_j) \leq l \leq m} Q(t_j,l)\mathsf{P}(Y_j = m-l), \text{ if } g(t_{j+1}) \leq m \leq h(t_{j+1}), \\ 0, \qquad \text{otherwise,} \end{cases} \tag{2.10}$$

where Y_j denotes a Poisson random variable with parameter $n(t_{j+1} - t_j)$. The required non-crossing probability is obtained by computing $Q(1,n)$ following (2.10). This is illustrated by Figure 2.2, where $g(t)$ and $h(t)$ are obtained based on (2.6), with $F(x)$ defined in (2.23) as part of Example 2.2.8. The black dots illustrate the mesh of points (t_{j+1},m), $j = 0,1,...,6$, $m = 0,1,2,...,5$, at which non-crossing of the trajectory of $\xi_n(t)$ with the boundaries $g(t)$, $h(t)$ may occur and the corresponding probabilities, $Q(t_{j+1},m)$ need to be computed, following (2.10).

As shown by Khmaladze and Shinjikashvili (2001), the recurrent computation following (2.10) requires total running time of order at most $\mathscr{O}(n^3)$. In the next step we employ FFT in order to improve this rate.

[2]Both Khmaladze and Shinjikashvili (2001) and Moscovich and Nadler (2017) assume $F(x)$ is continuous and consider strict inequalities in (2.8) i.e., they do not allow the process to touch the boundaries.

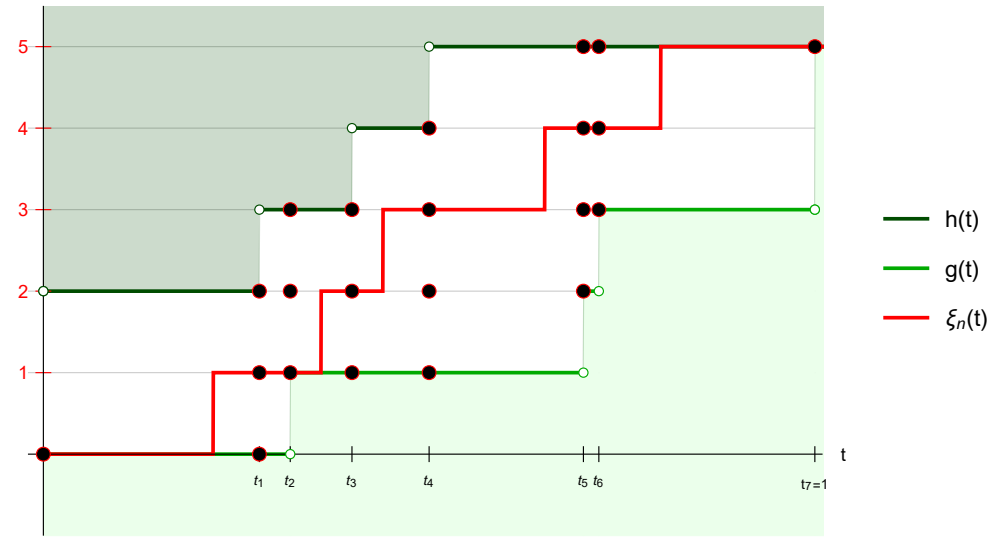


Fig. 2.2 Illustration of a trajectory of the Poisson process $\xi_n(t)$ staying in the corridor between the boundaries $h(t)$ and $g(t)$ defined as in (2.6) using $F(x)$ given in (2.23) (cf., Example 2.2.8). The black dots illustrate the mesh of points (t_{j+1}, m), $j = 0, 1, ..., 6$, $m = 0, 1, 2, ..., 5$, at which non-crossing of the trajectory of $\xi_n(t)$ with the boundaries $g(t)$, $h(t)$ may occur and the corresponding probabilities, $Q(t_{j+1}, m)$ need to be computed, following (2.10).

Step 4. Apply FFT to compute the truncated linear convolution of the vectors $\mathbf{Q}_{t_j} = (Q(t_j,0), Q(t_j,1), ..., Q(t_j,n))$ and $\boldsymbol{\pi}_{n(t_{j+1}-t_j)} = (\mathsf{P}(Y_j = 0), \mathsf{P}(Y_j = 1), ..., \mathsf{P}(Y_j = n))$ in order to solve (2.10), as proposed by Moscovich and Nadler (2017), see Section 2 therein. As shown by these authors, the total running time of this method is of order at most $\mathscr{O}(n^2 \log n)$, which is faster than $\mathscr{O}(n^3)$ especially for large n.

In summary, our approach to computing the exact $\mathsf{P}(D_n \geq q)$ when $F(x)$ is discontinuous is outlined in the following procedure (Procedure Exact-KS-FFT).

(i) Specify a discontinuous CDF $F(x)$, a sample size n, and a quantile q.

(ii) As detailed in Step 1, compute A_i and B_i for $i = 1, ..., n$, based on (2.4), where the limites are coded using a very small ε, e.g., $\varepsilon = 10^{-10}$.

(iii) As detailed in Step 2, compute the upper and lower boundaries $g(t)$, $h(t)$ using (2.6).

(iv) Following Steps 3 and 4, apply FFT to compute $Q(1,n)$ defined in (2.10). Hence, calculate the double-boundary non-crossing probability with respect to the Poisson process on the right-hand-side of (2.8) and respectively obtain the double-boundary non-crossing probability with respect to $\eta_n(t)$ on the left-hand-side of (2.8).

(v) Finally, compute the exact $\mathsf{P}(D_n \geq q)$ using (2.5) (cf., Steps 2 and 3).

Remark 2.2.1. Let us note that $\mathsf{P}(D_n \geq q)$, $0 \leq q \leq 1$, can directly be computed using (2.3) and (2.4), applying the determinantal formula for the rectangle probability in (2.3), due to Steck (1971), or the recurrence formula of Niederhausen (1981). However, such computations are slow, and may become unstable for sample sizes $n \geq 100$, as shown in Section 2.3.2, Example 2.3.5. We also note that $\mathsf{P}(D_n \geq q)$ is the p value corresponding to a fixed critical level $q \in [0,1]$. Thus, if $q = d_n$, where d_n is the value of the KS test statistic computed based on a sample $\{x_1,...,x_n\}$, then the corresponding exact p value, $\mathsf{P}(D_n \geq d_n)$ can be obtained through (2.3) and (2.4).

Remark 2.2.2. We have described the Procedure Exact-KS-FFT for computing the complementary CDF of the two-sided KS statistic, D_n, defined in (2.1). It should be noted that by selecting the lower boundary $g(t) \equiv 0, \forall t$, and the upper boundary $h(t)$ as specified in (2.6) one can compute the complementary CDF for the one-sided KS statistic $D_n^+ = \sup_x(F_n(x) - F(x))$. By selecting the upper boundary $h(t) \equiv n, \forall t$, and the lower boundary $g(t)$ as specified in (2.6), one can compute the complementary CDF for the one-sided KS statistic $D_n^- = \sup_x(F(x) - F_n(x))$ (see e.g., Gleser, 1985). For the sake of consistency, in what follows, we illustrate the proposed FFT-based method for the two-sided version of the KS statistic.

As noted and also demonstrated in Section 2.3, the proposed FFT-based method for computing exact $\mathsf{P}(D_n \geq q)$ is highly numerically efficient and could be easily applied to sample sizes n up to hundreds of thousands (see also Moscovich and Nadler, 2017). Nevertheless, it is still beneficial to know the asymptotic distribution of D_n as $n \to \infty$, since as demonstrated in Section 2.3, it can be efficiently applied to approximate $\mathsf{P}(D_n \geq q)$ for large and even moderate sample sizes and hypothesized distributions

with small number of jumps. The asymptotic distribution of D_n will be considered in the next section.

2.2.2 The asymptotic distribution of D_n

Schmid (1958) has studied the asymptotic distribution of the form

$$\Phi(\lambda) = \lim_{n\to\infty} \mathsf{P}(D_n < \lambda n^{-\frac{1}{2}}) = \lim_{n\to\infty} \mathsf{P}(\sup_{-\infty<x<\infty} |F_n(x) - F(x)| < \lambda n^{-\frac{1}{2}}), \qquad (2.11)$$

where n denotes the sample size, and $F(x)$ is a CDF with countable number of jumps J, at $x = x_l$, $l = 1,2,...,J$ and increasing continuous segments between the jumps. Let $F(x_l-) = f_{2l-1}, F(x_l) = f_{2l}$, $l = 1,2,...,J$, with $f_0 = 0, f_{2J+1} \equiv 1$, and $f_{2l} < f_{2l+1}$, $l = 0,...,J$. Under these assumptions on $F(x)$, Theorem 1 of Schmid (1958) states that

$$\Phi(\lambda) = \sum_{j_1=-\infty}^{\infty} \cdots \sum_{j_{J+1}=-\infty}^{\infty} (-1)^{j_1+\cdots+j_{J+1}}$$
$$\times c \int_{-\lambda}^{\lambda} \cdots \int_{-\lambda}^{\lambda} \exp\Big[-\frac{1}{2}\sum_{l=1}^{J} \frac{(z_{2l} - z_{2l-1})^2}{f_{2l} - f_{2l-1}} - \frac{1}{2}\sum_{l=0}^{J} \frac{(z_{2l+1} - (-1)^{j_{l+1}} z_{2l} - 2\lambda j_{l+1})^2}{f_{2l+1} - f_{2l}}\Big] dz_1 \cdots dz_{2J}, \qquad (2.12)$$

where

$$z_0 = z_{2J+1} = 0, \text{and } c = (2\pi)^{-J} \prod_{j=1}^{2J+1} (f_j - f_{j-1})^{-1/2}.$$

In view of (2.12), when the sample size n is large, the limiting $\mathsf{P}(D_n \geq q)$ for mixed $F(x)$ can be calculated as

$$\lim_{n\to\infty} \mathsf{P}(D_n \geq q) = 1 - \Phi(\lambda), \qquad (2.13)$$

where $\Phi(\lambda)$ is expressed as in (2.12), and $\lambda = qn^{\frac{1}{2}}$. However, Schmid's formula cannot be applied if the condition $f_{2l} < f_{2l+1}, l = 0,...,J$ is not satisfied, since there will be division by 0 in the second denominator in (2.12). Therefore, (2.12) is not applicable if $F(x)$ has constant segments between (some of) the jumps, as is the case when $F(x)$ is purely discrete, or if $F(x)$ starts (ends) with a jump at 0 (at 1), as is the case for zero-inflated (mixed) distributions. Carnal (1962) has generalized Schmid (1958)'s

formula to the case of arbitrary discontinuous $F(x)$ with finite number of jumps (cf., expression (5.1) therein). However, there is notational ambiguity (e.g., in the fourth summation in (5.1)) and because the embedded index structure is rather implicit, it is not straightforward to implement formula (5.1) numerically. Therefore, in what follows, we will derive an alternative formula for $\Phi(\lambda)$, for any discontinuous $F(x)$ with finite number of jumps (see Proposition 2.2.3). The latter formula may look cumbersome, but as we will see, it is notationally explicit and therefore easier to implement numerically. In addition, we believe that the clearer and more intuitive proof of Proposition 2.2.3 will facilitate better understanding of the structure underlying (2.15). However, one should note that formula (2.15) (respectively (2.20) and (2.22)) is only practically implementable for small/moderate number of jumps, J, in the null distribution, as otherwise the multidimensional integration becomes infeasible.

It is not difficult to see that any jump structure in $F(x)$ can be represented through only two different types of continuous segments of $F(x)$ followed by jumps. The first one is a segment of $F(x)$ increasing on $[x_{l-1}, x_l-]$, i.e., $f_{2l-2} < f_{2l-1}$, followed by a jump at x_l, and the second one is a constant segment of $F(x)$ on $[x_{l-1}, x_l-]$, i.e., $f_{2l-2} = f_{2l-1}$, followed by a jump at x_l. We will refer to these two types of segments as increasing-jump segment and flat-jump segment, respectively.

We will use the notation $\nu_1, \nu_2, ...$ to denote the sizes of groups of consecutive increasing-jump segments, i.e., ν_i denotes the number of consecutive jumps, preceded by an increasing segment, in the i^{th} group. Similarly, by ω_k, $k = 1, 2, ...$, we denote the number of consecutive jumps preceded by a flat segment, in the k^{th} group. Without loss of generality, we assume that there are m groups of increasing-jump and flat-jump segments, i.e., $\nu_1, ..., \nu_m$ and $\omega_1, ..., \omega_m$, and that these groups of jumps points, x_l, appear in the CDF in the following order:

$$\begin{aligned}\Big\{&x_1, \ldots, x_{\nu_1}, x_{\nu_1+1}, \ldots, x_{\nu_1+\omega_1}, x_{\nu_1+\omega_1+1}, \ldots, x_{\nu_1+\omega_1+\nu_2}, x_{\nu_1+\omega_1+\nu_2+1}, \ldots, \\ &x_{\nu_1+\omega_1+\nu_2+\omega_2}, \ldots, x_{\nu_1+\omega_1+\cdots+\omega_{m-1}+1}, \ldots, x_{\nu_1+\omega_1+\cdots+\omega_{m-1}+\nu_m}, \\ &x_{\nu_1+\omega_1+\cdots+\omega_{m-1}+\nu_m+1}, \ldots, x_{\nu_1+\omega_1+\cdots+\omega_{m-1}+\nu_m+\omega_m}\Big\},\end{aligned} \tag{2.14}$$

where $\nu_1+\omega_1+\cdots+\nu_m+\omega_m=J$ is the total number of jumps in $F(x)$, and

$$\nu_1\geq 0;\omega_1\geq 0;\nu_1+\omega_1>0;\nu_l>0,2\leq l\leq m;\omega_l>0,2\leq l\leq m-1;\omega_m\geq 0;\nu_m+\omega_m>0.$$

It can be seen that (2.14) covers any possible order of the jumps of different type in $F(x)$ as illustrated on some examples below (see e.g., Corollary 2.2.6 and Example 2.2.8). Under these general assumptions on $F(x)$, in the following proposition we give a formula for $\Phi(\lambda)$ which generalizes (2.12).

Proposition 2.2.3. *Assuming that a CDF $F(x)$ has the structure of jumps as in* (2.14) *and that $f_{2J}=f_{2J+1}\equiv 1$, we have*

$$\Phi(\lambda)=\sum_{j_1=-\infty}^{\infty}\cdots\sum_{j_{\nu_m}=-\infty}^{\infty}\left((-1)^{j_1+\cdots+j_{\nu_m}}\right)c\int_{-\lambda}^{\lambda}\cdots\int_{-\lambda}^{\lambda}\exp\{\psi\}dz_1\cdots dz_{2\nu_m+w_m-1}, \tag{2.15}$$

where

$$\begin{aligned}c=\prod_{i=1}^{m}\Bigg(\prod_{l=1}^{\nu_i}&\left(f_{2(\nu_{i-1}+w_{i-1}+l)-1}-f_{2(\nu_{i-1}+w_{i-1}+l)-2}\right)^{-1/2}\\&\left(f_{2(\nu_{i-1}+w_{i-1}+l)}-f_{2(\nu_{i-1}+w_{i-1}+l)-1}\right)^{-1/2}\Bigg)\\&\times\left(\prod_{l=1}^{\omega_i}\left(f_{2(\nu_i+w_{i-1}+l)}-f_{2(\nu_i+w_{i-1}+l)-1}\right)^{-1/2}\right)(2\pi)^{-\frac{2\nu_m+w_m-1}{2}},\end{aligned} \tag{2.16}$$

and

$$\begin{aligned}\psi=-\frac{1}{2}\sum_{i=1}^{m}\Bigg\{\sum_{l=1}^{\nu_i}\Bigg[&\frac{\left(z_{2(\nu_{i-1}+l)+w_{i-1}}-z_{2(\nu_{i-1}+l)+w_{i-1}-1}\right)^2}{f_{2(\nu_{i-1}+w_{i-1}+l)}-f_{2(\nu_{i-1}+w_{i-1}+l)-1}}\\&+\frac{\left(z_{2(\nu_{i-1}+l)+w_{i-1}-1}-(-1)^{j_{(\nu_{i-1}+l)}}z_{2(\nu_{i-1}+l)+w_{i-1}-2}-2\lambda j_{(\nu_{i-1}+l)}\right)^2}{f_{2(\nu_{i-1}+w_{i-1}+l)-1}-f_{2(\nu_{i-1}+w_{i-1}+l)-2}}\Bigg]\\&+\sum_{l=1}^{\omega_i}\left[\frac{\left(z_{2\nu_i+w_{i-1}+l}-z_{2\nu_i+w_{i-1}+l-1}\right)^2}{f_{2(\nu_i+w_{i-1}+l)}-f_{2(\nu_i+w_{i-1}+l)-1}}\right]\Bigg\},\end{aligned} \tag{2.17}$$

with $\nu_0=\omega_0=0;v_0=w_0=0;v_i=\sum_{k=1}^{i}\nu_k;w_i=\sum_{k=1}^{i}\omega_k,v_m+w_m=J$, *and* $z_0=z_{2\nu_m+w_m}=0$.

Proof: The reasoning in the proof follows that of Schmid (1958) with some necessary adjustments to account for the fact that $f_{2l} \leq f_{2l+1}$ as opposed to $f_{2l} < f_{2l+1}$, $l = 0, ..., J$. So, here we only give details related to those parts of the proof which are affected by the relaxed assumption on $F(x)$. Thus, following Schmid (1958), page 1014, denote by I the union of the closed intervals $[f_{2l}, f_{2l+1}]$, $l = 0, ..., J$ and let M_n be the set of integers j such that $j/n \in I$,

$$M_n = \left\{k_0 = 0, \ldots, k_1; k_2, k_2 + 1, \ldots, k_3; \ldots; k_{2J}, k_{2J} + 1, \ldots, k_{2J+1} = n\right\},$$

where k_i is such that $k_i/n \to f_i$, as $n \to \infty$. Note that if $f_{2l} = f_{2l+1}$ i.e., if we have a constant segment in the CDF $F(x)$, then $k_{2l} \equiv k_{2l+1}$ and both are included in the set M_n. Now, as demonstrated by Schmid (1958) (see expressions (20), (21) therein), the probability $\mathsf{P}_{0n} := \mathsf{P}(D_n < \lambda n^{-1/2})$ in (2.11), can be calculated as

$$\mathsf{P}_{0n} = \frac{n! e^n}{n^n} R_{0n},$$

where

$$R_{ik_{2l+1}} = \sum_{|j| < \lambda N^{1/2}} R_{jk_{2l}} \mathsf{P}[\mathscr{D}_{ik_{2l+1}} | \mathscr{D}_{ik_{2l}}], \quad l = 0, \ldots, J, \tag{2.18}$$

and

$$R_{ik_{2l}} = \sum_{|j| < \lambda N^{1/2}} R_{jk_{2l-1}} \frac{(k_{2l} - k_{2l-1})^{i-j+k_{2l}-k_{2l-1}}}{(i - j + k_{2l} - k_{2l-1})! e^{k_{2l}-k_{2l-1}}}, \quad l = 0, \ldots, J, \tag{2.19}$$

and $R_{00} = 1, R_{i0} = 0$ for $i \neq 0$. Note that recursion (2.19) is related to the l^{th} jump in $F(x)$, whereas recursion (2.18) is related to the continuous (increasing or flat) segment on $[x_l, x_{l+1}-]$ in $F(x)$. The events $\mathscr{D}_{ik}$ are specified in details in Schmid (1958) (see page 1016), but what is important here is to observe that when $k_{2l} = k_{2l+1}$ in M_n, $\mathsf{P}[\mathscr{D}_{ik_{2l+1}} | \mathscr{D}_{ik_{2l}}] = \mathbb{1}_{(i=j)}$. Thus, for a constant segment on $[x_l, x_{l+1}-]$ in $F(x)$, we have $R_{ik_{2l+1}} = R_{ik_{2l}}$ and so, recursion (2.18) is obsolete. Therefore, asymptotically, when $k_{2l} = k_{2l+1}$, we only need to consider the convergence of recursion (2.19) for a flat-jump segment in $F(x)$, whereas for increasing-jump segment, both recursions (2.18)

and (2.19) generate terms in the resulting expression, in particular (2.16) and (2.17). Now, applying the asymptotic arguments outlined on page 1018 of Schmid (1958), one easily obtains formula (2.15). □

Let us note that Proposition 2.2.3 does not cover the case when $f_{2J} < f_{2J+1} \equiv 1$. This case is addressed in the following proposition, which follows by similar reasoning.

Proposition 2.2.4. *Assuming that a CDF $F(x)$ has the structure of jumps as in* (2.14) *and that $f_{2J} < f_{2J+1} \equiv 1$, $v_m + w_m = J$, we have*

$$\Phi(\lambda) = \sum_{j_1=-\infty}^{\infty} \cdots \sum_{j_{v_m}=-\infty}^{\infty} \sum_{j_{v_m+1}=-\infty}^{\infty} \Big((-1)^{j_1+\cdots+j_{v_m}+j_{v_m+1}}\Big)\, c' \int_{-\lambda}^{\lambda} \cdots \int_{-\lambda}^{\lambda} \exp\{\psi'\} dz_1 \cdots dz_{2v_m+w_m}, \tag{2.20}$$

where

$$c' = c(f_{2J+1} - f_{2J})^{-1/2}(2\pi)^{-1/2}, \quad \textit{and} \quad \psi' = \psi + \frac{(-(-1)^{j_{v_m+1}} z_{2v_m+w_m} - 2\lambda j_{v_m+1})^2}{f_{2J+1} - f_{2J}}, \tag{2.21}$$

with c and ψ in (2.21) *defined in* (2.16), (2.17), *noting that $z_{2v_m+w_m} \neq 0$.*

Remark 2.2.5. Let us note that (2.12) is a special case of (2.20) when $m = 1, \omega_1 \equiv w_1 = 0, \nu_1 \equiv v_1 = J$.

Corollary 2.2.6. *When $F(x)$ is purely discrete with J jumps, the limiting distribution $\Phi(\lambda)$ in* (2.15) *becomes*

$$\Phi(\lambda) = (2\pi)^{-\frac{J-1}{2}} \prod_{l=1}^{J} (f_{2l} - f_{2l-1})^{-\frac{1}{2}} \int_{-\lambda}^{\lambda} \cdots \int_{-\lambda}^{\lambda} \exp\Big[-\frac{1}{2}\Big(\sum_{l=1}^{J} \frac{(z_l - z_{l-1})^2}{f_{2l} - f_{2l-1}}\Big)\Big] dz_1 \cdots dz_{J-1}, \tag{2.22}$$

where $z_0 = z_J = 0$.

Proof: Since the jump structure in this case includes only one group of flat-jump segments of size J, the first group of increasing-jump segments in (2.14) is empty, i.e., $m = 1, \nu_1 \equiv v_1 = 0$, $\omega_1 \equiv w_1 = J$, and by convention, $\prod_{l=1}^{\nu_1=0}(\cdot) = 1$, $\sum_{l=1}^{\nu_1=0}(\cdot) = 0$. Substituting these in (2.15), (2.16), and (2.17), we have

$$c = (2\pi)^{-\frac{J-1}{2}} \prod_{l=1}^{J} (f_{2l} - f_{2l-1})^{-\frac{1}{2}}, \quad \psi = -\frac{1}{2}\Big(\sum_{l=1}^{J} \frac{(z_l - z_{l-1})^2}{f_{2l} - f_{2l-1}}\Big),$$

and (2.15) becomes (2.22). □

Remark 2.2.7. It should be noted that (2.22) is the formula for the distribution of an $(J-1)$ dimensional Brownian bridge between $-\lambda$ and λ. The Brownian bridge interpretation has been used by Wood and Altavela (1978) to compute via Monte Carlo (MC) simulation the asymptotic distribution of D_n, without relating the interpretation to an explicit expression such as (2.22).

Next, we give an illustrative example on how to use the asymptotic distribution formula (2.15) given by Proposition 2.2.3, for mixed $F(x)$. Similarly, one can employ expressions (2.20) and (2.22) on appropriate specific examples.

Example 2.2.8. Consider a random variable X with CDF

$$F(x) = \begin{cases} 0 \text{ if } x < 0, \\ 0.2 + x \text{ if } 0 \le x < 0.2, \\ 0.5 \text{ if } 0.2 \le x < 0.8, \\ x - 0.1 \text{ if } 0.8 \le x < 1, \\ 1 \text{ if } x \ge 1. \end{cases} \tag{2.23}$$

Clearly, $F(x)$ is a CDF with four jumps, i.e., $J = 4$, at $x_1 = 0$, $x_2 = 0.2$, $x_3 = 0.8$, $x_4 = 1.0$, and $f_0 = f_1 = 0$, $f_2 = 0.2$, $f_3 = 0.4$, $f_4 = f_5 = 0.5$, $f_6 = 0.7$, $f_7 = 0.9$, $f_8 = f_9 = 1$. Since the jump structure of $F(x)$ in (2.23) is flat-jump, increasing-jump, flat-jump, increasing-jump segments, the first set of increasing-jump segments and the last set of flat-jump segments in (2.14) should be omitted. Therefore, $m = 3$, $\nu_1 = 0$, $\omega_1 = 1$, $\nu_2 = 1$, $\omega_2 = 1$, $\nu_3 = 1$, $\omega_3 = 0$, and $v_0 = 0$, $v_1 = 0$, $v_2 = 1$, $v_3 = 2$, $w_0 = 0$, $w_1 = 1$, $w_2 = 2$, $w_3 = 2$. Substituting these in (2.15), (2.16), and (2.17), we obtain

$$\Phi(\lambda) = \sum_{j_1=-\infty}^{\infty} \sum_{j_2=-\infty}^{\infty} c(-1)^{j_1+j_2} \int_{-\lambda}^{\lambda} \cdots \int_{-\lambda}^{\lambda} \exp\{\psi\} dz_1 \cdots dz_5, \tag{2.24}$$

where

$$c = (2\pi)^{-\frac{5}{2}} (f_2 - f_1)^{-1/2} (f_3 - f_2)^{-1/2} (f_4 - f_3)^{-1/2} (f_6 - f_5)^{-1/2} (f_7 - f_6)^{-1/2} (f_8 - f_7)^{-1/2},$$

and

$$\psi = -\frac{1}{2}\Bigg(\frac{z_1^2}{f_2-f_1} + \frac{(z_2-(-1)^{j_1}z_1-2\lambda j_1)^2}{f_3-f_2} + \frac{(z_3-z_2)^2}{f_4-f_3} + \frac{(z_4-z_3)^2}{f_6-f_5} + \frac{(z_5-(-1)^{j_2}z_4-2\lambda j_2)^2}{f_7-f_6} + \frac{z_5^2}{f_8-f_7}\Bigg).$$

2.3 Software implementation and numerical analysis

In this section, we introduce the C++ and the R implementation of the proposed FFT-based method for computing $\mathsf{P}(D_n \geq q)$, described in Section 2.2.1 and study its numerical properties. In the sequel, we will refer to it as the Exact-KS-FFT method. The method is implemented in the R package **KSgeneral** which can be downloaded from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=KSgeneral. In order to build the **KSgeneral** package from source, a C++ compiler is required. The latter is contained in the Windows program **Rtools** (R Core Team, 2018), or under MacOS in Xcode, downloadable from the App Store. The package **KSgeneral** uses **Rcpp** (Eddelbuettel and François, 2011) in R, and utilizes the C++ code that efficiently computes $\mathsf{P}(D_n \geq q)$ using the Exact-KS-FFT method (see the replication material to this paper available online). Since the latter requires computation of FFT, the **FFTW3** library developed by Frigo and Johnson (2005) needs to be installed from http://www.fftw.org/index.html. It should be noted that both the **Rtools** and **FFTW3** should be installed in the system `PATH`.

In this section, we also study the asymptotic formulae (2.15) and (2.22) given in Section 2.2.2, which have been implemented in Mathematica 10. For the purpose, in the next Sections 2.3.1 and 2.3.2, we compute the complementary CDF, $\mathsf{P}(D_n \geq q)$, for different values of n and q, and also compute related p values when $F(x)$ is mixed and discrete, respectively. Then, in Section 2.3.3 we consider $\mathsf{P}(D_n < q)$ and $\mathsf{P}(D_n \geq q)$ in the case of continuous $F(x)$. For the examples given in all three Sections 2.3.1, 2.3.2, 2.3.3 (and in the replication material), we give the lines of code that should be executed in C++ or R using **KSgeneral**. Furthermore, in the case when $F(x)$ is mixed (cf., Section 2.3.1), we compare the exact probabilities $\mathsf{P}(D_n \geq q)$, $q \in [0,1]$, obtained using the Exact-KS-FFT approach with those obtained using the asymptotic formula (2.15).

In addition, when $F(x)$ is purely discrete (cf., Section 2.3.2), we also compare with the results of the Brownian bridge simulation-based algorithm of Wood and Altavela (1978). When $F(x)$ is continuous, in Section 2.3.3, Appendix A.2, Appendix A.3, and Appendix A.4, we compare the accuracy and speed of the Exact-KS-FFT method to the results obtained from the R program of Carvalho (2015), and the C program due to Simard and L'Ecuyer (2011). The reported CPU times are obtained running the related C++ code on a machine with an 2.5GHz Intel Core i5 processor with 4GB RAM, running Mac OS X Yosemite.

2.3.1 Complementary CDF of D_n when $F(x)$ is mixed

In order to illustrate the performance of the Exact-KS-FFT method of Section 2.2.1, we consider first the following example from excess-of-loss reinsurance.

Example 2.3.1. Consider an excess-of-loss reinsurance contract with a retention level M and a limiting level L, where $0 < M < L$ are positive constants. Under such a contract, given a loss amount random variable X with a continuous CDF $F_X(\cdot)$ on $[0,+\infty)$, the insurer and the reinsurer pay correspondingly the amounts Z and Y, where

$$Z = \begin{cases} X \text{ if } X \le M, \\ M \text{ if } M < X \le L, \\ M + X - L \text{ if } L < X, \end{cases} \quad \text{and} \quad Y = \begin{cases} 0 \text{ if } X \le M, \\ X - M \text{ if } M < X \le L, \\ L - M \text{ if } L < X. \end{cases}$$

Clearly, both Z and Y are mixed random variables with correspondingly, one and two jumps in their CDFs. For illustrative purposes, assume that the CDF of Y, $F_Y(y)$ is of the form

$$F_Y(y) = \begin{cases} 0 \text{ if } y < 0, \\ 1 - 0.5e^{-y} \text{ if } 0 \le y < \log 2.5, \\ 1 \text{ if } y \ge \log 2.5, \end{cases} \tag{2.25}$$

where $M = \log 2$, $L = \log 5$, $F_X(x) = 1 - e^{-x}$. Assuming D_n in (2.1) is defined with respect to $F_Y(y)$, i.e., $F(x) \equiv F_Y(y)$ in (2.1), we have computed exact probabilities $\mathsf{P}(D_n \ge q)$, for different values of n and q, applying the Exact-KS-FFT method and

also, the asymptotic formula (2.15). In order to apply (2.15), one should note that $F_Y(y)$ has two jumps, (i.e., $J = 2$) at $x_1 = 0, x_2 = \log 2.5$, and $f_0 = f_1 = 0, f_2 = 0.5, f_3 = 0.8, f_4 = f_5 = 1$. Since the jump structure of $F_Y(\cdot)$ in (2.25) is flat-jump, increasing-jump segments, the first set of increasing-jump segments and the last set of flat-jump segments in (2.14) should be omitted. Therefore, one should apply formula (2.15) with $m = 2, \nu_1 = 0, \omega_1 = 1, \nu_2 = 1, \omega_2 = 0$, and $v_0 = 0, v_1 = 0, v_2 = 1, w_0 = 0, w_1 = 1, w_2 = 1$.

The results for $\mathsf{P}(D_n \geq q)$ calculated using the proposed FFT-based method and the asymptotic formula (2.15), for different values of n, q, and respectively $\lambda = qn^{1/2}$, are shown in Tables 2.1 and 2.2. For example, to obtain the probability $\mathsf{P}(D_n \geq q)$ using C++, for $n = 25$, $q = 0.60$ as shown in the column Exact-KS-FFT of Table 2.1, according to step (i) of the Procedure Exact-KS-FFT, we first define the mixed CDF in (2.25) in the file "crossprob.cc" using the following code.

```
vector<double> MixDF (vector<double> obs){
    vector<double> observed = obs;
    set<double> s;
    for (int i = 0; i < obs.size(); ++i){
        s.insert(obs[i]);
    }
    obs.assign(s.begin(), s.end());
    vector<double> DF(obs.size());
    /* The distribution in the reinsurance example in (25) */
        for (int i = 0; i < obs.size(); ++i){
            if (obs[i] < 0.0){
                DF[i] = 0.0;
            }
            else if (obs[i] < log(2.5)){
                DF[i] = 1 - 0.5 * exp(-1.0 * obs[i]);
            }
            else
            {
                DF[i] = 1.0;
            }
        }
    return DF;
}
```

Also, since the mixed CDF in (2.25) has jumps at $y = 0$ and $y = \log 2.5$, we need to specify this by inputting `vector_input3 = {0.0, log(2.5)};` to the `int main()` function in the file "crossprob.cc".

Next, we first run `make` in one of the command line tools (e.g., bash) to build the program for the Exact-KS-FFT method developed in this paper, based on the code provided by Moscovich and Nadler (2017). Then, in the command line tool, we run the following line `./bin/crossprob ecdf 25 Boundary_Crossing_Time.txt`, where 25 is the input for the sample size. We will have the following screen prompts.

```
Please enter the distribution type: 1 for Continuous Distribution,
2 for Discontinuous Distributions:
```

We enter 2 since the CDF in (2.25) is not continuous.

```
2
```

Then, we can choose whether to calculate the KS complementary CDF, $\mathsf{P}(D_n \geq q)$, or the p value, $\mathsf{P}(D_n \geq d_n)$ corresponding to a value d_n computed based on a user provided data sample.

```
Please enter the objective: 1 for KS Complementary Distribution,
2 for P-Values:
```

Since we want to obtain the probability $\mathsf{P}(D_n \geq q)$, for $n = 25$, $q = 0.6$, we will enter 1.

```
1
```

Here, we enter the sample size n and the quantile q.

```
Please enter the sample size and quantile:
```

```
25
0.6
```

```
Probability: 0.0000000019082332
Time taken: 0.0000720000000000
```

Now, steps (ii), (iii), (iv) and (v) of the Procedure Exact-KS-FFT are performed. The result for $\mathsf{P}(D_n \geq q)$, for $n = 25$, $q = 0.60$, is 1.90823×10^{-9} as shown in the column Exact-KS-FFT of Table 2.1. The corresponding computation time is also printed.

Remark 2.3.2. Note that the distribution of the KS test statistic D_n depends on the hypothesized distribution $F(x)$ when $F(x)$ is not continuous. Hence, to obtain $\mathsf{P}(D_n \geq q)$ for different mixed $F(x)$, the users should: 1) define the mixed CDF in the file "crossprob.cc" each time, and 2) in the file "crossprob.cc", define the vector containing points where $F(x)$ has jumps, `vector_input3`.

In order to compute $\mathsf{P}(D_n \geq q)$, when $F(x)$ is mixed using the R package **KS-general**, one needs to input `mixed_ks_c_cdf(q, n, jump_points, Mixed_dist, ..., tol = 1e - 10)`, where `jump_points` is a numeric vector of the x coordinates of the jumps of $F(x)$, `Mixed_dist` specifies the mixed CDF $F(x)$, possibly followed by a list of parameters `...` specifying $F(x)$, and `tol` is the value of ε that is used to compute the values A_i and B_i, $i = 1,...,n$, as detailed in equations (2.4) in Step 1 of Section 2.2.1. By default, `tol = 1e - 10`. Note that a value `NA` or `0` will lead to an error. For instance, if one wants to use the R package **KSgeneral** to compute $\mathsf{P}(D_n \geq q)$, when $F(x)$ is the mixed CDF specified in Example 2.3.1 by equation (2.25), with $n = 25$, $q = 0.1$, one needs to run the following code in order to obtain the corresponding result, as shown in Table 2.2 for $n = 25, q = 0.1$.

```
R> Mixed_cdf_example <- function(x)
{
  result <- 0
  if (x < 0){
    result <- 0
  }
  else if (x == 0){
    result <- 0.5
  }
  else if (x < log(2.5)){
    result <- 1 - 0.5 * exp(-x)
  }
  else{
    result <- 1
  }
  return (result)
}
R> mixed_ks_c_cdf(0.1, 25, c(0, log(2.5)), Mixed_cdf_example)
```

```
[1] 0.76768489
```

From Tables 2.1 and 2.2, one can first see that the Exact-KS-FFT method effectively computes $\mathsf{P}(D_n \geq q)$ for small, medium and large sample sizes n and various levels q, and gives exact probabilities in the range of 10^{-10} to 1. It should be noted though that the method could become numerically unstable (producing negative values) when calculating probabilities of 10^{-11} or smaller. Similar issue has been observed by Simard and L'Ecuyer (2011) in the case of continuous $F(x)$. The column Rel.err. (%) quantifies

the relative error of the asymptotic value (for fixed λ) compared to the exact values in the column Exact-KS-FFT (for various combinations of n and q resulting in the same λ). Furthermore, we see that when the sample size n is large, results using formula (2.15) approximate closely the exact $\mathsf{P}(D_n \geq q)$, except when $\mathsf{P}(D_n \geq q)$ is nearly zero (when $\lambda = 2,3$ in Table 2.1). Also, asymptotic formula (2.15) gives better approximations to the exact values of $\mathsf{P}(D_n \geq q)$ as q decreases, or equivalently, as $\mathsf{P}(D_n \geq q)$ increases. Moreover, as λ decreases, values obtained from asymptotic formula (2.15) become better approximations to the exact $\mathsf{P}(D_n \geq q)$. Let us recall however that formula (2.15) (respectively (2.20) and (2.22)) is only practically implementable for small/moderate number of jumps, J, in the null distribution (which is the case with (2.25) illustrated in Tables 2.1 and 2.2), as otherwise the multidimensional integration becomes infeasible.

In addition, as mentioned in Section 2.1, a null hypothesis that a sample comes from a discontinuous distribution will be accepted more often if one uses the continuous KS test, as opposed to using the discontinuous KS test. To illustrate this, assume that a random sample of size $n = 25$ follows $F(x) \equiv F_Y(y)$ in (2.25) under H_0, and that the KS test statistic for the sample is $d_n = 0.25$. Then, the exact p value of the test is $\mathsf{P}(D_n \geq 0.25|H_0) = 0.04496610$ and, with a significance level of 5%, one should reject H_0. On the other hand, a p value calculated using the complementary CDF of the distribution-free continuous KS test statistic D_n (i.e., when $F(x)$ in (2.1) is continuous) is $0.07360597 > 0.05$. Therefore, based on the latter p value, one will not reject H_0. Similar situations are illustrated in Table 2.3 for larger sample sizes and different values of the test statistic D_n, where one can see that the differences between the values in the last two columns are higher than 58% (our experience shows that these are typically in the range 50% - 65%) and do not decrease with n. To the best of our knowledge, the KS test in softwares such as R, SPSS, Stata, MATLAB, Mathematica is based on the distribution-free continuous KS test statistic and the discontinuous (mixed and purely discrete) version is not implemented due to the lack of efficient and robust method such as the Exact-KS-FFT method we propose here.

Table 2.1 Exact and asymptotic values of $\mathsf{P}(D_n \geq q)$ obtained via the Exact-KS-FFT method and the asymptotic formula (2.15),when $\lambda = qn^{1/2} = 3,\ 2,\ 1$, respectively, when the underlying CDF $F(x)$ follows $F_Y(y)$ in (2.25). Numbers in () are run times in seconds.

λ	n	q	Exact-KS-FFT		Asympt. (15)	Rel.err. (%)
3	25	0.60	1.90823×10^{-9}	(0.000)	1.72031×10^{-8}	801.52
	100	0.30	9.49583×10^{-9}	(0.000)	(5155.54)	81.17
	400	0.15	1.41586×10^{-8}	(0.015)		21.50
	2500	0.06	1.62830×10^{-8}	(0.202)		5.65
	10000	0.03	1.67952×10^{-8}	(2.932)		2.43
	40000	0.015	1.69539×10^{-8}	(59.86)		1.49
	90000	0.01	1.70076×10^{-8}	(351.9)		1.16
	250000	0.006	1.74648×10^{-8}	(3524)		1.43
2	25	0.4	2.13209×10^{-4}	(0.000)	3.98459×10^{-4}	86.89
	100	0.2	3.27304×10^{-4}	(0.000)	(1.17)	21.74
	400	0.1	3.66979×10^{-4}	(0.015)		8.58
	2500	0.04	3.86968×10^{-4}	(0.195)		2.97
	10000	0.02	3.92912×10^{-4}	(2.707)		1.41
	40000	0.01	3.95740×10^{-4}	(57.14)		0.69
	90000	1/150	3.96661×10^{-4}	(341.3)		0.45
	250000	0.004	3.97390×10^{-4}	(3465)		0.27
1	25	0.2	0.151510006	(0.000)	0.174525238	15.19
	100	0.1	0.164499986	(0.000)	(0.73)	6.09
	400	0.05	0.169049900	(0.015)		3.24
	2500	0.02	0.172221536	(0.171)		1.34
	10000	0.01	0.173354312	(2.511)		0.68
	40000	0.005	0.173934996	(54.94)		0.34
	90000	1/300	0.174130680	(330.3)		0.23
	250000	0.002	0.174287993	(3423)		0.14

Table 2.2 Exact and asymptotic values of $\mathsf{P}(D_n \geq q)$ obtained via the Exact-KS-FFT method and the asymptotic formula (2.15),when $\lambda = qn^{1/2} = 0.5$, 0.2 and 0.15, respectively, when the underlying CDF $F(x)$ follows $F_Y(y)$ in (2.25). Numbers in () are run times in seconds.

λ	n	q	Exact-KS-FFT		Asympt. (15)	Rel.err. (%)
0.5	25	0.1	0.767684886	(0.000)	0.801033877	4.35
	100	0.05	0.782681427	(0.000)	(5.63)	2.35
	400	0.025	0.790339869	(0.015)		1.35
	2500	0.01	0.796406211	(0.156)		0.58
	10000	0.005	0.798664879	(2.441)		0.30
	40000	0.0025	0.799837547	(54.27)		0.15
	90000	1/600	0.800234794	(326.5)		0.12
	250000	0.001	0.800554870	(3410)		0.06
0.2	25	0.04	0.999798067	(0.000)	0.999961812	0.016
	100	0.02	0.999888190	(0.000)	(5.03)	0.007
	400	0.01	0.999925985	(0.015)		0.004
	2500	0.004	0.999948507	(0.156)		0.001
	10000	0.002	0.999955380	(2.364)		0.001
	40000	0.001	0.999958655	(53.62)		0.000
	90000	1/1500	0.999959721	(324.4)		0.000
	250000	0.0004	0.999960564	(3383)		0.000
0.15	25	0.03	0.999998692	(0.000)	0.999999978	0.000
	100	0.015	0.999999682	(0.000)	(0.51)	0.000
	400	0.0075	0.999999905	(0.015)		0.000
	2500	0.003	0.999999956	(0.156)		0.000
	10000	0.0015	0.999999969	(2.355)		0.000
	40000	0.00075	0.999999974	(53.45)		0.000
	90000	0.0005	0.999999975	(324.7)		0.000
	250000	0.0003	0.999999977	(3372)		0.000

Table 2.3 Discontinuous and continuous KS p values under null hypothesis $H_0 : F(x) \equiv F_Y(y)$, obtained via the Exact-KS-FFT method.

n	$D_n = d_n$	Discontinuous KS p values	Continuous KS p values
25	0.25	0.04496610	0.07360597
100	0.13	0.03913182	0.06209234
400	0.065	0.04090172	0.06511744
2500	0.026	0.04200207	0.06690821
10000	0.013	0.04237475	0.06750119
40000	0.0065	0.04256212	0.06779695

Example 2.3.3. Another possible application of KS tests on mixed distributions appears in testing the goodness-of-fit in zero-inflated or/and one-inflated models. Many real data contain zeros and ones, i.e., have masses at zero and one, and therefore zero- and one-inflated distributions need to be applied. For example, Ospina and Ferrari (2010) have used the zero-and-one-inflated beta distribution to model the proportion of inhabitants living within a 200 kilometer wide costal strip in 232 countries in the year 2000, denoted as Y. The data for years 1990, 2000 and 2010 are supplied by the Columbia University Centre for International Earth Science Information Network, see CIESIN (2012), and are available at http://sedac.ciesin.columbia.edu/data/set/nagdc-population-landscape-climate-estimates-v3. The zero-and-one-inflated beta distribution considered by Ospina and Ferrari (2010) is of the following form

$$G_Y(y;\alpha,\gamma,\mu,\phi) = \alpha\text{Bernoulli}(y;\gamma) + (1-\alpha)F(y;\mu,\phi), \quad 0 \le y \le 1,$$

where $\text{Bernoulli}(\cdot;\gamma)$ denotes the CDF of a Bernoulli random variable with parameter γ, $0 < \gamma < 1$, and $F(\cdot;\mu,\phi)$ denotes the CDF of a beta random variable with parameters μ, $0 < \mu < 1$, and $\phi > 0$. Hence, the zero-and-one-inflated distribution can be seen as a mixture of a (discrete) Bernoulli distribution and a (continuous) beta distribution, with weights α and $(1-\alpha)$, respectively, $0 < \alpha < 1$.

According to Ospina and Ferrari (2010), the random variable Y has the following distribution

$$G_Y(y) = \begin{cases} 0 \text{ if } y < 0, \\ 0.1141 + 0.4795F_Y(y;\mu,\phi) \text{ if } 0 \le y < 1, \\ 1 \text{ if } y \ge 1, \end{cases}$$

where $F_Y(y;\mu,\phi)$ has a density function

$$f_Y(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

with μ and ϕ estimated as $\mu = 0.6189, \phi = 0.6615$ based on the population data in 2000, and $\Gamma(\cdot)$ is the gamma function. Then, we can examine the goodness-of-fit of the distribution to the population data in 2010, denoted by $\tilde{Y}$, hypothesizing that $\tilde{Y}$ has the same distribution as $G_Y(y)$. Using (2.1) with $F(x) \equiv G_Y(y)$ and $F_n(x) \equiv G_n(y)$, where $G_n(y)$ is the EDF of $\tilde{Y}$ computed from the population data in 2010, we obtain the KS test statistic $d_n = 0.09047$. Using the Exact-KS-FFT method, we compute a p value of $0.03403 < 0.05$. Alternatively, applying the asymptotic formula (2.15), we obtain a p value of 0.03641, which is reasonably accurate, given the sample size of 232. Therefore, the KS test indicates that the zero-and-one-inflated beta distribution estimated using population data in 2000 does not fit the population data in 2010 at a significance level of 5%, providing evidence for a change in the proportion of inhabitants in the decade.

In order to perform the one-sample two-sided KS test, when $F(x)$ is `mixed_ks_test(x, jump_points, Mixed_dist, ..., tol = 1e - 10)`, where `x` is a numeric vector of data sample values, and where other arguments are defined similarly as in the function `mixed_ks_c_cdf()`. For instance, if one wants to use the R package **KSgeneral** to calculate the p value for the KS test, when $F(x)$ follows a zero-and-one-inflated beta distribution as in Example 2.3.3, with a sample of size $n = 232$, one should run the following R code.

```
R> data("Population_Data")
R> mu <- 0.6189
R> phi <- 0.6615
R> a <- mu * phi
R> b <- (1 - mu) * phi
R> Mixed_cdf_example <- function(x)
{
  result <- 0
  if (x < 0){
    result <- 0
  }
  else if (x == 0){
    result <- 0.1141
  }
  else if (x < 1){
    result <- 0.1141 + 0.4795 * pbeta(x, a, b)
  }
```

```
  else{
    result <- 1
  }
  return (result)
}
R> ksgeneral::mixed_ks_test(Population_Data, c(0, 1), Mixed_cdf_example)
```

```
       One-sample Kolmogorov-Smirnov test

data: Population
D = 0.0904737, p-value = 0.034025
alternative hypothesis: two-sided
```

In the next section, assuming $F(x)$ is purely discrete, we apply the FFT-based methodology and the asymptotic formula (2.22) (cf., Corollary 2.2.6) to compute correspondingly, exact and approximate values of $\mathsf{P}(D_n \geq q)$.

2.3.2 Complementary CDF of D_n when $F(x)$ is purely discrete

There is an abundance of real-life applications in which purely discrete distributions are used to model count data, such as number of claims to an insurance company, number of jumps in stock returns, number of trades on the stock exchange, number of manufacturing defects, number of diseased species and plants in biology and agricultural research, and many other count data applications. In all such cases, examining the goodness-of-fit of the model requires computing p values or $\mathsf{P}(D_n \geq q)$ for various of n and q. As an illustration, using the proposed FFT-based method, we will compute exact probabilities $\mathsf{P}(D_n \geq q)$ when the underlying $F(x)$ follows Binomial(r, π) distribution (see Example 2.3.4) and when it follows a discrete uniform distribution (see Example 2.3.5). In Example 2.3.4, we compare these exact probabilities with approximate ones obtained using the asymptotic distribution of D_n, given by (2.22), and using the asymptotic MC simulation-based method of Wood and Altavela (1978). In Example 2.3.5, we compare the exact results with those obtained using the R function `ks.test` of Arnold and Emerson (2011). The latter is a revised version of the same function from the recommended package **stats**.

Wood and Altavela (1978)'s approach utilizes the connection between the asymptotic distribution of D_n and a multi-variate Brownian bridge (cf., Remark 2.2.7), and they directly simulate the latter, thus avoiding the necessity to derive and evaluate an explicit expression such as (2.22). Following the Wood and Altavela (1978)'s method, one should simulate from the $(J-1)$-variate normal random vector $(Z_1, Z_2, ..., Z_{J-1})$, where

$$\mathsf{E}(Z_i) = 0, \quad \mathsf{E}(Z_i, Z_k) = \min(f_{2i}, f_{2k}) - f_{2i} f_{2k}, \quad i,k = 1, ..., J-1, \tag{2.26}$$

and estimate the probability in $\Phi(\lambda)$ in (2.11) as

$$\frac{\sum_{i=1}^{N} \mathbb{1}_{\{(Z_1, Z_2, ..., Z_{J-1}) \in [-\lambda, \lambda]^{J-1}\}}}{N},$$

where N is the number of simulations, $\mathbb{1}_{\{\cdot\}}$ is an indicator function, and $[-\lambda, \lambda]^{J-1}$ is the $(J-1)$ dimensional hypercube. The authors further suggest a continuity correction for λ in (2.11), as $\lambda = qn^{1/2} - 0.5n^{-1/2}$. In the remainder of this section, we will refer to this method as W&A(a) method and to its version without the continuity correction, as W&A(b) method.

Example 2.3.4. Assume that $F(x)$ in (2.1) is Binomial(r, π) with $r = 3, 7, 15$ (i.e., with $J = r+1$ number of jumps), and $\pi = 0.5$. In Tables 2.4, 2.5, and 2.6, for different values of n, q, and respectively $\lambda = qn^{1/2}$, we give the exact $\mathsf{P}(D_n \geq q)$ obtained with the Exact-KS-FFT method, and compare with the asymptotic probabilities obtained using (2.22) (combined with (2.13)), and using the Wood and Altavela (1978) simulation-based approach. We have coded both the W&A(a) and W&A(b) versions in R as part of the **KSgeneral** R package and have simulated 1000000 realizations of the random vector $(Z_1, Z_2, ..., Z_{J-1})$. As before, the numbers in parentheses show the computation (run) times, in seconds. Let us note that the multidimensional numerical integration in (2.22) becomes unstable as the number of jumps, $J = r+1$, in $F(x)$ increases, and so we only use W&A(a) and W&A(b) to obtain approximate asymptotic probabilities in the case of $r = 15$ and $\pi = 0.5$ (see Table 2.6).

In order to compute $\mathsf{P}(D_n \geq q)$, when $F(x)$ is purely discrete using the R package **KSgeneral**, one needs to input `disc_ks_c_cdf(q, n, y, ..., exact = NULL, tol = 1e - 08, sim.size = 1e + 06, num.sim = 10)`, where `y` specifies the purely discrete CDF $F(x)$, possibly followed by a list of parameters `...` specifying $F(x)$, the input parameter `exact` is a logical variable specifying whether one wants to compute exact values for $\mathsf{P}(D_n \geq q)$ using the FFT-based method, `exact = TRUE` or wants to compute the approximate values for $\mathsf{P}(D_n \geq q)$ using the simulation-based algorithm of Wood and Altavela (1978), in which case `exact = FALSE`. When `exact = NULL` and `n <= 100000`, the exact $\mathsf{P}(D_n \geq q)$ will be computed using the FFT-based method. The input parameter `tol` is the value of ε that is used to compute the values A_i and B_i, $i = 1,...,n$, as detailed in equations (2.4) in Step 1 of Section 2.2.1. By default, `tol = 1e - 08`. Note that a value of `NA` or `0` will lead to an error. The input parameter `sim.size` is the required number of simulated trajectories in order to produce one MC estimate (one MC run) of the asymptotic p value using the algorithm of Wood and Altavela (1978). By default, `sim.size = 1e + 06`. The input parameter `num.sim` is the number of MC runs, each producing one estimate (based on `sim.size` number of trajectories), which are then averaged in order to produce the final estimate for the asymptotic p value. This is done in order to reduce the variance of the final estimate. By default, `num.sim = 10`. For instance, if one wants to use the R package **KSgeneral** to compute the exact value for $\mathsf{P}(D_n \geq q)$, when $F(x)$ follows a Binomial$(3, 0.5)$ distribution as in Example 2.3.4, with $n = 400$, $q = 0.05$, one should run the following R code and obtain the corresponding result as shown in the column Exact-KS-FFT of Table 2.4.

```
R> binom_3 <- stepfun(c(0 : 3), c(0, pbinom(0 : 3, 3, 0.5)))
R> disc_ks_c_cdf(0.05, 400, binom_3)
```

```
[1] 0.05611849
```

On the other hand, if one wants to use the simulation-based method of Wood and Altavela (1978) in order to approximate the asymptotic value for $\mathsf{P}(D_n \geq q)$, when $F(x)$ follows a Binomial$(3, 0.5)$ distribution, with $n = 400$, $q = 0.05$, one should use the

W&A(a) method, by running the following R code and obtain the corresponding result as shown in the column W&A(a) of Table 2.4.

```
R> binom_3 <- stepfun(c(0 : 3), c(0, pbinom(0 : 3, 3, 0.5)))
R> disc_ks_c_cdf(0.05, 400, binom_3, exact = FALSE, tol = 1e-08,
+ sim.size = 1e+06, num.sim = 10)
```

```
[1] 0.0561864
```

Looking at Tables 2.4, 2.5, and 2.6, one can see that the Exact-KS-FFT method effectively computes $\mathsf{P}(D_n \geq q)$ for small, medium and large sample sizes n and various levels q, and gives exact probabilities in the range 10^{-12} to 1. We also see that when the sample size n is large, results using formula (2.22) approximate closely the exact $\mathsf{P}(D_n \geq q)$, except when $\mathsf{P}(D_n \geq q)$ is nearly zero (when $\lambda = 2,3$ in Tables 2.4 and 2.5). Similarly to the mixed $F(x)$ case, asymptotic formula (2.22) gives better approximations to the exact values of $\mathsf{P}(D_n \geq q)$ as q decreases, or equivalently, as $\mathsf{P}(D_n \geq q)$ increases. Moreover, as λ decreases, values obtained from asymptotic formula (2.22) become better approximations to the exact $\mathsf{P}(D_n \geq q)$. One can further observe that asymptotic formula (2.22) and W&A(b) method provide similar results. In particular, as the number of jumps in $F(x)$ increases, results obtained from these two methods almost coincide. In addition, when the number of jumps in $F(x)$ is small (in our case $J = 4$ or 8), we see that values obtained from W&A(a) method provide more accurate approximations to the exact probabilities. On the other hand, when the number of jumps in $F(x)$ is large (in our case $J = 16$), values obtained from W&A(b) method give closer approximations. In comparison with the Exact-KS-FFT method, W&A(a) and W&A(b) deviate stronger from the exact probabilities for moderate values of λ, e.g., $\lambda = 0.5$, 1, and this is more pronounced for small sample sizes, see $n \leq 400$.

With regards to computation time, looking at Tables 2.4, 2.5, and 2.6, for fixed sample size n and number of jumps J, as λ decreases, the computation time for the Exact-KS-FFT method, W&A(a) method and W&A(b) method decreases. Furthermore, when the sample size n and q are fixed, as the number of jumps in $F(x)$, J, increases, the computation time for the Exact-KS-FFT method decreases, whereas the computation

time for W&A(a) and W&A(b) methods increases. And, as expected, when the sample size n increases, the Exact-KS-FFT method becomes more time-consuming.

Table 2.4 Exact and asymptotic values of $\mathsf{P}(D_n \geq q)$ obtained via the Exact-KS-FFT method, the asymptotic formula (2.22) and W&A(a), W&A(b) methods for $\lambda = qn^{1/2} = 3, 2, 1, 0.5, 0.2$ and 0.1, respectively, when the underlying CDF $F(x)$ follows Binomial$(3, 0.5)$ distribution. Numbers in () are run times in seconds.

			$\lambda = 3$		
n	q	Exact-KS-FFT	Time	Asympt.(2.22)	Rel.err. (%)
25	0.60	1.15052×10^{-12}	(0.000)	1.97318×10^{-9}	
400	0.15	2.04622×10^{-9}	(0.015)	(10.55)	3.570
10000	0.03	2.07657×10^{-9}	(3.291)		4.979
90000	0.01	1.89810×10^{-9}	(427.5)		-3.955
n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.60	1.15052×10^{-12}	0	(6.06)	
400	0.15	2.04622×10^{-9}	0	(6.14)	
10000	0.03	2.07657×10^{-9}	0	(6.14)	
90000	0.01	1.89810×10^{-9}	0	(6.13)	
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.60	1.15052×10^{-12}	0	(6.29)	
400	0.15	2.04622×10^{-9}	0	(6.29)	
10000	0.03	2.07657×10^{-9}	0	(6.29)	
90000	0.01	1.89810×10^{-9}	0	(6.29)	
			$\lambda = 2$		
n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.40	1.99454×10^{-5}	(0.000)	6.33453×10^{-5}	
400	0.10	7.43068×10^{-5}	(0.015)	(10.82)	14.752

Continued on next page

Table 2.4 – continued from previous page

10000	0.02	6.59391×10^{-5}	(3.010)		3.934
90000	1/150	6.42285×10^{-5}	(414.1)		1.375
n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.40	1.99454×10^{-5}	1.41933×10^{-4}	(6.06)	
400	0.10	7.43068×10^{-5}	7.50667×10^{-5}	(6.13)	-1.023
10000	0.02	6.59391×10^{-5}	6.33667×10^{-5}	(6.14)	3.901
90000	1/150	6.42285×10^{-5}	6.18333×10^{-5}	(6.16)	3.729
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.40	1.99454×10^{-5}	6.11333×10^{-5}	(6.16)	
400	0.10	7.43068×10^{-5}	6.11333×10^{-5}	(6.16)	17.728
10000	0.02	6.59391×10^{-5}	6.11333×10^{-5}	(6.16)	7.288
90000	1/150	6.42285×10^{-5}	6.11333×10^{-5}	(6.16)	4.819
			$\lambda = 1$		
n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.20	0.046850021	(0.000)	0.049438582	
400	0.05	0.056118495	(0.015)	(5.30)	11.903
10000	0.01	0.050721030	(2.776)		2.528
90000	1/300	0.049863086	(400.2)		0.851
n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.20	0.046850021	0.08178503	(5.92)	
400	0.05	0.056118495	0.05618643	(6.00)	-0.121
10000	0.01	0.050721030	0.05073470	(6.01)	-0.027
90000	1/300	0.049863086	0.04986763	(6.02)	-0.009
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)

Continued on next page

Table 2.4 – continued from previous page

25	0.20	0.046850021	0.04944507	(6.04)	
400	0.05	0.056118495	0.04944507	(6.04)	11.892
10000	0.01	0.050721030	0.04944507	(6.04)	2.516
90000	1/300	0.049863086	0.04944507	(6.04)	0.838

$\lambda = 0.5$

n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.10	0.532599669	(0.000)	0.46014460	
400	0.025	0.500282800	(0.015)	(10.65)	8.023
10000	0.005	0.468139770	(2.574)		1.708
90000	1/600	0.462807932	(392.2)		0.575

n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.10	0.532599669	0.62989733	(5.15)	
400	0.025	0.500282800	0.50094213	(5.38)	-0.132
10000	0.005	0.468139770	0.46828113	(5.40)	-0.030
90000	1/600	0.462807932	0.46293110	(5.42)	-0.027

n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.10	0.532599669	0.46026040	(5.44)	
400	0.025	0.500282800	0.46026040	(5.44)	8.000
10000	0.005	0.468139770	0.46026040	(5.44)	1.683
90000	1/600	0.462807932	0.46026040	(5.44)	0.550

$\lambda = 0.2$

n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.04	0.935407699	(0.000)	0.92701801	
400	0.01	0.949180930	(0.015)	(10.84)	2.335

Continued on next page

Table 2.4 – continued from previous page

10000	0.002	0.931797646	(2.527)		0.513
900001	1/1500	0.928630334	(389.4)		0.174
n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.04	0.935407699	0.98963853	(4.13)	
400	0.01	0.949180930	0.94915067	(4.33)	0.003
10000	0.002	0.931797646	0.93180017	(4.35)	0.000
900001	1/1500	0.928630334	0.92863450	(4.39)	0.000
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.04	0.935407699	0.92702207	(4.40)	
400	0.01	0.949180930	0.92702207	(4.40)	2.335
10000	0.002	0.931797646	0.92702207	(4.40)	0.513
900001	1/1500	0.928630334	0.92702207	(4.40)	0.173
			$\lambda = 0.1$		
n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.020	0.999999999	(0.000)	0.98963108	
400	0.005	0.995546700	(0.015)	(10.84)	0.594
10000	0.001	0.991072365	(2.480)		0.145
900001	1/3000	0.990126719	(388.8)		0.050
n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.020	0.999999999	1	(3.88)	
400	0.005	0.995546700	0.99553633	(4.02)	0.001
10000	0.001	0.991072365	0.99106897	(4.10)	0.000
900001	1/3000	0.990126719	0.99013033	(4.11)	0.000
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)

Continued on next page

Table 2.4 – continued from previous page

25	0.020	0.999999999	0.98963853	(4.13)	
400	0.005	0.995546700	0.98963853	(4.13)	0.593
10000	0.001	0.991072365	0.98963853	(4.13)	0.145
90000	1/3000	0.990126719	0.98963853	(4.13)	0.049

Table 2.5 Exact and asymptotic values of $\mathsf{P}(D_n \geq q)$ obtained via the Exact-KS-FFT method, the asymptotic formula (2.22) and W&A(a), W&A(b) methods for $\lambda = qn^{1/2} = 3, 2, 1, 0.5, 0.2$ and 0.1, respectively, when the underlying CDF $F(x)$ follows Binomial$(7, 0.5)$ distribution. Numbers in () are run times in seconds.

$\lambda = 3$					
n	q	Exact-KS-FFT	Time	Asympt.(2.22)	Rel.err. (%)
25	0.60	2.74894×10^{-10}	(0.000)	6.90809×10^{-4}	
400	0.15	2.06159×10^{-9}	(0.015)	(32.43)	
10000	0.03	2.08064×10^{-9}	(2.074)		
90000	0.01	1.91281×10^{-9}	(259.6)		
n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.60	2.74894×10^{-10}	0	(11.43)	
400	0.15	2.06159×10^{-9}	0	(10.92)	
10000	0.03	2.08064×10^{-9}	0	(10.99)	
90000	0.01	1.91281×10^{-9}	0	(10.94)	
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.60	2.74894×10^{-10}	0	(10.93)	
400	0.15	2.06159×10^{-9}	0	(10.93)	
10000	0.03	2.08064×10^{-9}	0	(10.93)	

Continued on next page

Table 2.5 – continued from previous page

90000	0.01	1.91281×10^{-9}	0	(10.93)	
			$\lambda = 2$		
n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.40	4.20725×10^{-5}	(0.000)	-6.69185×10^{-5}	
400	0.10	7.91684×10^{-5}	(0.015)	(34.01)	
10000	0.02	6.93244×10^{-5}	(1.840)		
90000	1/150	6.75595×10^{-5}	(244.9)		
n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.40	4.20725×10^{-5}	1.54867×10^{-4}	(10.95)	
400	0.10	7.91684×10^{-5}	8.06000×10^{-5}	(10.93)	-1.808
10000	0.02	6.93244×10^{-5}	6.81333×10^{-5}	(10.94)	1.718
90000	1/150	6.75595×10^{-5}	6.63333×10^{-5}	(10.98)	1.815
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.40	4.20725×10^{-5}	6.51×10^{-5}	(10.98)	
400	0.10	7.91684×10^{-5}	6.51×10^{-5}	(10.98)	17.770
10000	0.02	6.93244×10^{-5}	6.51×10^{-5}	(10.98)	6.094
90000	1/150	6.75595×10^{-5}	6.51×10^{-5}	(10.98)	3.640
			$\lambda = 1$		
n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.20	0.068266018	(0.000)	0.070168353	
400	0.05	0.074899103	(0.015)	(35.39)	6.316
10000	0.01	0.070933439	(1.606)		1.079
90000	1/300	0.070290581	(233.2)		0.174

Continued on next page

Table 2.5 – continued from previous page

n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.20	0.068266018	0.11542800	(10.58)	
400	0.05	0.074899103	0.07965367	(10.65)	-6.348
10000	0.01	0.070933439	0.07187410	(10.71)	-1.326
90000	1/300	0.070290581	0.07064190	(10.74)	-0.500
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.20	0.068266018	0.070035433	(10.75)	
400	0.05	0.074899103	0.070035433	(10.75)	6.494
10000	0.01	0.070933439	0.070035433	(10.75)	1.266
90000	1/300	0.070290581	0.070035433	(10.75)	0.363
			$\lambda = 0.5$		
n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.10	0.619487745	(0.000)	0.56366243	
400	0.025	0.583754412	(0.015)	(45.96)	3.442
10000	0.005	0.567662656	(1.481)		0.705
90000	1/600	0.564996352	(221.4)		0.236
n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.10	0.619487745	0.73754833	(8.31)	
400	0.025	0.583754412	0.60684037	(8.76)	-3.955
10000	0.005	0.567662656	0.57234977	(8.93)	-0.826
90000	1/600	0.564996352	0.56664517	(8.95)	-0.292
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.10	0.619487745	0.56379917	(8.96)	
400	0.025	0.583754412	0.56379917	(8.96)	3.418

Continued on next page

Table 2.5 – continued from previous page

10000	0.005	0.567662656	0.56379917	(8.96)	0.681
90000	1/600	0.564996352	0.56379917	(8.96)	0.212

$\lambda = 0.2$

n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.04	0.976334785	(0.000)	0.97713513	
400	0.01	0.983298737	(0.015)	(89.49)	0.627
10000	0.002	0.978475846	(1.404)		0.137
900001	/1500	0.977587940	(216.5)		0.046

n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.04	0.976334785	0.99938833	(6.01)	
400	0.01	0.983298737	0.98756200	(6.61)	-0.434
10000	0.002	0.978475846	0.97956150	(6.73)	-0.111
900001	/1500	0.977587940	0.97796337	(6.78)	-0.038

n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.04	0.976334785	0.97713387	(6.80)	
400	0.01	0.983298737	0.97713387	(6.80)	0.627
10000	0.002	0.978475846	0.97713387	(6.80)	0.137
900001	/1500	0.977587940	0.97713387	(6.80)	0.046

$\lambda = 0.1$

n	q	Exact-KS-FFT	Time	Asympt. (2.22)	Rel.err. (%)
25	0.020	0.999999999	(0.000)	0.99938850	
400	0.005	0.999745396	(0.015)	(11.75)	0.036
10000	0.001	0.999472182	(1.388)		0.008
900001	/3000	0.999417006	(214.6)		0.003

Continued on next page

Table 2.5 – continued from previous page

n	q	Exact-KS-FFT	W&A(a)	Time	Rel.err. (%)
25	0.020	0.999999999	1	(4.97)	
400	0.005	0.999745396	0.99989767	(5.72)	-0.015
10000	0.001	0.999472182	0.99955080	(5.90)	-0.008
900001	/3000	0.999417006	0.99944667	(5.97)	-0.003
n	q	Exact-KS-FFT	W&A(b)	Time	Rel.err. (%)
25	0.020	0.999999999	0.99938833	(6.00)	
400	0.005	0.999745396	0.99938833	(6.00)	0.036
10000	0.001	0.999472182	0.99938833	(6.00)	0.008
900001	/3000	0.999417006	0.99938833	(6.00)	0.003

Table 2.6 Exact and asymptotic values of $\mathsf{P}(D_n \geq q)$ obtained via the Exact-KS-FFT method and W&A(a), W&A(b) methods for $\lambda = qn^{1/2} = 3, 2, 1, 0.5, 0.2$ and 0.1, respectively, when the underlying CDF $F(x)$ follows Binomial$(15, 0.5)$ distribution. Numbers in () are run times in seconds.

$\lambda = 3$				
n	q	Exact-KS-FFT	W&A(a)	Rel.err. (%)
25	0.60	4.08521×10^{-10} (0.000)	0 (21.75)	
400	0.15	2.32760×10^{-9} (0.015)	0 (21.78)	
10000	0.03	2.21527×10^{-9} (1.622)	0 (21.79)	
90000	0.01	2.07134×10^{-9} (186.3)	0 (21.86)	
n	q	Exact-KS-FFT	W&A(b)	Rel.err. (%)
25	0.60	4.08521×10^{-10} (0.000)	0 (21.87)	
400	0.15	2.32760×10^{-9} (0.015)	0 (21.87)	

Continued on next page

Table 2.6 – continued from previous page

10000	0.03	2.21527×10^{-9} (1.622)	0 (21.87)	
90000	0.01	2.07134×10^{-9} (186.3)	0 (21.87)	
		$\lambda = 2$		
n	q	Exact-KS-FFT	W&A(a)	Rel.err. (%)
25	0.40	6.62012×10^{-5} (0.000)	2.07367×10^{-4} (21.60)	
400	0.10	9.95661×10^{-5} (0.015)	1.06900×10^{-4} (21.63)	-7.366
10000	0.02	9.05026×10^{-5} (1.387)	8.92667×10^{-5} (21.70)	1.366
90000	1/150	8.88601×10^{-5} (173.5)	8.66333×10^{-5} (21.76)	2.506
n	q	Exact-KS-FFT	W&A(b)	Rel.err. (%)
25	0.40	6.62012×10^{-5} (0.000)	8.55×10^{-5} (21.80)	
400	0.10	9.95661×10^{-5} (0.015)	8.55×10^{-5} (21.80)	14.13
10000	0.02	9.05026×10^{-5} (1.387)	8.55×10^{-5} (21.80)	5.53
90000	1/150	8.88601×10^{-5} (173.5)	8.55×10^{-5} (21.80)	3.78
		$\lambda = 1$		
n	q	Exact-KS-FFT	W&A(a)	Rel.err. (%)
25	0.20	0.089163050 (0.000)	0.14505810 (20.65)	
400	0.05	0.093364526 (0.015)	0.10142243 (21.02)	-8.631
10000	0.01	0.090270911 (1.138)	0.09184050 (21.13)	-1.739
90000	1/300	0.089721687 (161.2)	0.09031193 (21.10)	-0.658
n	q	Exact-KS-FFT	W&A(b)	Rel.err. (%)
25	0.20	0.089163050 (0.000)	0.08956207 (21.11)	
400	0.05	0.093364526 (0.015)	0.08956207 (21.11)	4.073
10000	0.01	0.090270911 (1.138)	0.08956207 (21.11)	0.785
90000	1/300	0.089721687 (161.2)	0.08956207 (21.11)	0.178

Continued on next page

Table 2.6 – continued from previous page

$\lambda = 0.5$				
n	q	Exact-KS-FFT	W&A(a)	Rel.err. (%)
25	0.10	0.715781619 (0.000)	0.81817303 (15.09)	
400	0.025	0.659784355 (0.015)	0.69382393 (16.38)	-5.159
10000	0.005	0.652226764 (1.045)	0.65902720 (16.49)	-1.043
90000	1/600	0.650966899 (155.2)	0.65323800 (16.54)	-0.349
n	q	Exact-KS-FFT	W&A(b)	Rel.err. (%)
25	0.10	0.715781619 (0.000)	0.65034803 (16.57)	
400	0.025	0.659784355 (0.015)	0.65034803 (16.57)	1.430
10000	0.005	0.652226764 (1.045)	0.65034803 (16.57)	0.288
90000	1/600	0.650966899 (155.2)	0.65034803 (16.57)	0.095
$\lambda = 0.2$				
n	q	Exact-KS-FFT	W&A(a)	Rel.err. (%)
25	0.04	0.992964654 (0.000)	0.99996560 (11.24)	
400	0.01	0.994406641 (0.015)	0.99733230 (12.10)	-0.294
10000	0.002	0.993769635 (0.967)	0.99457260 (12.24)	-0.081
90000	1/1500	0.993672471 (151.6)	0.99396553 (12.34)	-0.029
n	q	Exact-KS-FFT	W&A(b)	Rel.err. (%)
25	0.04	0.992964654 (0.000)	0.99363783 (12.35)	
400	0.01	0.994406641 (0.015)	0.99363783 (12.35)	0.077
10000	0.002	0.993769635 (0.967)	0.99363783 (12.35)	0.013
90000	1/1500	0.993672471 (151.6)	0.99363783 (12.35)	0.003
$\lambda = 0.1$				

Continued on next page

Table 2.6 – continued from previous page

n	q	Exact-KS-FFT	W&A(a)	Rel.err. (%)
25	0.020	0.999999999 (0.000)	1 (7.52)	
400	0.005	0.999974260 (0.015)	0.99999750 (10.80)	-0.002
10000	0.001	0.999966686 (0.951)	0.99997897 (11.04)	-0.001
90000	1/3000	0.999965549 (150.1)	0.99997020 (11.14)	0.000
n	q	Exact-KS-FFT	W&A(b)	Rel.err. (%)
25	0.020	0.999999999 (0.000)	0.99996560 (11.22)	
400	0.005	0.999974260 (0.015)	0.99996560 (11.22)	0.001
10000	0.001	0.999966686 (0.951)	0.99996560 (11.22)	0.000
90000	1/3000	0.999965549 (150.1)	0.99996560 (11.22)	0.000

Example 2.3.5. Next, we consider another illustrative example where we compare the performance of the proposed Exact-KS-FFT method with the R function `ks.test` from the package **dgof** (Arnold and Emerson, 2011). Hypothesizing that the underlying $F(x)$ in (2.1) follows a discrete uniform distribution on $[1,10]$, we have simulated random samples of size n, $25 \leq n \leq 100000$, from the discrete uniform distribution on $[1,10]$ and have performed KS tests on the simulated samples. In Table 2.7, we compute p values corresponding to different values of the test statistic D_n for the simulated samples of size n.

In order to perform the one-sample two-sided KS test, when $F(x)$ is purely discrete, one needs to input the `disc_ks_test(x, y, ..., exact = NULL, tol = 1e - 08, sim.size = 1e + 06, num.sim = 10)`, where `x` is a numeric vector of data sample values, and where other arguments are defined similarly as in the function `disc_ks_c_cdf()`. For instance, in order to calculate the p value for the KS test, when $F(x)$ follows a discrete uniform distribution on $[1,10]$ as in Example 2.3.5, with a sample size $n = 1000$, one should run the following R code.

Table 2.7 p values obtained via the Exact-KS-FFT method, the R function `ks.test`, and W&A(a) method, when the underlying CDF $F(x)$ follows a discrete uniform distribution on $[1,10]$. Numbers in () are run times in seconds.

n	$D_n = d_n$	Exact-KS-FFT		`ks.test`		`ks.test`(simulation)		W&A(a)	
25	0.2	0.1523	(0.0000)	0.1523	(0.007)	0.1465	(0.79)	0.1910	(12.63)
30	0.2	0.1133	(0.0000)	0.1133	(0.007)	0.125	(0.84)	0.1194	(12.73)
50	0.22	0.007164	(0.0000)	0.007167	(0.014)	0.007	(1.10)	0.0078223	(13.36)
100	0.2	0.00021	(0.0000)	NU		0.0002	(4.10)	0.0002277	(13.80)
1000	0.02	0.5424	(0.0150)	NU		0.5385	(8.35)	0.5429	(11.08)
5000	0.0094	0.4779	(0.2340)	NU		0.509	(68.37)	0.4781	(10.92)
10000	0.0065	0.4975	(0.8890)	NU		0.4985	(123.98)	0.4977	(11.08)
100000	0.00241	0.3343	(118.85)	NU		-	-	0.3344	(11.80)

```
R> x4 <- sample(1 : 10, 1000, replace = TRUE)
R> disc_ks_test(x4, ecdf(1 : 10), exact = TRUE)
```

```
        One-sample Kolmogorov-Smirnov test

data: x4
D = 0.01, p-value = 0.97023
alternative hypothesis: two-sided
```

As can be seen from Table 2.7, the Exact-KS-FFT method produces exact p values for all sample sizes $25 \leq n \leq 100000$, whereas the function `ks.test` becomes numerically unstable (NU) for $n \geq 100$, as noted also by Arnold and Emerson (2011). To avoid instability, for large n the function `ks.test` allows for estimating p values via simulation, which may be insufficiently accurate or prohibitively time consuming, depending on the choice of the number of simulations (cf., the column `ks.test`(simulation) in Table 2.7 where the number of simulations is 2000). In contrast to the `ks.test` function, using the Exact-KS-FFT method, one obtains the exact p value 0.3343 for sample size $n = 100000$ in less than 2 minutes without any simulation. Moreover, note that the p values in the column `ks.test`(simulation) in Table 2.7 are based on the suggested default number of 2000 replicates (i.e., obtained by implementing the R code `dgof::ks.test(x, ecdf(1 : 10), simulated.p.value = TRUE, B = 2000)`). Thus, each estimated p value is likely to be different if we run another simulation and the relative error will also vary substantially, as we demonstrate in Table 2.8. To reduce the variation of the simulated p values, one may wish to increase the number of simulations but that will increase even more the computation time and make it pro-

hibitive even for $n > 1000$. In addition, mainly due to the way it has been implemented, for $n > 1000$ the number of simulations cannot be significantly increased, e.g., go beyond 4000 replicates.

Table 2.8 Differences between the exact and simulated values of $\mathsf{P}(D_n \geq q)$ obtained via the Exact-KS-FFT method and the R function `ks.test`, respectively, for certain $n > 100$ and q, when the underlying $F(x)$ follows Binomial$(3, 0.5)$ or Binomial$(7, 0.5)$ distribution.

$F(x)$	n, q	Exact-KS-FFT	`ks.test`(simulation)	Rel.err.
Binomial$(3, 0.5)$	10000, 0.02	0.0000659	0	100%
			0.0005	658%
			0	100%
Binomial$(3, 0.5)$	400, 0.05	0.05612	0.050	10.9%
			0.061	8.7%
			0.069	22.9%
Binomial$(7, 0.5)$	10000, 0.01	0.07093	0.0760	7.14%
			0.0895	26.2%
			0.0745	5.03%
Binomial$(7, 0.5)$	400, 0.05	0.07490	0.0825	10.1%
			0.0910	21.5%
			0.0885	18.2%

As can be seen from Table 2.8 (which extends Tables 2.4 and 2.5) and as also supported by many additional calculations we have run, even for $n \leq 10000$ the accuracy of the R function `dgof::ks.test` may vary substantially for p values in the (rather important) range $(0, 0.1)$.

For small, moderate to large sample sizes (e.g., $25 \leq n \leq 10000$), looking at the column W&A(a) of Table 2.7, one can see that the alternative MC simulation-based W&A(a) method produces less accurate results and can be significantly slower than the Exact-KS-FFT method. W&A(a) performs better in terms of the trade-off between accuracy and speed for very large sample sizes, e.g., $n = 100000$.

To conclude, the proposed method outperforms the R function `ks.test` from the package **dgof** in all of the tested cases. When the number of jumps in the underlying $F(x)$ is small, the asymptotic p value obtained from (2.22) may not be a good estimate

unless sample sizes are very large (e.g., ≥ 40000). Whereas when the number of jumps in $F(x)$ is large, one may use the asymptotic p values to approximate the exact ones for large samples. In the next section, we turn our attention to the case of KS tests with continuous null distributions, which has been widely studied in the literature and for which very efficient numerical procedures have been recently developed.

2.3.3 (Complementary)CDF of D_n when $F(x)$ is continuous

Our purpose in this section is to illustrate the numerical performance of the proposed FFT-based approach of Section 2.2.1 and compare it with the state-of-the-art routines of Simard and L'Ecuyer (2011) and Carvalho (2015) developed especially for the case when the underlying CDF, $F(x)$, is strictly continuous. These authors have summarized and enhanced further the most accurate and efficient methods for computing the distribution of D_n for $F(x)$ continuous, developed earlier in a series of papers e.g., by Durbin (1968), Durbin (1973), Pomeranz (1974), Ruben and Gambino (1982), Marsaglia et al. (2003) and Brown and Harvey (2008). For comparison and further details on the implementations of these methods in various statistical softwares, we refer to Simard and L'Ecuyer (2011) and Brown and Harvey (2007). In their recent paper, Simard and L'Ecuyer (2011) have combined into one state-of-the-art program different exact methods to compute the distribution of D_n for different combinations of n and q, based on the relative efficiency and accuracy of the methods. Moreover, for certain combinations of n and q, where the implementations of the exact methods break down (due to cancellation errors, loss of precision and/or prohibitive running time), e.g., for very large n or when the CDF of D_n is close to one, Simard and L'Ecuyer (2011) incorporate in their program various asymptotic formulae for the limiting distribution of D_n. We refer the reader to Section 4 in Simard and L'Ecuyer (2011) for further details. More recently, Carvalho (2015), by avoiding the direct calculation of powers of matrices as required by the approach of Durbin (1973), developed the R package **kolmim** with function `pkolmim` that produces results with similar accuracy as those obtained by the routine of Marsaglia et al. (2003), but much faster. However, the related

R function becomes too slow when $n > 10000$ as the running time is proportional to n^3 on average. We will show this in Appendix A.4.

Let us reemphasize that the proposed FFT-based method developed in Section 2.2.1 is general and thus, applicable also for the case when $F(x)$ is continuous. Hypothesizing on a continuous distribution $F(x)$ leads to certain simplifications. In particular, (2.3) of Step 1 simplifies to

$$\mathsf{P}(D_n \geq q) = 1 - \mathsf{P}\left(\frac{i}{n} - q \leq U_{(i)} \leq \frac{i-1}{n} + q,\ 1 \leq i \leq n\right), \tag{2.27}$$

which confirms that the distribution of D_n no longer depends on $F(x)$. Also, (2.5) of Step 2 simplifies to (2.27) since the boundaries in (2.6) become $g(t) = nt - nq$ and $h(t) = nt + nq$, $q \geq 0$ as shown by Durbin (1968). This special case of the proposed FFT-based method has been considered by Moscovich and Nadler (2017) in the general context of computing the probability of non-crossing an upper and a lower boundaries by a Poisson process.

Similarly to Simard and L'Ecuyer (2011) (see Sections 4 and 5 therein), we consider three regions of n, (i) $n \leq 140$, (ii) $140 < n \leq 10^5$, and (iii) $n > 10^5$, forming various sub-regions with respect to q, as specified in Appendix A.2 and Appendix A.3. Within these sub-regions Simard and L'Ecuyer (2011) use different methods to compute the distribution of D_n. We have performed a thorough numerical comparison across these regions with details given in Appendix A.2 and Appendix A.3, and can report that, with only a few exceptions, the Exact-KS-FFT method returns values that are of at least the same precision as those obtained from the R or C program.

2.4 Conclusions

We have provided a fast and accurate method to compute $\mathsf{P}(D_n \geq q)$ when $F(x)$ is arbitrary, discontinuous (i.e., mixed or purely discrete) or continuous. The approach we take is to express $\mathsf{P}(D_n \geq q)$ as an appropriate rectangle probability for uniform order statistics and to compute the latter probability using the FFT method. We demonstrate

that the proposed Exact-KS-FFT method is numerically efficient and robust when hypothesizing on either discontinuous or continuous $F(x)$. In particular, when $F(x)$ is purely discrete the proposed method outperforms in terms of speed and accuracy the R function of Arnold and Emerson (2011), especially for large sample sizes. Furthermore, in the case of continuous $F(x)$ the Exact-KS-FFT method represents a viable alternative to the state-of-the-art methods of Simard and L'Ecuyer (2011) and Carvalho (2015) as it returns values that are of at least the same precision. In the case when $F(x)$ is mixed, to the best of our knowledge no alternative methods have been proposed in the literature to compute the exact distribution of D_n.

In this paper, we have also derived a useful extension of Schmid (1958)'s asymptotic formula, relaxing his requirement for $F(x)$ to be increasing between jumps and thus allowing for any general mixed or purely discrete $F(x)$. As demonstrated numerically, the extended asymptotic formula provides reasonably close approximations to the exact values of $\mathsf{P}(D_n \geq q)$ and can successfully be used for small to moderate number of jumps in $F(x)$ and large sample sizes.

As part of a separate ongoing research, we have also demonstrated that the FFT-based method can be successfully applied to compute the complementary CDF of the weighted version of the KS test statistic

$$K_n = \sup_x \sqrt{n}|F_n(x) - F(x)|\sqrt{\psi[F(x)]},$$

where $\psi(t) \geq 0, \forall t \in [0,1]$ is a weight function, first considered by Anderson and Darling (1952). The result of this additional research will appear elsewhere. Finally, as noted in Remark A.1.3, the complementary CDFs $\mathsf{P}(D_n \geq q)$ and $\mathsf{P}(D_n > q)$ are non-increasing functions with jumps at some values of q. Characterizing in detail the distribution of D_n, in particular the points of discontinuity, in relation to $F(x)$ is also a subject of ongoing research.

Appendix A

Appendix for Chapter 2

A.1 Expressing complementary CDFs of D_n

In this appendix, we express $\mathsf{P}(D_n > q)$ and $\mathsf{P}(D_n \geq q)$ in terms of a rectangle probability with respect to the uniform order statistics.

Lemma A.1.1. *The following holds true*

$$\mathsf{P}(D_n > q) = 1 - \mathsf{P}(\tilde{A}_i \leq U_{(i)} \leq \tilde{B}_i, 1 \leq i \leq n),$$

where $\tilde{A}_i = F\Big(\big(F^{-1}(\frac{i}{n} - q)\big) - \Big)$ *and* $\tilde{B}_i = F\Big(F^{-1}\big((\frac{i-1}{n} + q) + \big)\Big)$ *and* $F^{-1}(y+) = \lim_{\varepsilon \downarrow 0} F^{-1}(y+\varepsilon)$.

Proof: We have

$$
\begin{aligned}
&\mathsf{P}(D_n > q) \\
&= \mathsf{P}(\sup_{-\infty<x<\infty} |F_n(x) - F(x)| > q) \\
&= 1 - \mathsf{P}(\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \leq q) \\
&= 1 - \mathsf{P}(|F_n(x) - F(x)| \leq q, \text{ for all } x) \\
&= 1 - \mathsf{P}(-q \leq F_n(x) - F(x) \leq q, \text{ for all } x) \\
&= 1 - \mathsf{P}(F(x) - q \leq F_n(x) \leq F(x) + q, \text{ for all } x) \\
&= 1 - \mathsf{P}\Big(F(X_{(i)}-) - q \leq F_n(X_{(i-1)}) \text{ and } F_n(X_{(i)}) \leq F(X_{(i)}) + q, \text{ for } 1 \leq i \leq n\Big) \\
&= 1 - \mathsf{P}\Big(F(X_{(i)}-) \leq \frac{i-1}{n} + q \text{ and } \frac{i}{n} - q \leq F(X_{(i)}), \text{ for } 1 \leq i \leq n\Big) \\
&= \mathsf{P}\Big(F(X_{(i)}-) > \frac{i-1}{n} + q \text{ or } \frac{i}{n} - q > F(X_{(i)}), \text{ for some } 1 \leq i \leq n\Big) \\
&= \mathsf{P}\left(F^{-1}\Big(\Big(\frac{i-1}{n} + q\Big) +\Big) < X_{(i)} \text{ or } F^{-1}\Big(\frac{i}{n} - q\Big) > X_{(i)} \text{ for some } 1 \leq i \leq n\right),
\end{aligned}
$$

where in the last equality we have applied that $u < F(x-)$ if and only if $F^{-1}(u+) < x$ and that $x < F^{-1}(u)$ if and only if $F(x) < u$ (see e.g., Lemma 1 (iii) and (v) of Gleser (1985)). Therefore, we now have

$$
\begin{aligned}
&\mathsf{P}(D_n > q) \\
&= 1 - \mathsf{P}\left(F^{-1}\Big(\frac{i}{n} - q\Big) \leq X_{(i)} \leq F^{-1}\Big(\Big(\frac{i-1}{n} + q\Big) +\Big) \text{ for } 1 \leq i \leq n\right) \\
&= 1 - \mathsf{P}\left(F\Big(\Big(F^{-1}\Big(\frac{i}{n} - q\Big)\Big) -\Big) \leq U_{(i)} \leq F\Big(F^{-1}\Big(\Big(\frac{i-1}{n} + q\Big) +\Big)\Big) \text{ for } 1 \leq i \leq n\right),
\end{aligned}
\tag{A.1}
$$

where in the last equality we have applied Lemma 1 of Dimitrova et al. (2017). The statement now follows noting that one can rewrite the last equality in terms of $\tilde{A}_i$ and $\tilde{B}_i$.

□

Remark A.1.2. The fact that the non-crossing probability

$$\begin{aligned}&\mathsf{P}(F(x)-q\le F_n(x)\le F(x)+q,\ \text{for all } x)\\&=\mathsf{P}\left(F^{-1}\left(\frac{i}{n}-q\right)\le X_{(i)}\le F^{-1}\left(\left(\frac{i-1}{n}+q\right)+\right)\ \text{for } 1\le i\le n\right)\end{aligned}$$

shown in the proof of Lemma A.1.1 is illustrated in Figure A.1 with $F(x)$ (the green piecewise linear function) defined in (2.23) (cf., Example 2.2.8), for $n = 5$.

Remark A.1.3. The statement of Lemma A.1.1 holds true also for $\mathsf{P}(D_n \ge q)$, as stated in (2.3), with A_i and B_i defined as in (2.4). The proof is similar but more involved than that of Lemma A.1.1 and is therefore omitted. It should also be noted that the complementary CDFs $\mathsf{P}(D_n \ge q)$ and $\mathsf{P}(D_n > q)$ are non-increasing functions with jumps at some values of q. In fact, these two functions coincide, except at the jumps where $\mathsf{P}(D_n \ge q)$ is left-continuous and $\mathsf{P}(D_n > q)$ is right-continuous. This is a consequence of the fact that the pairs A_i, B_i and $\tilde{A}_i$, $\tilde{B}_i$ coincide except at their points of discontinuity, where A_i, B_i are correspondingly right- and left- continuous, whereas $\tilde{A}_i$, $\tilde{B}_i$ are correspondingly left- and right- continuous.

Remark A.1.4. Let us note that the result of Lemma A.1.1 coincides with Theorem 1 of Gleser (1985).

A.2 Computing the CDF of D_n when $F(x)$ is continuous

In this appendix, we compute the values of the CDF $\mathsf{P}(D_n \le q)$ for different n and q using the Exact-KS-FFT method and compare the results to those obtained with the C program due to Simard and L'Ecuyer (2011) and R function `pkolmim` from the package **kolmim** by Carvalho (2015), which is claimed to be highly efficient and precise. Hence, we calculate an absolute error as the absolute difference between our results and the R outputs, from which we can infer the number of decimal digits of precision of our results.

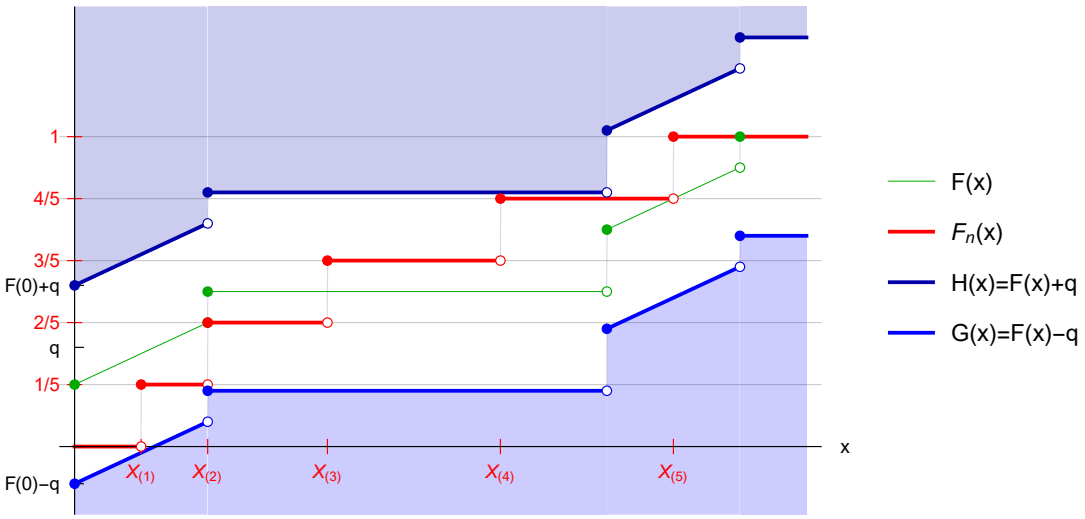


Fig. A.1 Illustration of the equivalence of $\mathsf{P}(F(x)-q \le F_n(x) \le F(x)+q, \text{ for all } x)$ to $\mathsf{P}\left(F^{-1}(\frac{i}{n}-q) \le X_{(i)} \le F^{-1}\left((\frac{i-1}{n}+q)+\right) \text{ for } 1 \le i \le n\right)$ (cf., Remark A.1.2), for $F(x)$ defined as in (2.23) with $n = 5$.

In order to compute $\mathsf{P}(D_n \le q)$, when $F(x)$ is continuous using the R package **KSgeneral**, one needs to input `cont_ks_cdf(q, n)`. For example, in order to compute the value for $\mathsf{P}(D_n \le q)$, when $F(x)$ is continuous, for $n = 40$, $nq^2 = 0.76$, one should run the following R code and obtain the corresponding result as shown in Table A.3 for $n = 40$ in the column Exact-KS-FFT.

```
R> cont_ks_cdf(sqrt(0.76/40), 40)
```

```
[1] 0.6032371
```

Simard and L'Ecuyer (2011) consider the following regions: 1) $n \le 140$ and $q \le 1/n$; 2) $n \le 140$ and $q \ge 1-1/n$; 3) $n \le 140$ and $1/n < nq^2 < 0.754693$; 4) $n \le 140$ and $0.754693 \le nq^2 < 4$; 5) $n \le 140$ and $4 \le nq^2 < 18$; 6) $n \le 140$ and $nq^2 \ge 18$; 7) $140 < n \le 10^5$ and $nq^{3/2} < 1.4$; 8) $140 < n \le 10^5$ and $nq^{3/2} \ge 1.4$; and 9) $n > 10^5$ where they use different methods to compute the distribution of D_n (cf., Simard and L'Ecuyer, 2011, Section 4).

Following the segmentation of regions, we have computed the distribution of D_n with the proposed FFT-based method and can report that for regions 1), 2), 3), 4), 7), our approach gives results that are of at least the same precision as those obtained from

the R or C program. In regions 5) and 6), when $n \leq 140$ and $nq^2 > 12$, our approach may be unsuitable due to numerical instabilities which may occur.

More specifically, when 1) $n \leq 140$ and $q \leq 1/n$, or when 2) $n \leq 140$ and $q \geq 1-1/n$, Simard and L'Ecuyer (2011) use the Ruben and Gambino (1982) formula to calculate the distribution of D_n, returning results with at least 13 decimal digits of precision. As can be seen from Table A.1, in these regions our method gives results that are of similar accuracy as those from the R function `pkolmim` or the C program of Simard and L'Ecuyer (2011).

Table A.1 Values of $\mathsf{P}(D_n \leq q)$ for $q = 1/n$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	2.320196159531E-08	2.320196159531E-08	2.320196159531E-08	9.9262E-23
40	6.749093037884E-17	6.749093037884E-17	6.749093037884E-17	1.7010E-30
60	1.702549809333E-25	1.702549809333E-25	1.702549809333E-25	3.0076E-39
80	4.050687717856E-34	4.050687717855E-34	4.050687717855E-34	3.2928E-47
100	9.332621544394E-43	9.332621544394E-43	9.332621544394E-43	3.4092E-56
120	2.106901932614E-51	2.106901932614E-51	2.106901932614E-51	2.5994E-64
140	4.690131222300E-60	4.690131222299E-60	4.690131222299E-60	1.0004E-72

When 3) $n \leq 140$ and $1/n < nq^2 < 0.754693$, Simard and L'Ecuyer (2011) use the Durbin matrix algorithm to calculate the distribution of D_n, returning results with at least 13 decimal digits of precision. As can be seen from Table A.2, in this region our method gives results of at least the same accuracy.

Table A.2 Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 0.75$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.6089841201379	0.6089841201379	0.6089841201379	2.9936E-15
40	0.5951497241008	0.5951497241008	0.5951497241008	1.9984E-15
60	0.5888010590107	0.5888010590107	0.5888010590107	1.9984E-15
80	0.5849488429478	0.5849488429478	0.5849488429478	4.7962E-14
100	0.5822897960080	0.5822897960080	0.5822897960080	2.2093E-14
120	0.5803108927579	0.5803108927579	0.5803108927579	7.2053E-14
140	0.5787632928760	0.5787632928760	0.5787632928760	1.0991E-14

When 4) $n \leq 140$ and $0.754693 \leq nq^2 < 4$, Simard and L'Ecuyer (2011) use the Pomeranz (1974) method to calculate the distribution of D_n, returning results with at

least 13 decimal digits of precision. In this region, again our method gives results of at least the same accuracy as shown in Tables A.3 and A.4 for $nq^2 = 0.76$ and $nq^2 = 3.9$, respectively.

Table A.3 Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 0.76$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.6169412955836	0.6169412955835	0.6169412955835	2.9976E-15
40	0.6032370735674	0.6032370735674	0.6032370735674	7.9936E-15
60	0.5969494784897	0.5969494784898	0.5969494784897	9.9920E-16
80	0.5931349807275	0.5931349807274	0.5931349807274	4.2966E-14
100	0.5905022875562	0.5905022875562	0.5905022875562	3.0087E-14
120	0.5885431553286	0.5885431553286	0.5885431553285	6.0063E-14
140	0.5870111081551	0.5870111081552	0.5870111081551	1.3989E-14

Table A.4 Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 3.9$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.9995468293485	0.9995468293485	0.9995468293485	4.9960E-15
40	0.9994205337332	0.9994205337332	0.9994205337332	1.7097E-14
60	0.9993680770022	0.9993680770022	0.9993680770022	1.2990E-14
80	0.9993382289964	0.9993382289964	0.9993382289964	7.4940E-14
100	0.9993185558110	0.9993185558110	0.9993185558110	3.9968E-14
120	0.9993044245859	0.9993044245858	0.9993044245857	1.1902E-13
140	0.9992936831012	0.9992936831013	0.9992936831012	1.9096E-14

When 5) $n \leq 140$ and $4 \leq nq^2 < 18$, Simard and L'Ecuyer (2011) first use the Miller (1956) approximation to estimate $\mathsf{P}(D_n \geq q)$, and then calculate the distribution of D_n by $\mathsf{P}(D_n \leq q) = 1 - \mathsf{P}(D_n \geq q)$. The authors claim that the approximated values of $\mathsf{P}(D_n \leq q)$ have 14 decimal digits of precision. As illustrated in Tables A.5 and A.6 for $nq^2 = 4.1$ and $nq^2 = 12$, our method gives results of at least the same accuracy when $n \leq 140$ and $4 \leq nq^2 \leq 12$. For $n \leq 140$ and $12 < nq^2 < 18$, since our implementation uses floating numbers in C++, numerical instabilities may occur.

When 6) $n \leq 140$ and $nq^2 \geq 18$, $\mathsf{P}(D_n \geq q) < 5 \times 10^{-16}$. Equivalently, $\mathsf{P}(D_n \leq q) = 1 - \mathsf{P}(D_n \geq q) > 1 - 5 \times 10^{-16}$. Hence, returning $\mathsf{P}(D_n \leq q) = 1$ will give results with 15 decimal digits of precision.

Table A.5 Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 4.1$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.99970981546296	0.99970981546295	0.99970981546295	5.3291E-15
40	0.99962025405236	0.99962025405235	0.99962025405235	1.6209E-14
60	0.99958292108831	0.99958292108830	0.99958292108830	1.6764E-14
80	0.99956168530875	0.99956168530868	0.99956168530868	7.5717E-14
100	0.99954770168480	0.99954770168484	0.99954770168484	4.3188E-14
120	0.99953766763972	0.99953766763961	0.99953766763961	1.1346E-13
140	0.99953004813548	0.99953004813546	0.99953004813546	1.8430E-14

Table A.6 Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 12$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	0.99999999999963	0.99999999999962	0.99999999999962	7.5495E-15
40	0.99999999999135	0.99999999999134	0.99999999999134	1.5210E-14
60	0.99999999998168	0.99999999998167	0.99999999998167	1.3656E-14
80	0.99999999997415	0.99999999997407	0.99999999997407	7.6827E-14
100	0.99999999996823	0.99999999996827	0.99999999996827	3.8192E-14
120	0.99999999996388	0.99999999996376	0.99999999996376	1.1702E-13
140	0.99999999996020	0.99999999996017	0.99999999996017	2.3981E-14

When 7) $140 < n \leq 10^5$ and $nq^{3/2} < 1.4$, Simard and L'Ecuyer (2011) use the Durbin matrix algorithm to obtain the exact distribution of D_n, returning probabilities with at least 13 decimal digits of precision. As illustrated in Table A.7, our method returns values of at least the same accuracy.

Table A.7 Values of $\mathsf{P}(D_n \leq q)$ for $nq^{3/2} = 1.3$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
140	6.378698330645E-02	6.378698330644E-02	6.378698330644E-02	9,9920E-16
200	3.847020660831E-02	3.847020660831E-02	3.847020660831E-02	4.9960E-16
500	7.365490405433E-03	7.365490405433E-03	7.365490405433E-03	3.9899E-17
1000	1.383862966203E-03	1.383862966202E-03	1.383862966202E-03	3.7015E-16
2000	1.629201120187E-04	1.629201120188E-04	1.629201120188E-04	1.5501E-16
5000	3.811342214264E-06	3.811342214276E-06	3.811342214276E-06	1.1910E-17
10000	8.999089573402E-08	8.999089573401E-08	8.999089573401E-08	1.2308E-20
100000	5.388085736386E-17	5.388085736343E-17	5.388085736345E-17	4.0739E-28

In region 8), when $140 < n \leq 10^5$, $nq^{3/2} \geq 1.4$, and $nq^2 \leq 18$, Simard and L'Ecuyer (2011) apply the Pelz and Good (1976) approximation that gives five decimal digits of

precision for values of $\mathsf{P}(D_n \leq q)$. In contrast, when $140 < n \leq 10^5$, $nq^{3/2} \geq 1.4$, and $nq^2 \leq 10$, our approach gives results with at least 11 decimal digits of precision even though it is using floating numbers in calculation. The results when $nq^{3/2} = 1.4$ and when $nq^2 = 10$ are shown in Tables A.8 and A.9, respectively. However, in region 8), when $140 < n \leq 10^5$, $nq^{3/2} \geq 1.4$, and $nq^2 > 10$, our approach may be unsuitable due to numerical instabilities. In particular, it will return results with at least 11 decimal digits of precision, but the resulting values of $\mathsf{P}(D_n \leq q)$ may not be decreasing in n, due to the errors in calculations with floating numbers. When $140 < n \leq 10^5$ and $nq^2 \geq 18$, returning $\mathsf{P}(D_n \leq q) = 1$ will give results with 15 decimal digits of precision.

Table A.8 Values of $\mathsf{P}(D_n \leq q)$ for $nq^{3/2} = 1.4$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
140	9.0262329475006E-02	9.025921823E-02	9.0262329475004E-02	1.9013E-15
500	1.3024254002106E-02	1.302426466E-02	1.3024254002106E-02	4.0072E-16
1000	2.8949372516988E-03	2.89496818E-03	2.8949372516981E-03	6.7004E-16
5000	1.4235508314598E-05	1.42356151E-05	1.4235508314645E-05	4.7100E-17
10000	4.8334541076751E-07	4.83345438E-08	4.8334541076707E-07	4.3506E-19
50000	3.7148003980197E-12	3.71479094E-12	3.7147909440549E-12	9.4540E-18
100000	2.2123605255202E-15	2.21229903E-15	2.2123605254766E-15	4.3560E-26

Table A.9 Values of $\mathsf{P}(D_n \leq q)$ for $nq^2 = 10$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
141	0.99999999743970	0.99999999743965	0.99999999743964	5.6066E-14
500	0.99999999654196	0.99999999654197	0.99999999654196	9.9920E-16
1000	0.99999999629650	0.99999999629831	0.99999999629630	1.9806E-13
5000	0.99999999602730	0.99999999603085	0.99999999603074	3.4379E-12
10000	0.99999999597940	0.99999999597986	0.99999999597981	4.1001E-13
50000	0.99999999592690	0.99999999591965	0.99999999591967	7.2330E-12
100000	0.99999999592133	0.99999999590672	0.99999999590684	1.4486E-11

Finally, in region 9), Simard and L'Ecuyer (2011) apply the Pelz and Good (1976) approximation to obtain values of $\mathsf{P}(D_n \leq q)$ when $nq^2 < 18$, and set $\mathsf{P}(D_n \leq q) = 1$ when $nq^2 \geq 18$. As illustrated in Table A.10 for $n = 100001$, our approach tends to be more accurate when $\mathsf{P}(D_n \leq q)$ is very small. However, Pelz and Good (1976) approximation may provide higher accuracy when $\mathsf{P}(D_n \leq q)$ tends to one.

Table A.10 Values of $\mathsf{P}(D_n \leq q)$ for $n = 100001$.

q	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
$\frac{1}{10\sqrt{n}}$	2.350089150939E-52	2.269812367E-52	2.350089151281E-52	3.4177E-62
$\frac{1}{8\sqrt{n}}$	1.969026572915E-33	1.962478061E-33	1.969026573193E-33	2.7816E-43
$\frac{1}{6\sqrt{n}}$	1.018454527586E-18	1.018350563E-18	1.018454527742E-18	1.5595E-28
$\frac{1}{4\sqrt{n}}$	2.907074248741E-08	2.9070737934E-08	2.907074249157E-08	4.1588E-18
$\frac{1}{2\sqrt{n}}$	3.639199759592E-02	3.639199759592E-02	3.639199760172E-02	5.7979E-12
$\frac{1}{\sqrt{n}}$	7.305646850557E-01	7.305646847185E-01	7.305646847159E-01	3.3980E-10
$\frac{2}{\sqrt{n}}$	9.993319331457E-01	9.993319333086E-01	9.993319333086E-01	1.6290E-10

To conclude, apart from the regions where $n \leq 140$ and $12 < nq^2 < 18$; or $140 < n \leq 10^5$, $nq^{3/2} \geq 1.4$, and $10 < nq^2 < 18$; or $nq^2 \geq 18$, the Exact-KS-FFT method returns values of $\mathsf{P}(D_n \leq q)$ that are at least as accurate as those obtained by Simard and L'Ecuyer (2011). This is shown in Figure A.2. Moreover, for $n > 10^5$, the proposed method may be accurate when $\mathsf{P}(D_n \leq q)$ is very small.

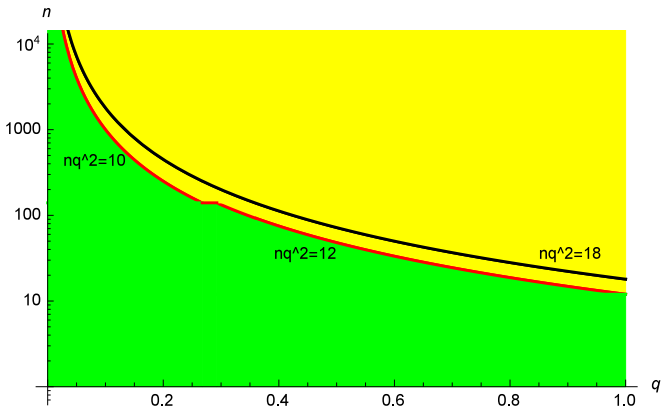


Fig. A.2 Approximate regions where the Exact-KS-FFT method returns $\mathsf{P}(D_n \leq q)$ efficiently and accurately.

A.3 Computing the complementary CDF when $F(x)$ is continuous

It is well known that

$$D_n = \sup_x |F_n(x) - F(x)| \longrightarrow 0 \quad a.s.,$$

as $n \to \infty$. Hence, when n is very large, $\mathsf{P}(D_n \leq q)$ is close to one. Also, it can be seen that $D_n \in [0,1]$, so $\mathsf{P}(D_n \leq q)$ is close to one when q is close to one. In these cases, cancellation errors may occur when trying to numerically compute the p value

$$\mathsf{P}(D_n \geq q) = 1 - \mathsf{P}(D_n \leq q). \tag{A.2}$$

Similarly to previous section, we compute the values of $\mathsf{P}(D_n \geq q)$ for different n and q using the Exact-KS-FFT method and compare the results to those obtained with the R program of Carvalho (2015), and the C program due to Simard and L'Ecuyer (2011).

In order to compute $\mathsf{P}(D_n \geq q)$, when $F(x)$ is continuous using the R package **KSgeneral**, one needs to input `cont_ks_c_cdf(q, n)`. For instance, in order to compute the value for $\mathsf{P}(D_n \geq q)$, for $n = 141$, $nq^2 = 2.1$, one should run the following R code and obtain the corresponding result as shown in Table A.12 for $n = 141$ in the column Exact-KS-FFT.

```
R> cont_ks_c_cdf(sqrt(2.1/141), 141)
```

```
[1] 0.02743689
```

Simard and L'Ecuyer (2011) consider the following regions: 1) $n \leq 140$ and $nq^2 < 4$; 2) $n \leq 140$ and $nq^2 \geq 4$; 3) $n > 140$ and $nq^2 < 2.2$; and 4) $n > 140$ and $nq^2 \geq 2.2$ where they use different methods to compute the complementary CDF of D_n (cf., Simard and L'Ecuyer, 2011, Section 5).

Following the segmentation of regions, we have computed the complementary CDF of D_n with the proposed FFT-based method. Consequently, we can report that for region

1), our approach gives results that are of at least the same accuracy as those obtained from the R or C program. In region 2), $\mathsf{P}(D_n \leq q)$ is close to one and our method may be unsuitable due to cancellation errors which may occur when calculating the complementary CDF via (A.2). A comparison for $nq^2 = 4$ is shown in Table A.11.

Table A.11 Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 4$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
20	3.627396978E-04	3.627396978E-04	3.627396978E-04	5.0590E-15
40	4.691487961E-04	4.691487961E-04	4.691487961E-04	1.5461E-14
60	5.134182982E-04	5.134182982E-04	5.134182982E-04	1.3937E-14
80	5.386021475E-04	5.386021476E-04	5.386021476E-04	7.4480E-14
100	5.551927328E-04	5.551927328E-04	5.551927328E-04	3.9403E-14
120	5.671032850E-04	5.671032851E-04	5.671032851E-04	1.0974E-13
140	5.761521040E-04	5.761521040E-04	5.761521040E-04	1.0433E-14

In region 3), when $140 < n \leq 10^5$ and $nq^2 < 2.2$, Simard and L'Ecuyer (2011) use the Pelz and Good (1976) approximation and apply (A.2) to calculate the complementary CDF, returning results with at least five decimal digits of precision. Our approach also applies (A.2), but returns results with at least nine decimal digits of precision. A comparison for $nq^2 = 2.1$ is given in Table A.12.

Table A.12 Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 2.1$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
141	0.02743688914	0.02743688914	0.02743688914	5.0990E-13
500	0.02866250067	0.02866250067	0.02866250073	5.9554E-11
1000	0.02905830855	0.02905830855	0.02905830828	2.6492E-10
5000	0.02957796836	0.02957796836	0.02957796797	3.9119E-10
10000	0.02969964497	0.02969964497	0.02969964418	7.9672E-10
50000	0.02986114255	0.02986114255	0.02986114263	7.2066E-11
100000	0.02989926133	0.02989926162	0.02989926162	2.8962E-10

In region 4), when $140 < n \leq 10^5$ and $nq^2 \geq 2.2$, Simard and L'Ecuyer (2011) use the Miller (1956) approximation and obtain complementary CDF with at least six decimal digits of precision. In this region, the proposed FFT-based method may give more accurate results when $140 < n \leq 10^5$ and $2.2 \leq nq^2 \leq 7$. For example, for $nq^2 = 2.2$

and $nq^2 = 7$, Tables A.13 and A.14 show that the Exact-KS-FFT method returns complementary CDF with at least 10 decimal digits of precision. When $140 < n \leq 10^5$ and $nq^2 > 7$, our method may be unsuitable due to cancellation errors as previously discussed.

Table A.13 Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 2.2$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
141	0.02239633302	0.0223963592	0.02239633302	5.2000E-14
500	0.02343606481	0.0234361007	0.02343606481	2.6201E-14
1000	0.02377033994	0.0237703789	0.02377033994	2.0260E-13
10000	0.02431016270	0.0243102062	0.02431016270	1.3636E-12
100000	0.02447768608	0.0244777310	0.02447768610	1.8812E-11

Table A.14 Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 7$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
141	1.2484862E-06	1.2484863E-06	1.2484863E-06	5.7535E-14
500	1.4796907E-06	1.4796906E-06	1.4796907E-06	2.1112E-14
1000	1.5434598E-06	1.5434599E-06	1.5434600E-06	1.9722E-13
10000	1.6309268E-06	1.6309265E-06	1.6309266E-06	1.9895E-13
100000	1.6534902E-06	1.6534983E-06	1.6534982E-06	7.9321E-12

Finally, when $n > 10^5$ and $nq^2 < 370$, Simard and L'Ecuyer (2011) use the Miller (1956) approximation and obtain complementary CDF with a few correct decimal digits. These authors have shown that complementary CDF can be set to be zero when $nq^2 \geq 370$. Recall that in Table A.10, we have shown that the Exact-KS-FFT method tends to be more accurate when $\mathsf{P}(D_n \leq q)$ is very small, or when q is small. In this case, we can apply (A.2) to calculate the complementary CDF, without incurring large cancellation errors. More specifically, when $n > 10^5$ and $nq^2 \leq 3$, the Exact-KS-FFT method returns complementary CDF with at least seven decimal digits of precision as demonstrated in Table A.15. The accuracy of course deteriorates when $n > 10^5$ and $3 < nq^2 < 370$.

To summarize, apart from the regions where $n \leq 140$ and $nq^2 \geq 4$; or $n \leq 140$ and $q \geq 1 - 1/n$; or $140 < n \leq 10^5$ and $nq^2 > 7$, the Exact-KS-FFT method returns values of

Table A.15 Values of $\mathsf{P}(D_n \geq q)$ for $nq^2 = 3$.

n	Exact-KS-FFT	Simard & L'Ecuyer	Carvalho	Abs. err.
100001	4.939303411E-03	4.939303336E-03	4.939303263053E-03	1.4795E-10
200000	4.944654927E-03	4.944654662E-03	4.944654584319E-03	3.4268E-10
300000	4.947020044E-03	4.947020013E-03	4.947019946709E-03	9.7291E-11

the probability $\mathsf{P}(D_n \geq q)$ that are at least as accurate as those obtained by Simard and L'Ecuyer (2011). This is shown in Figure A.3. Moreover, when $n > 10^5$ and $nq^2 \leq 3$, the proposed approach may be more accurate than Simard and L'Ecuyer (2011) method.

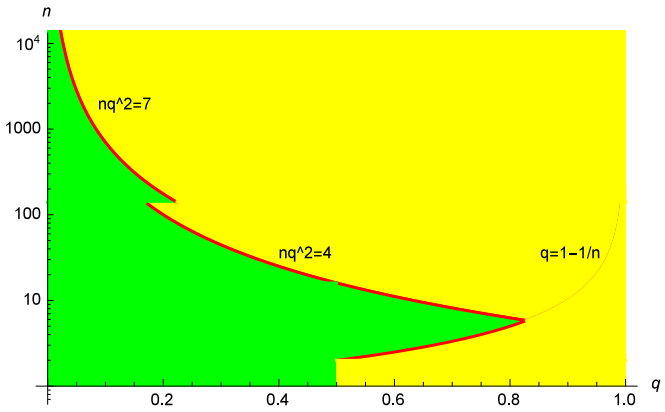


Fig. A.3 Approximate regions where the Exact-KS-FFT method returns $\mathsf{P}(D_n \geq q)$ efficiently and accurately

A.4 Speed comparison

Tables A.16, A.17 and A.18 report the CPU times to compute $\mathsf{P}(D_n \geq q)$ 100 times, for selected values of n and λ. Note that Carvalho (2015) procedure cannot be used with the chosen values of q and $n = 100000$ as it is prohibitively slow. As expected, Simard and L'Ecuyer (2011) C program which combines the most efficient methods for computing the distribution of D_n for $F(x)$ continuous, is the fastest among the three procedures. However, the Exact-KS-FFT method proves to be a viable alternative especially given its generality and applicability to the case of discontinuous $F(x)$.

Table A.16 CPU time (seconds) to compute $\mathsf{P}(D_n \geq q)$ 100 times with the Simard and L'Ecuyer (2011) C program.

$n\backslash\lambda$	0.25	0.5	1	2	3	4
10	0.00034	0.00059	0.00065	0.00069	0.00087	0.00014
100	0.00524	0.01318	0.01835	0.03242	0.04765	0.00107
140	0.00615	0.01915	0.03474	0.06172	0.08618	0.11874
141	0.00673	0.01955	0.03529	0.06657	0.09285	0.11886
1000	0.15040	0.00013	0.00014	0.00019	0.00894	0.00894
10000	0.00013	0.00014	0.00012	0.00015	0.08124	0.08080
100000	0.00014	0.00015	0.00014	0.00019	0.78912	0.75099

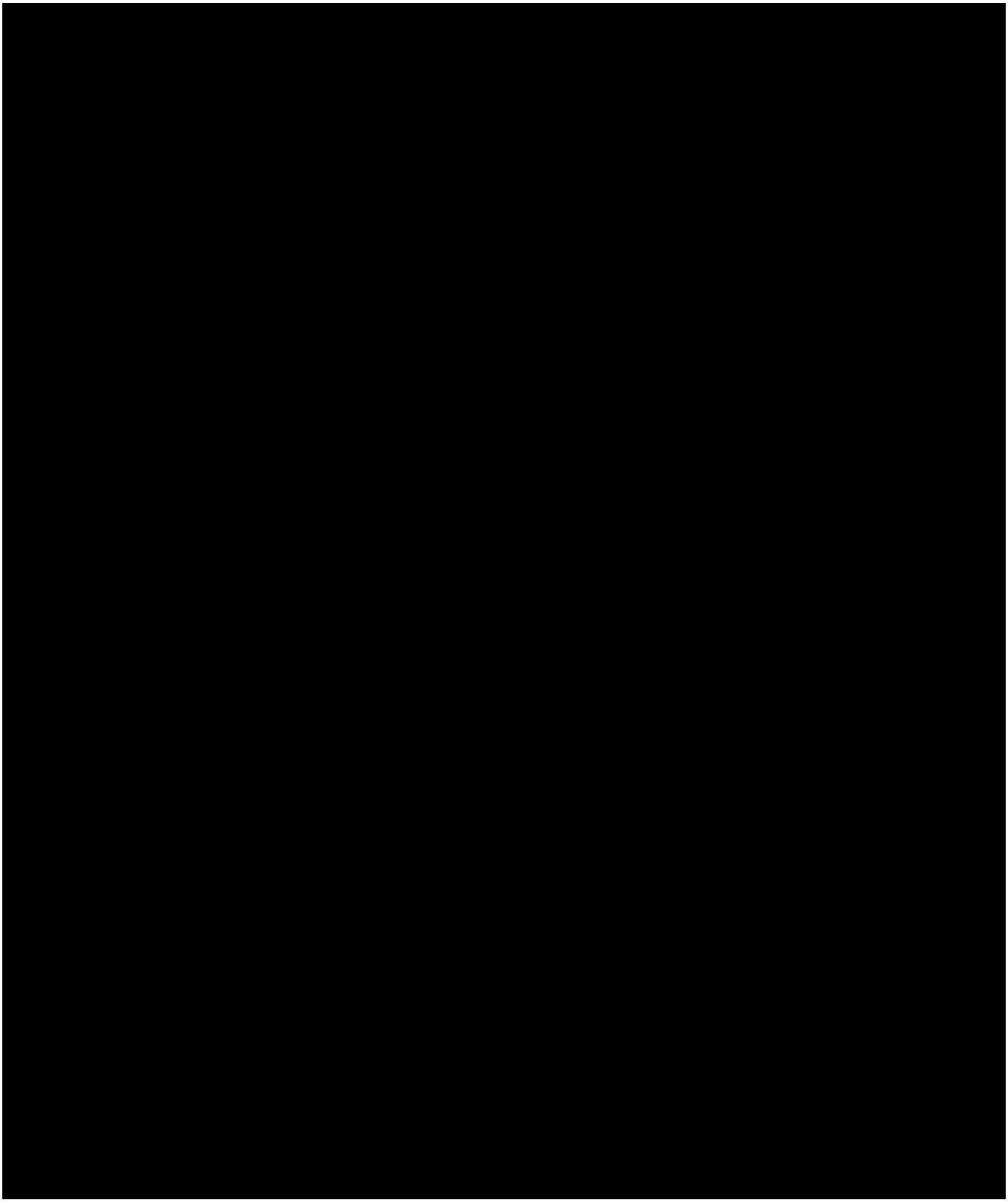
Table A.17 CPU time (seconds) to compute $\mathsf{P}(D_n \geq q)$ 100 times with the Exact-KS-FFT method.

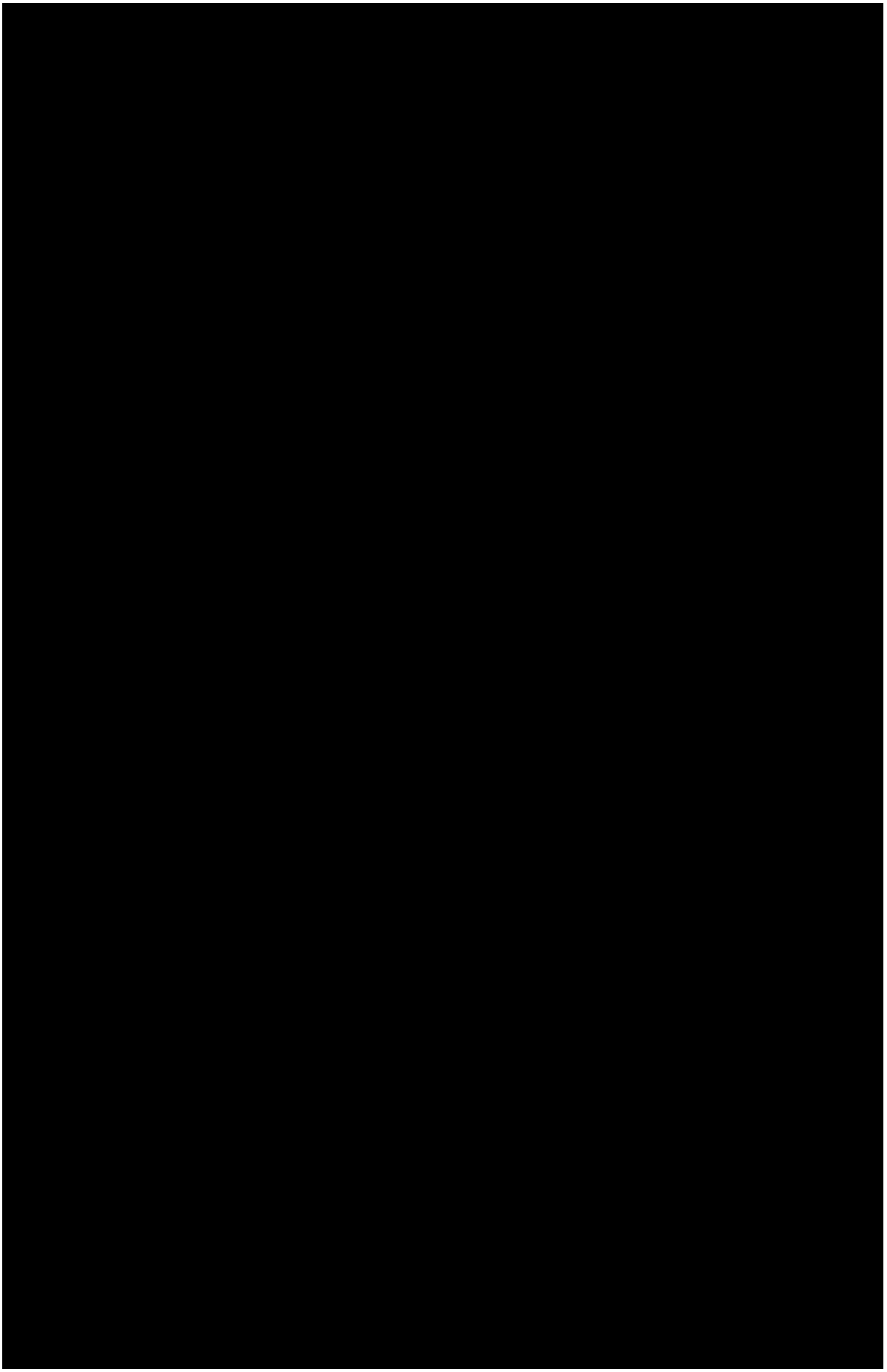
$n\backslash\lambda$	0.25	0.5	1	2	3	4
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	0.0150	0.0150	0.0150	0.0380	0.0380	0.0550
140	0.0150	0.0150	0.0310	0.0620	0.0780	0.1090
141	0.0150	0.0150	0.0310	0.0620	0.0780	0.1090
1000	0.1400	0.2960	0.6550	1.1700	1.9340	2.2990
10000	5.6310	8.5320	19.500	45.100	52.890	94.700
100000	182.29	333.31	672.16	1466.6	2503.3	3211.7

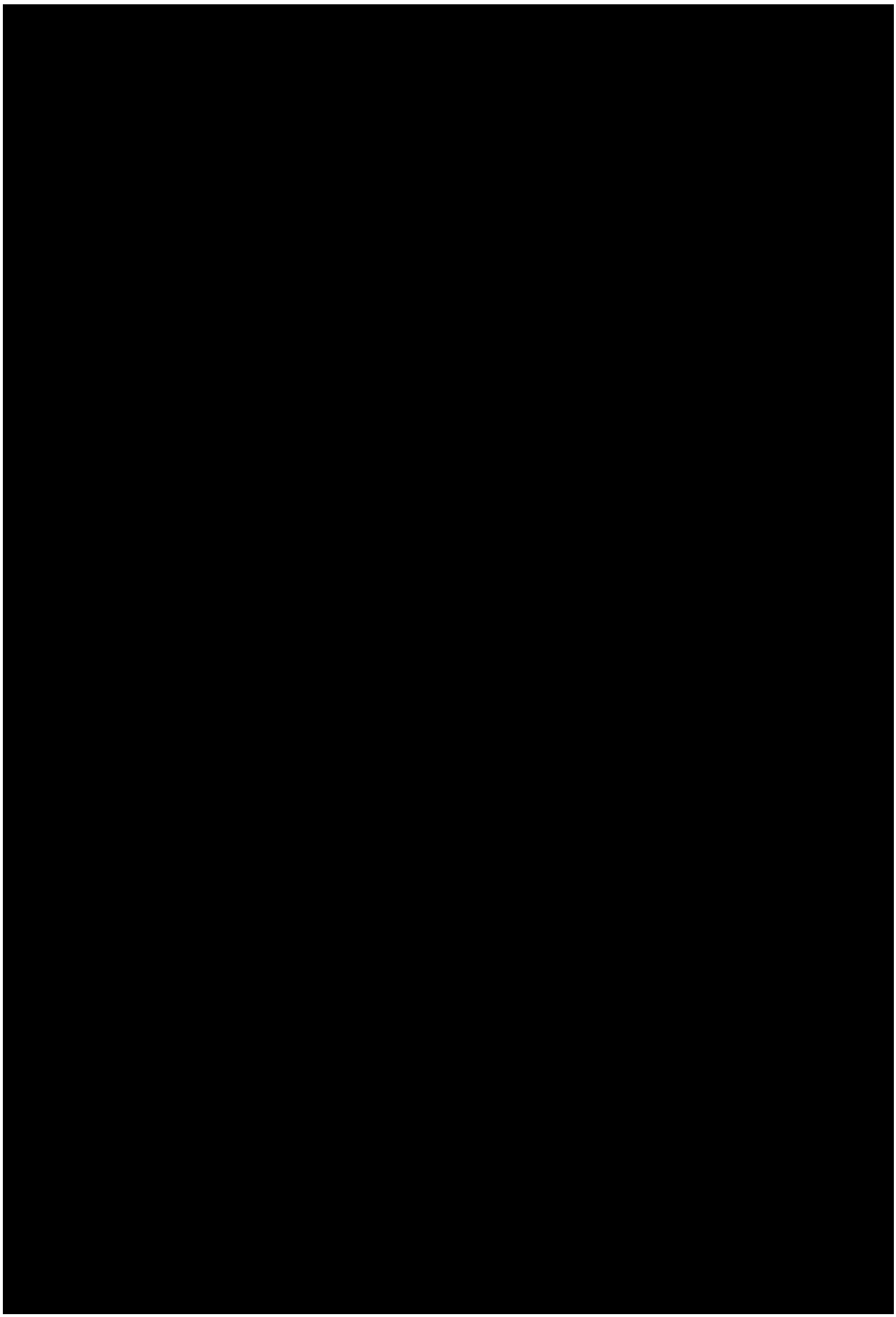
Table A.18 CPU time (seconds) to compute $\mathsf{P}(D_n \geq q)$ 100 times with the Carvalho (2015) R program.

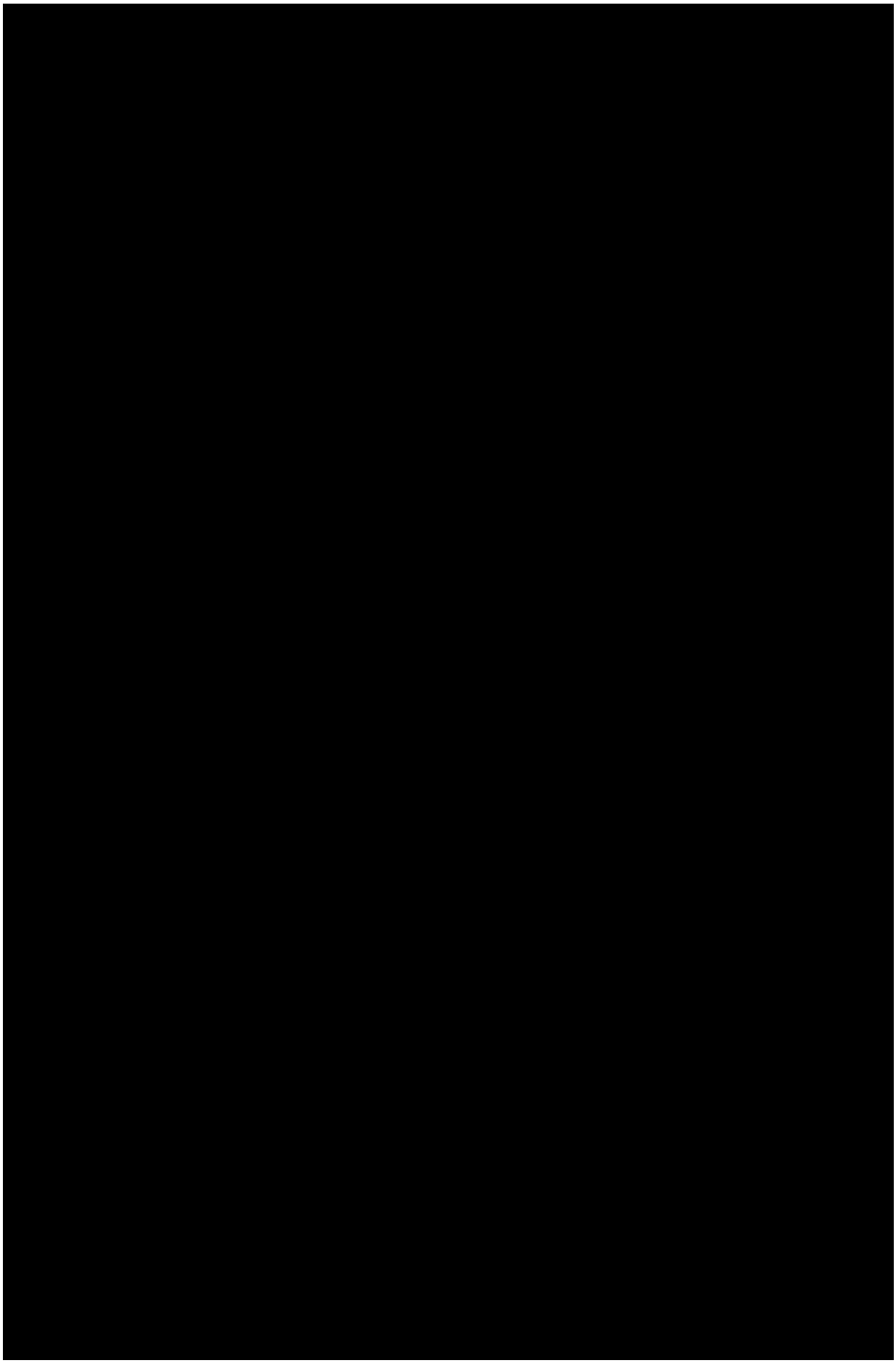
$n \backslash \lambda$	0.25	0.5	1	2	3	4
10	0.001	0.001	0.001	0.001	0.001	0.001
100	0.003	0.004	0.006	0.009	0.013	0.017
140	0.004	0.006	0.009	0.014	0.020	0.023
141	0.004	0.006	0.009	0.014	0.020	0.024
1000	0.086	0.155	0.268	0.499	0.747	1.066
10000	6.250	13.16	40.22	97.18	145.5	188.4
100000	na	na	na	na	na	na

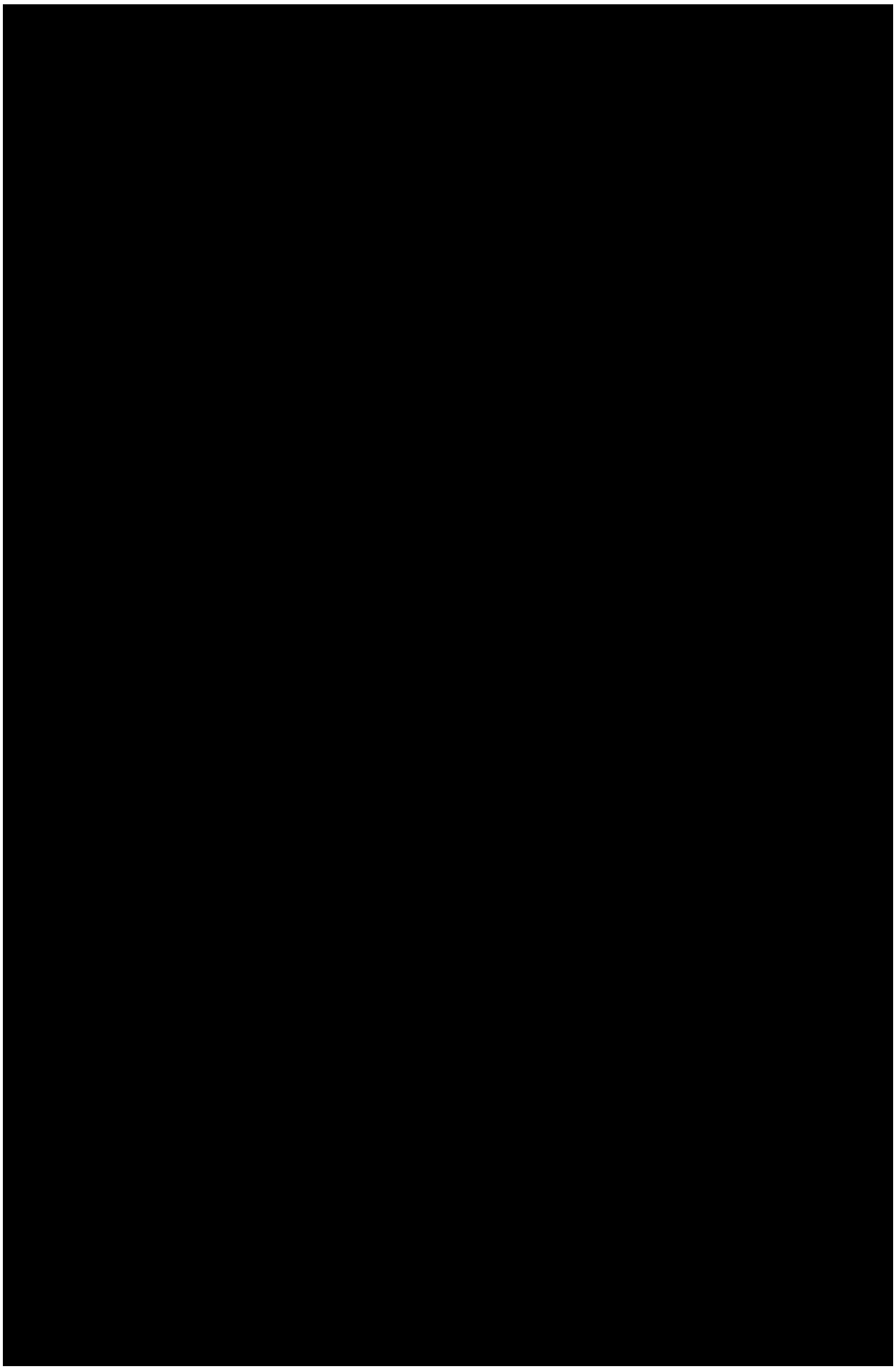
The full text of these chapters have been removed for copyright reasons

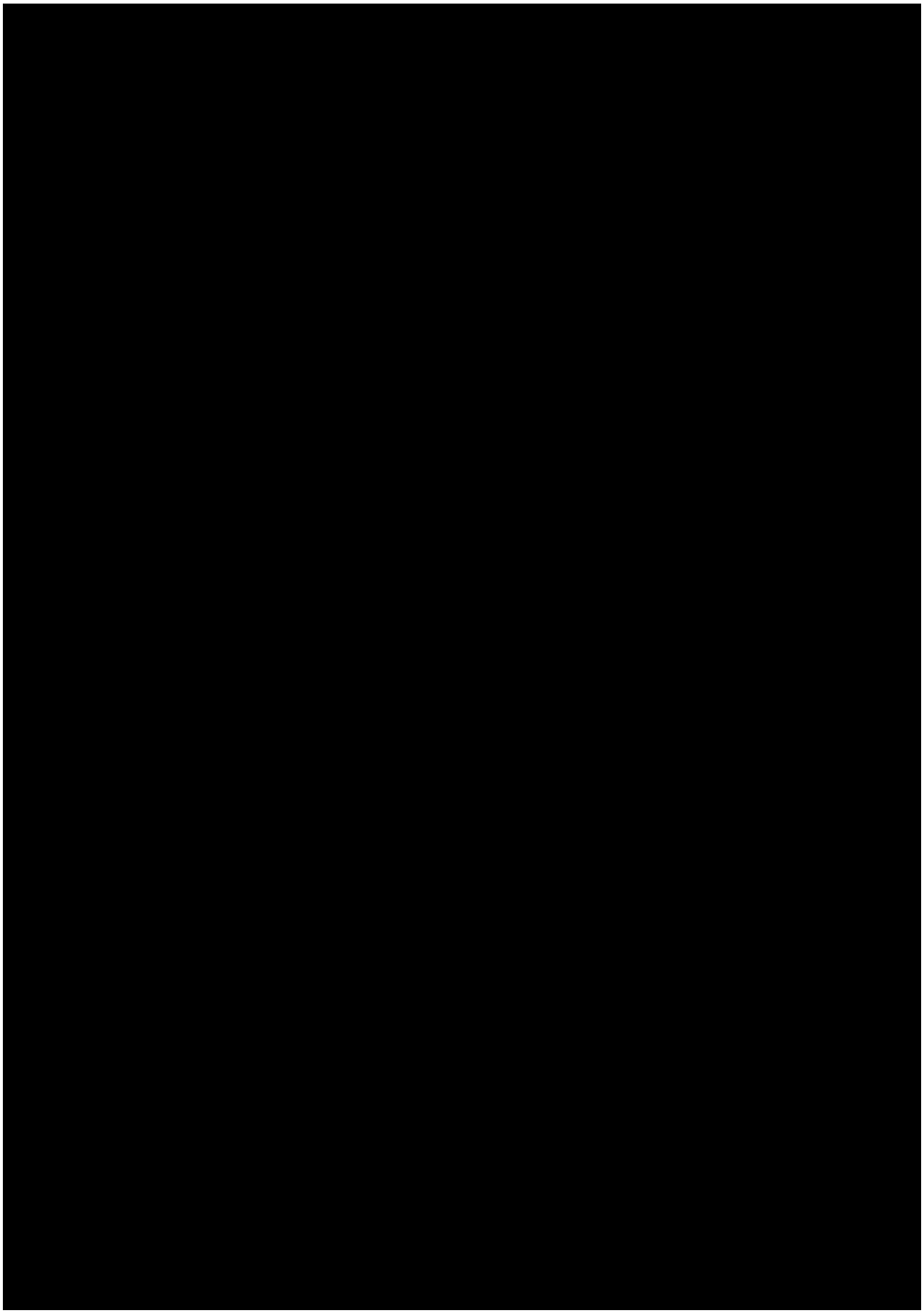


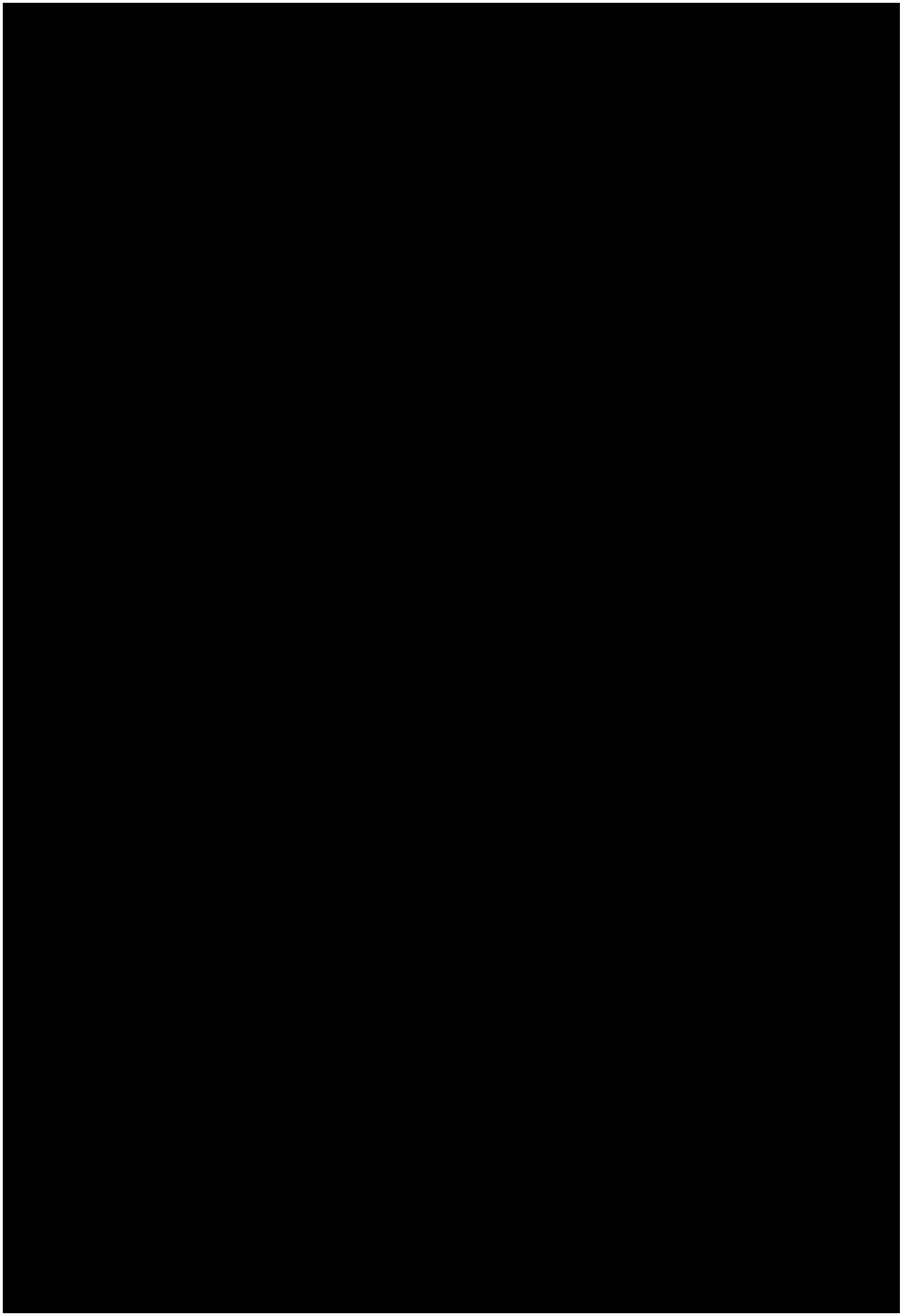


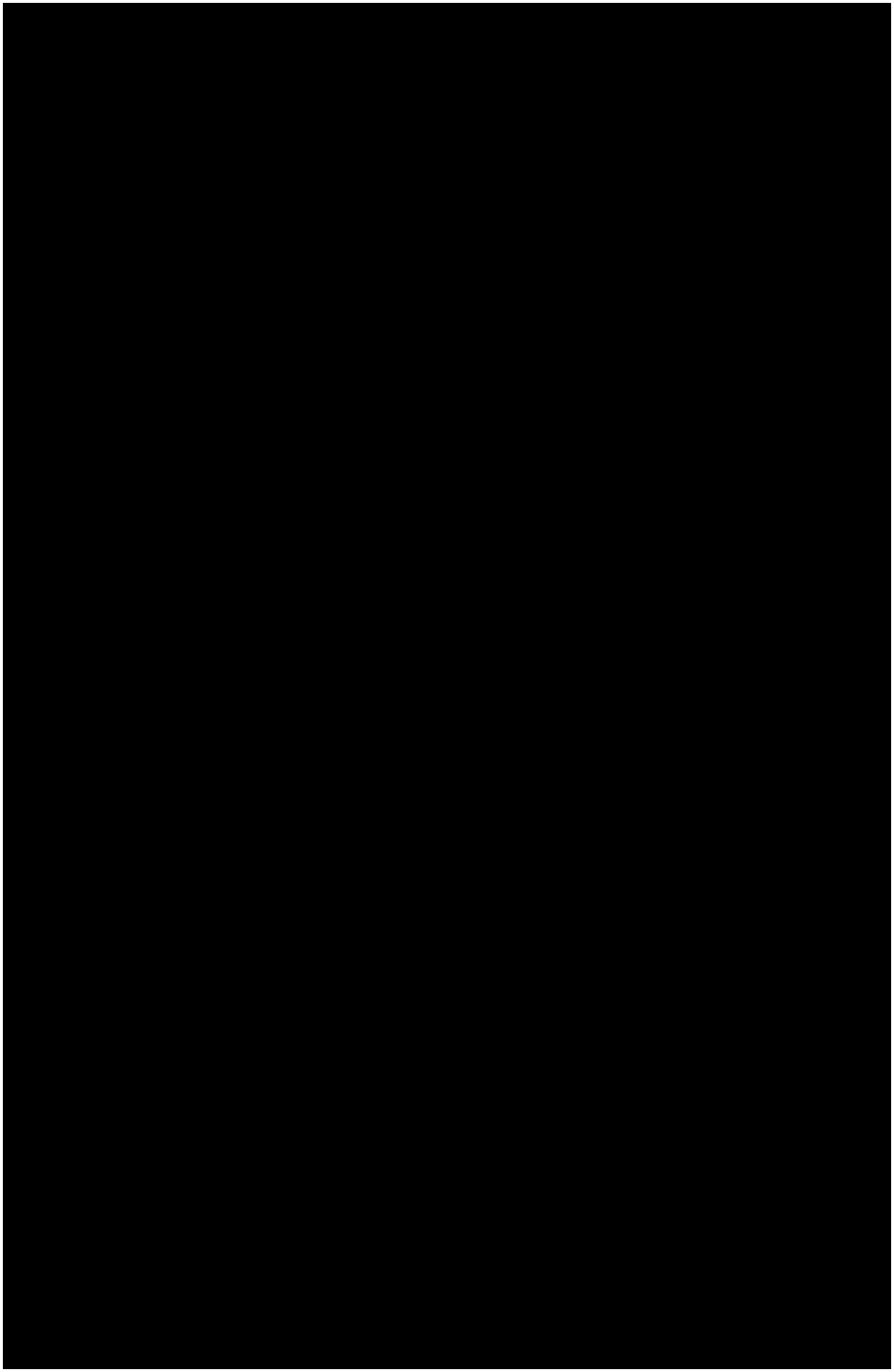


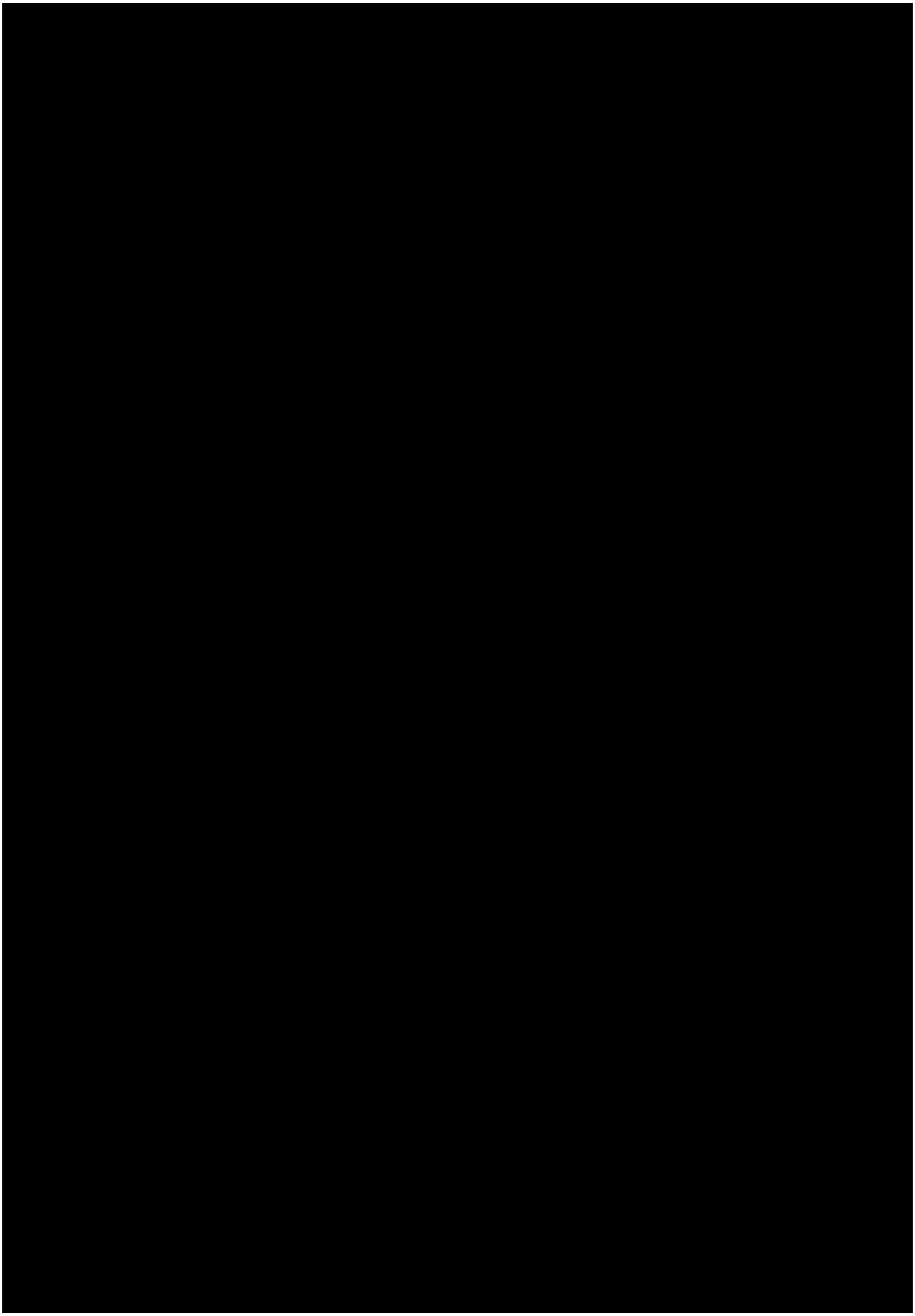


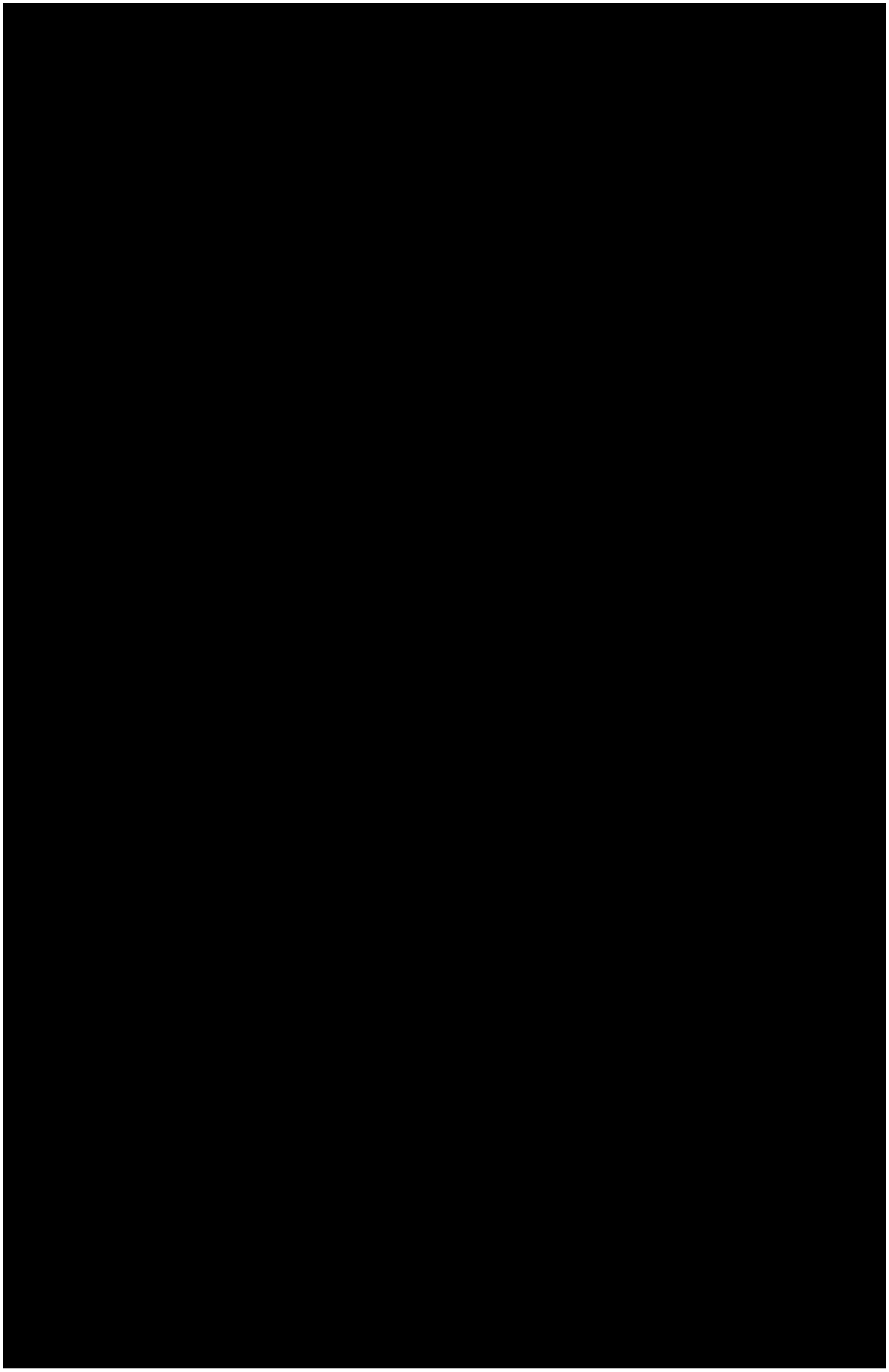


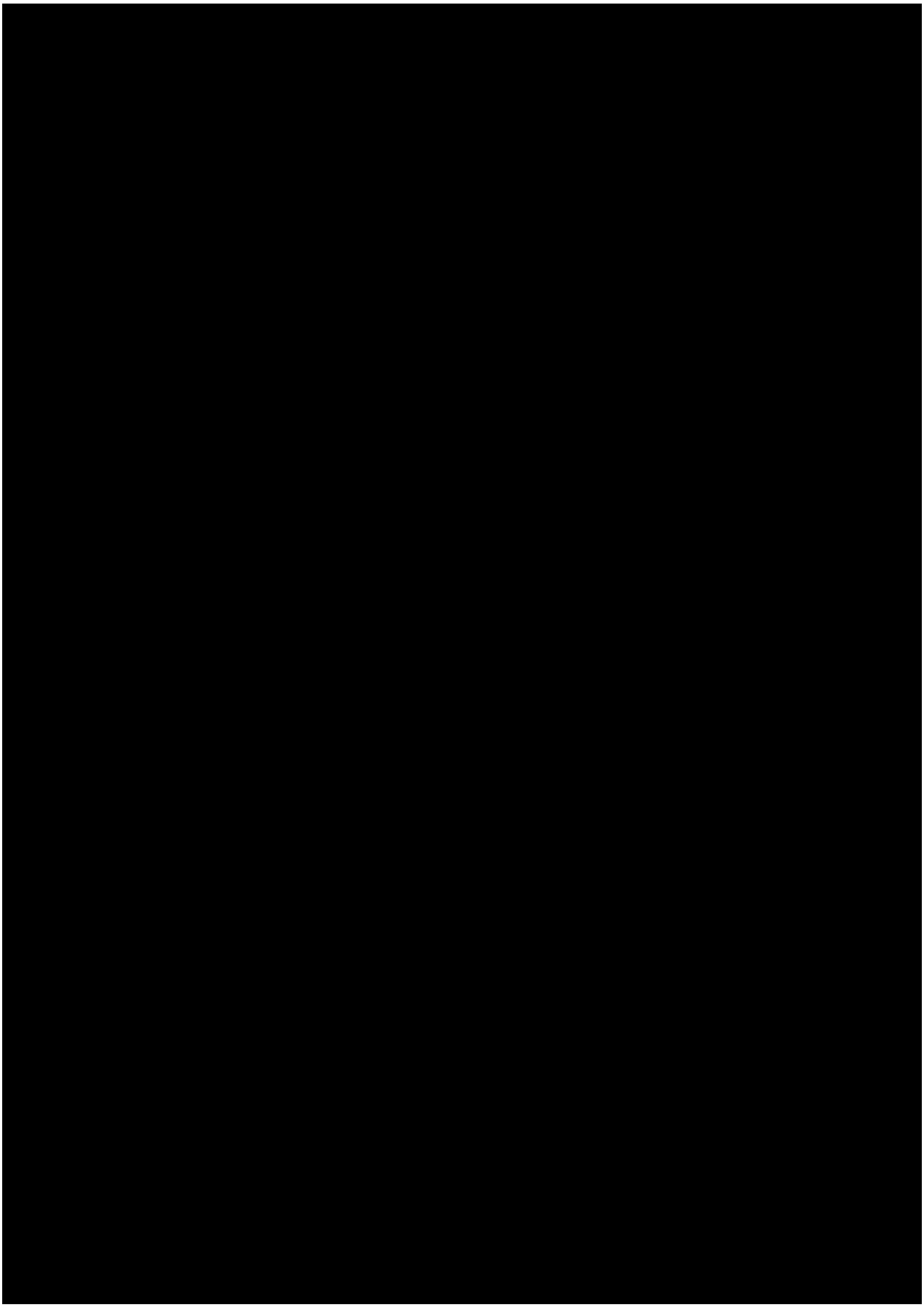


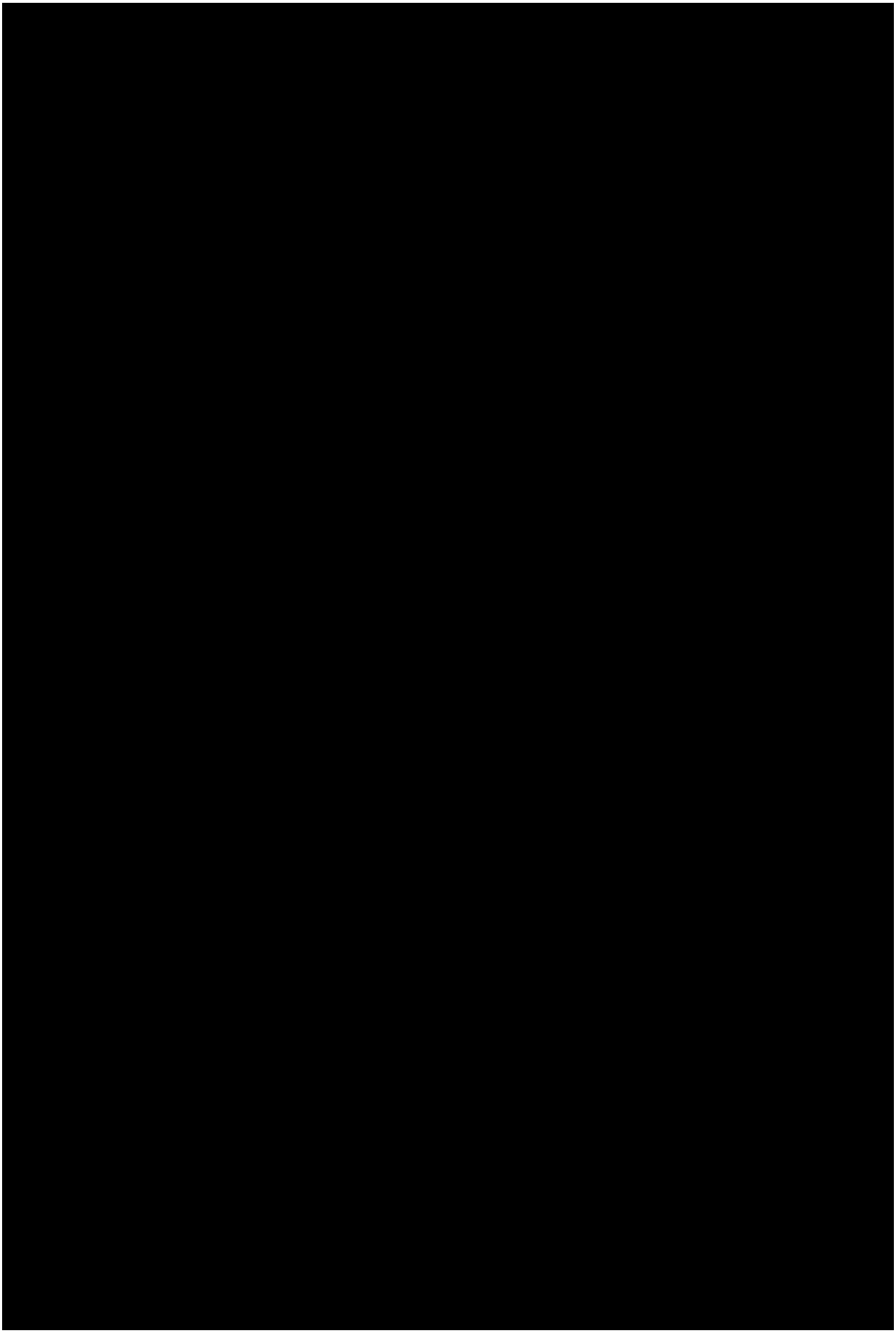


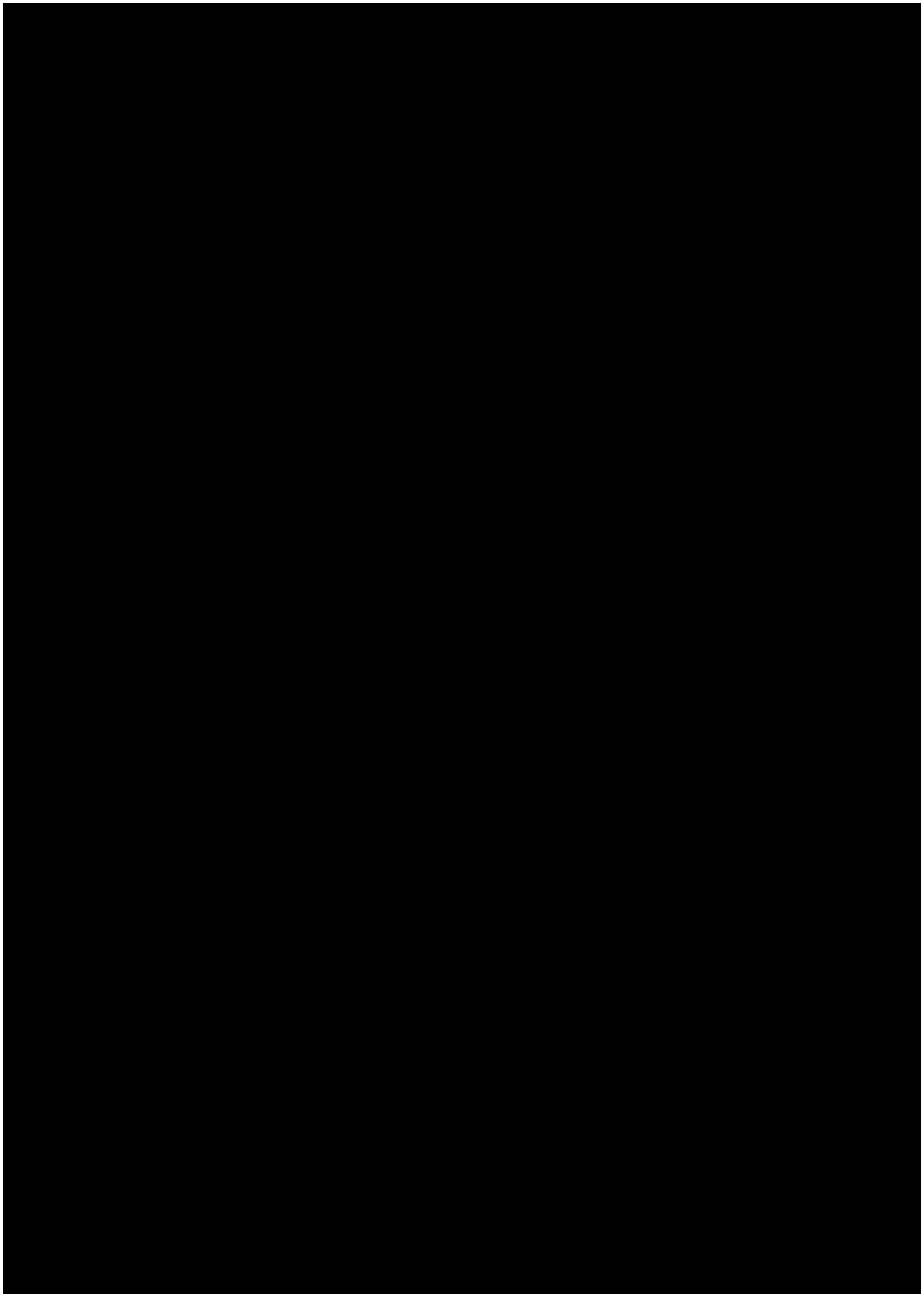


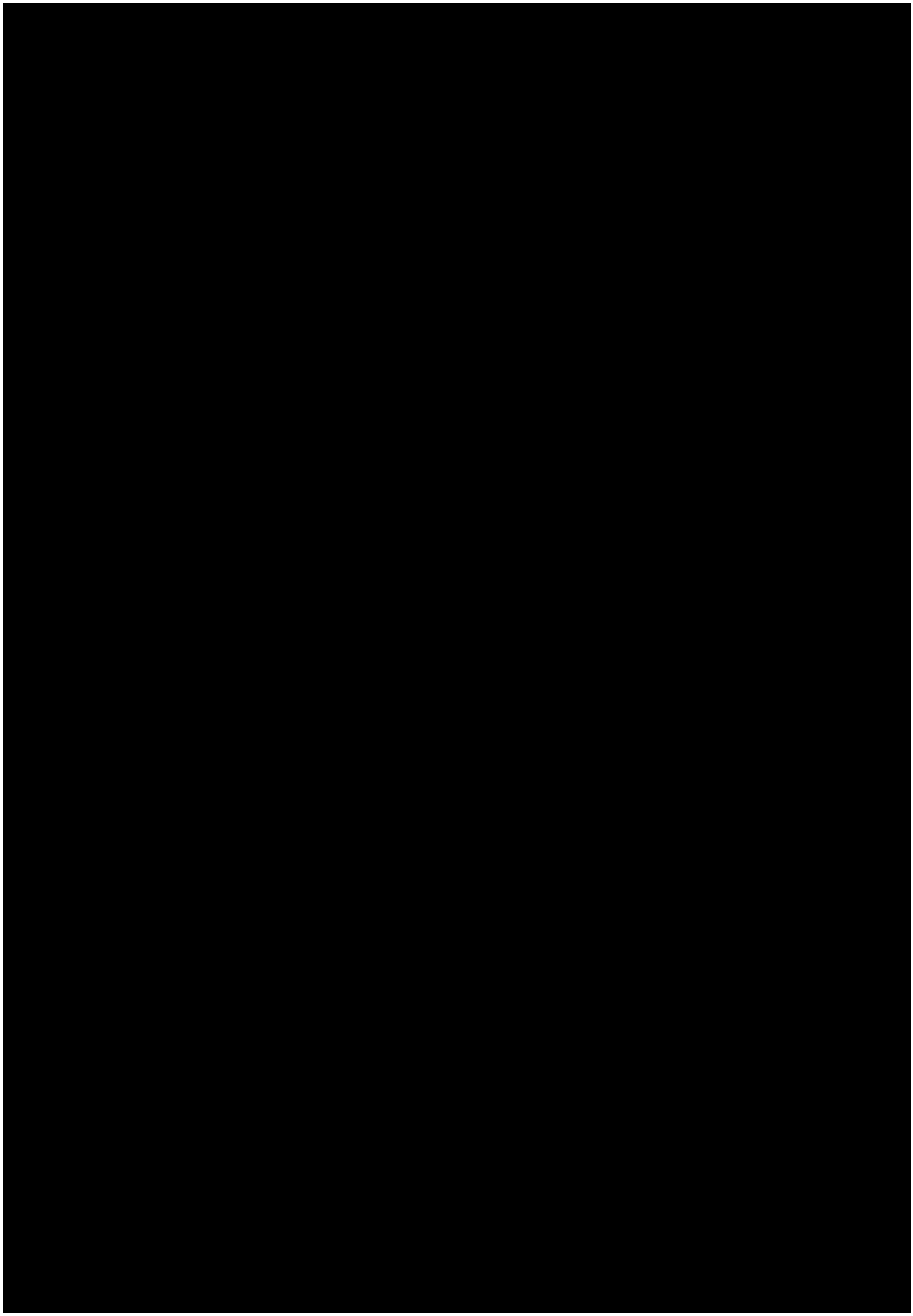


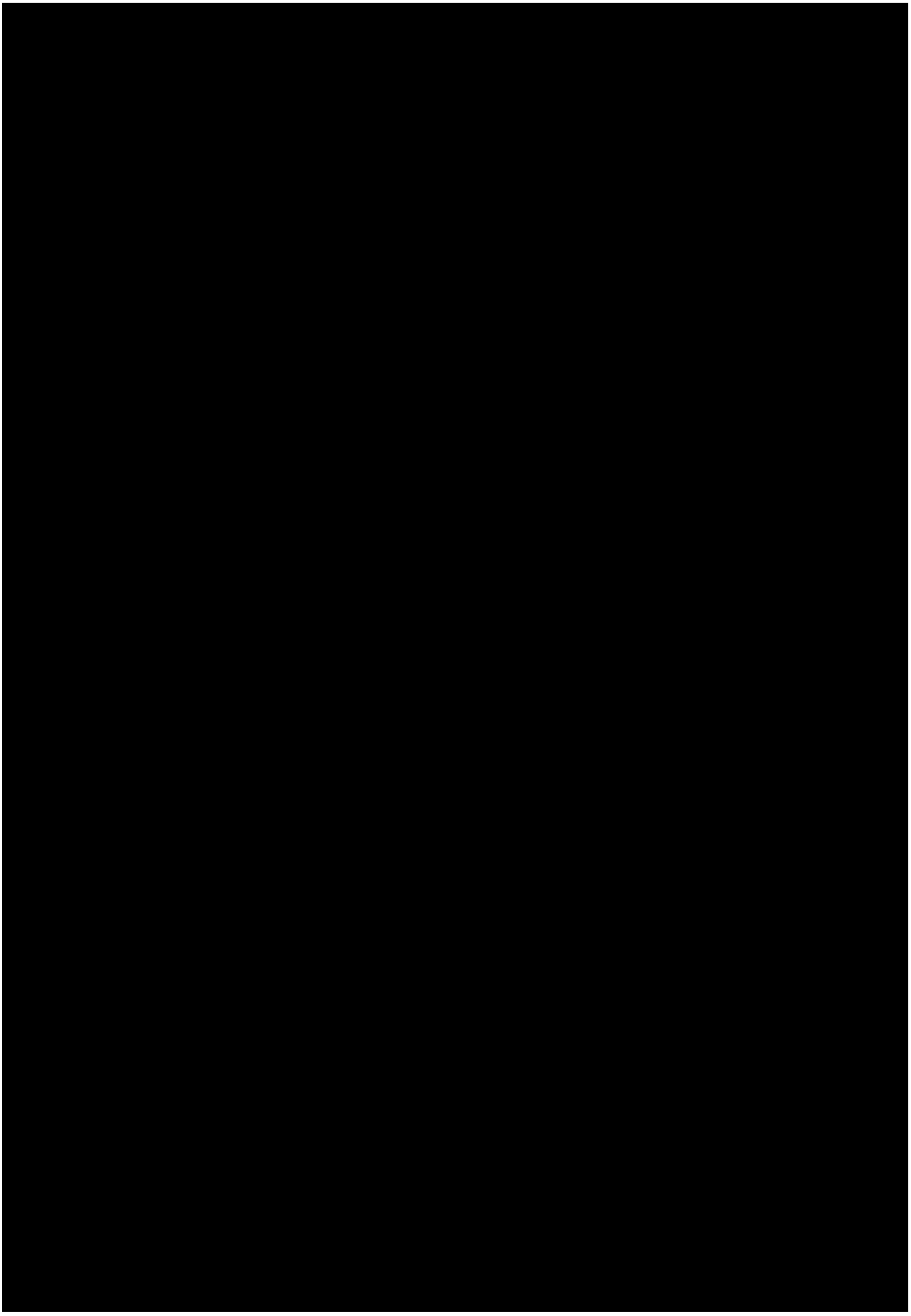


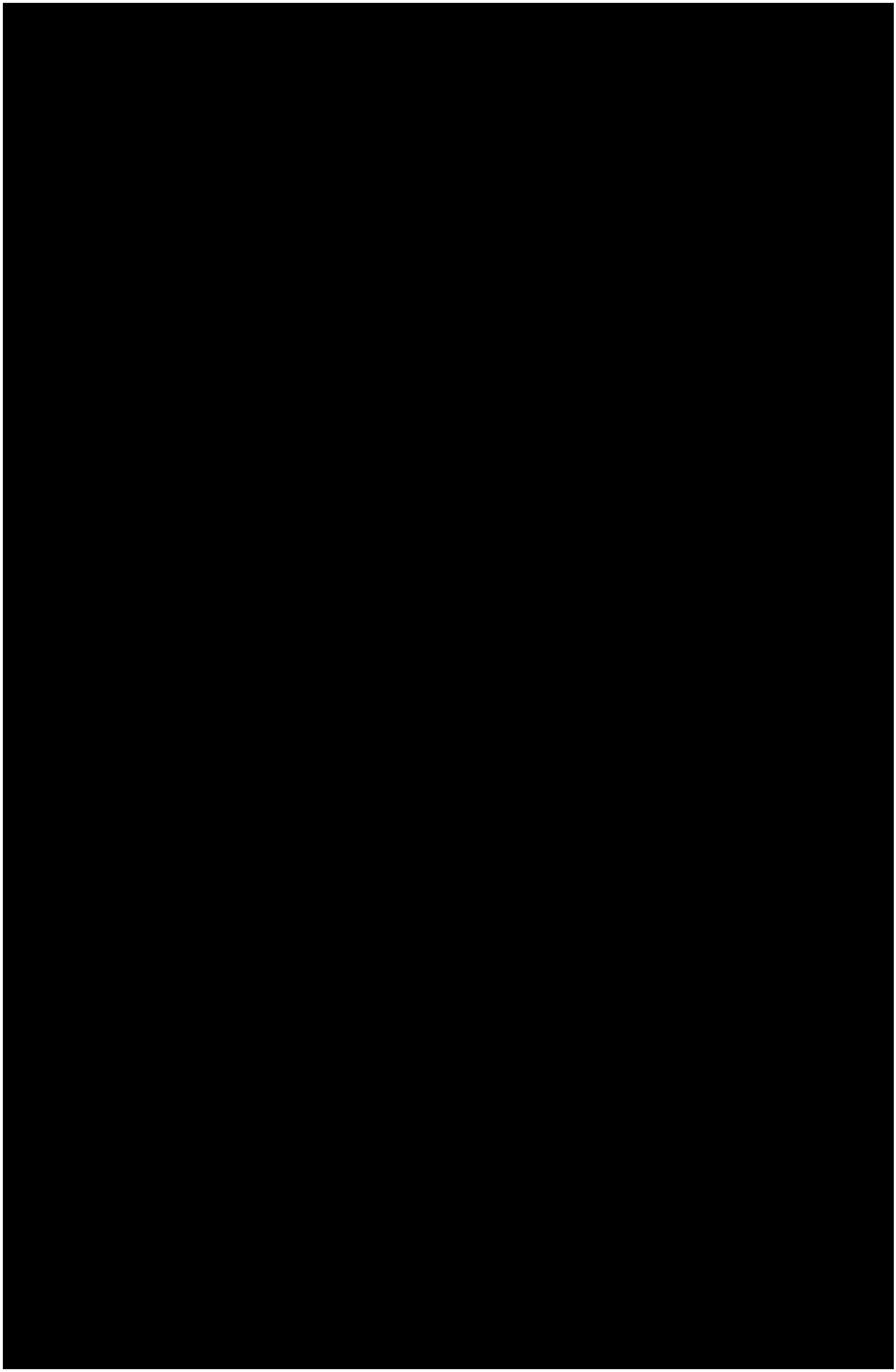


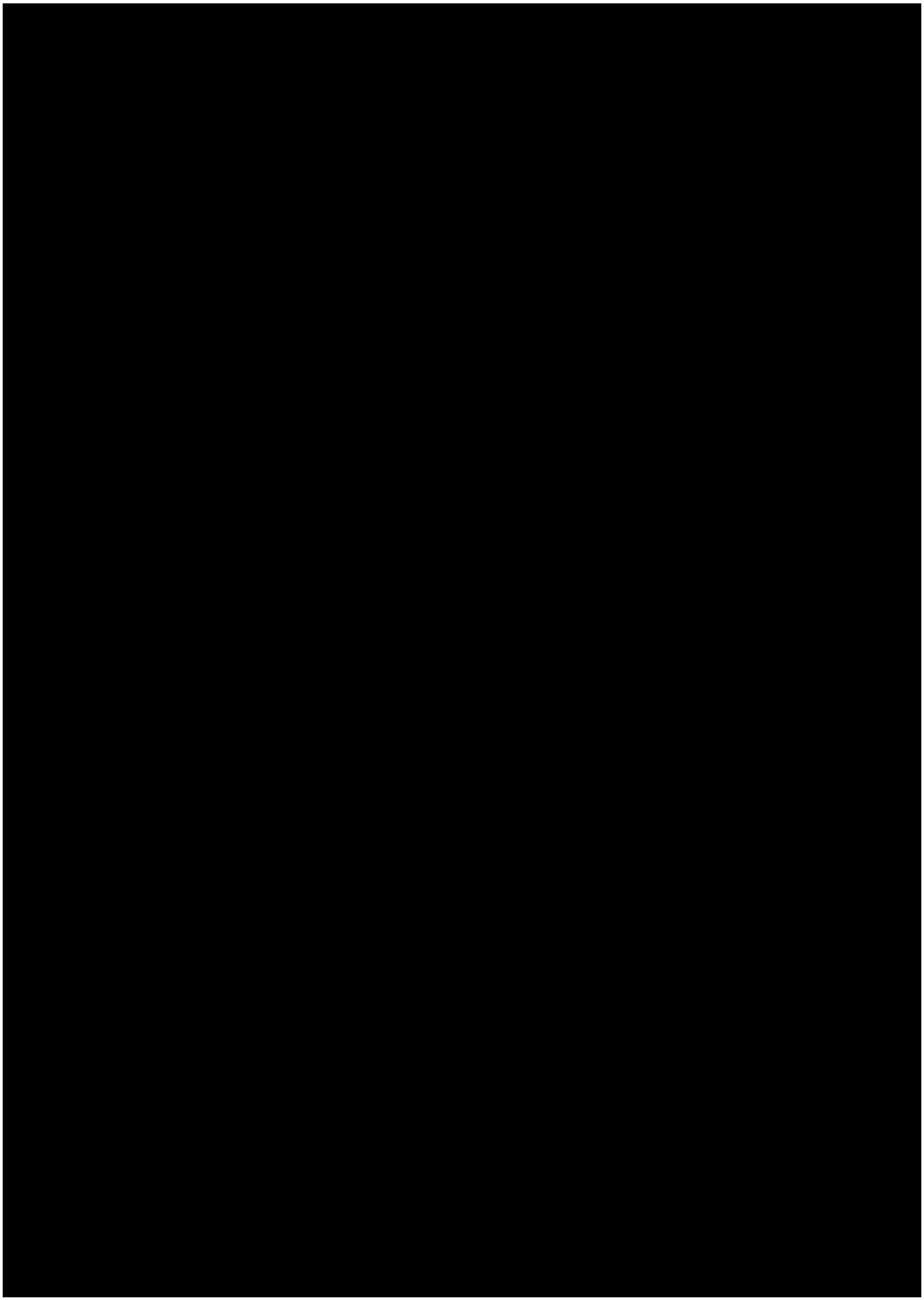


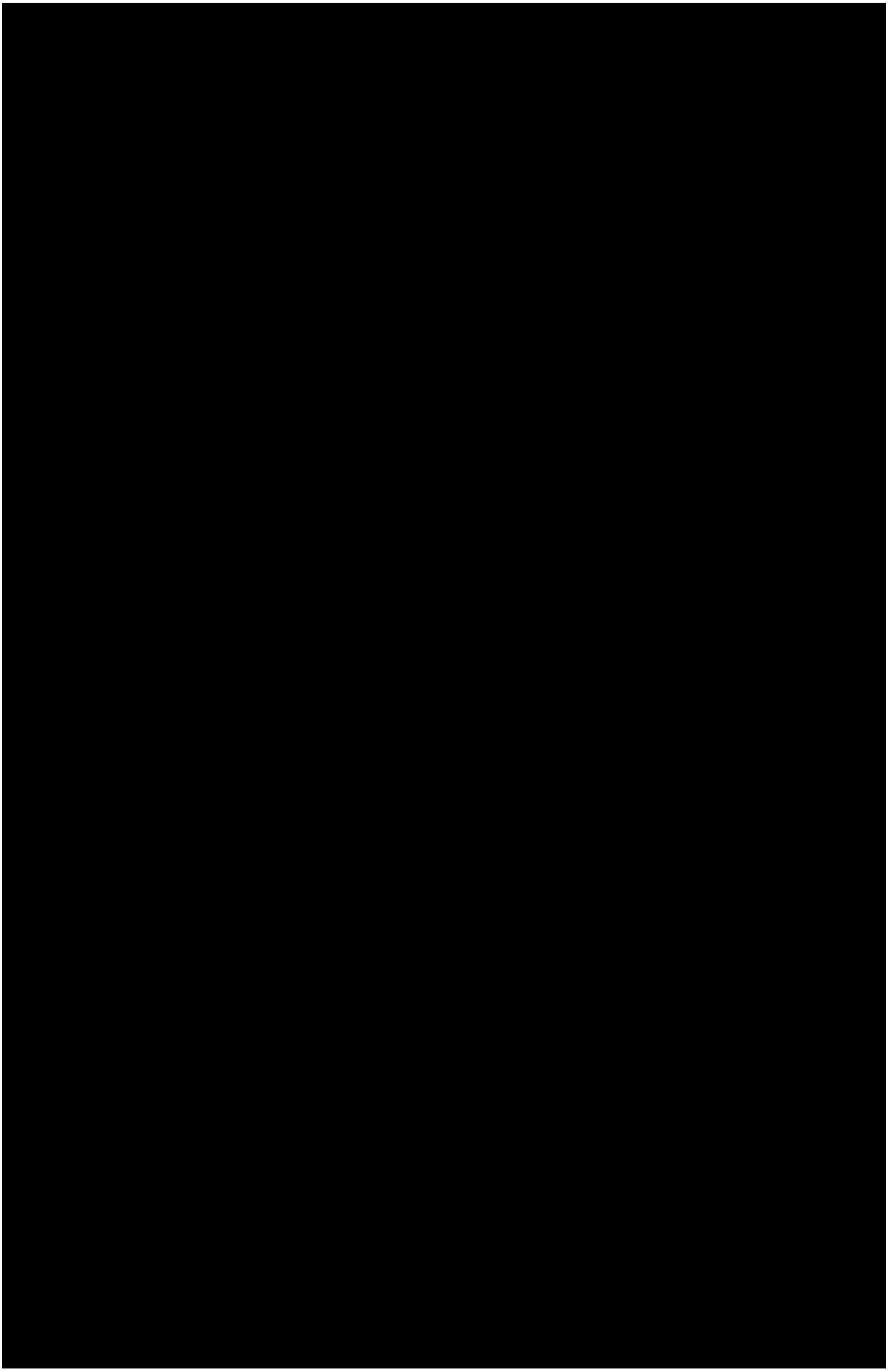


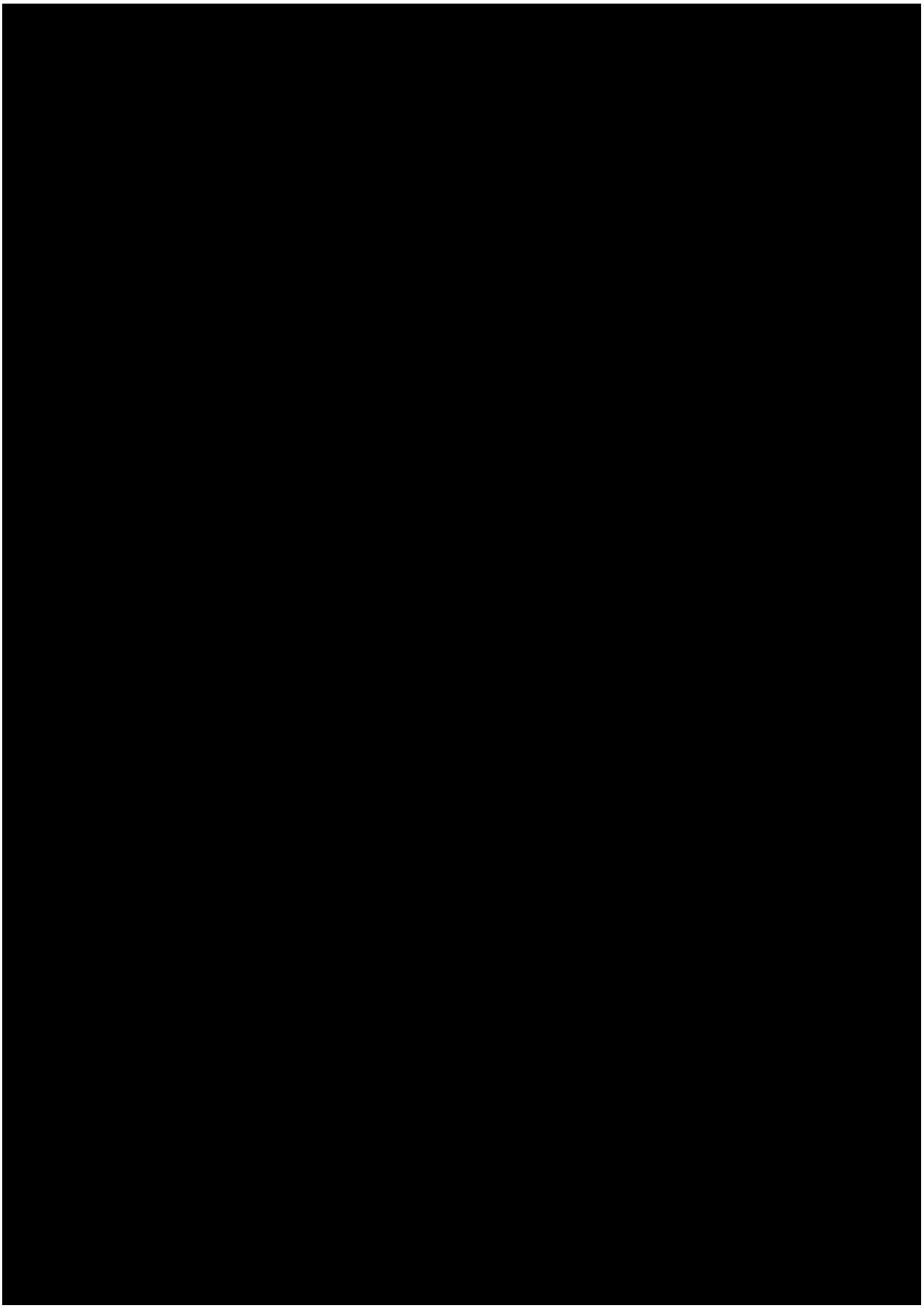


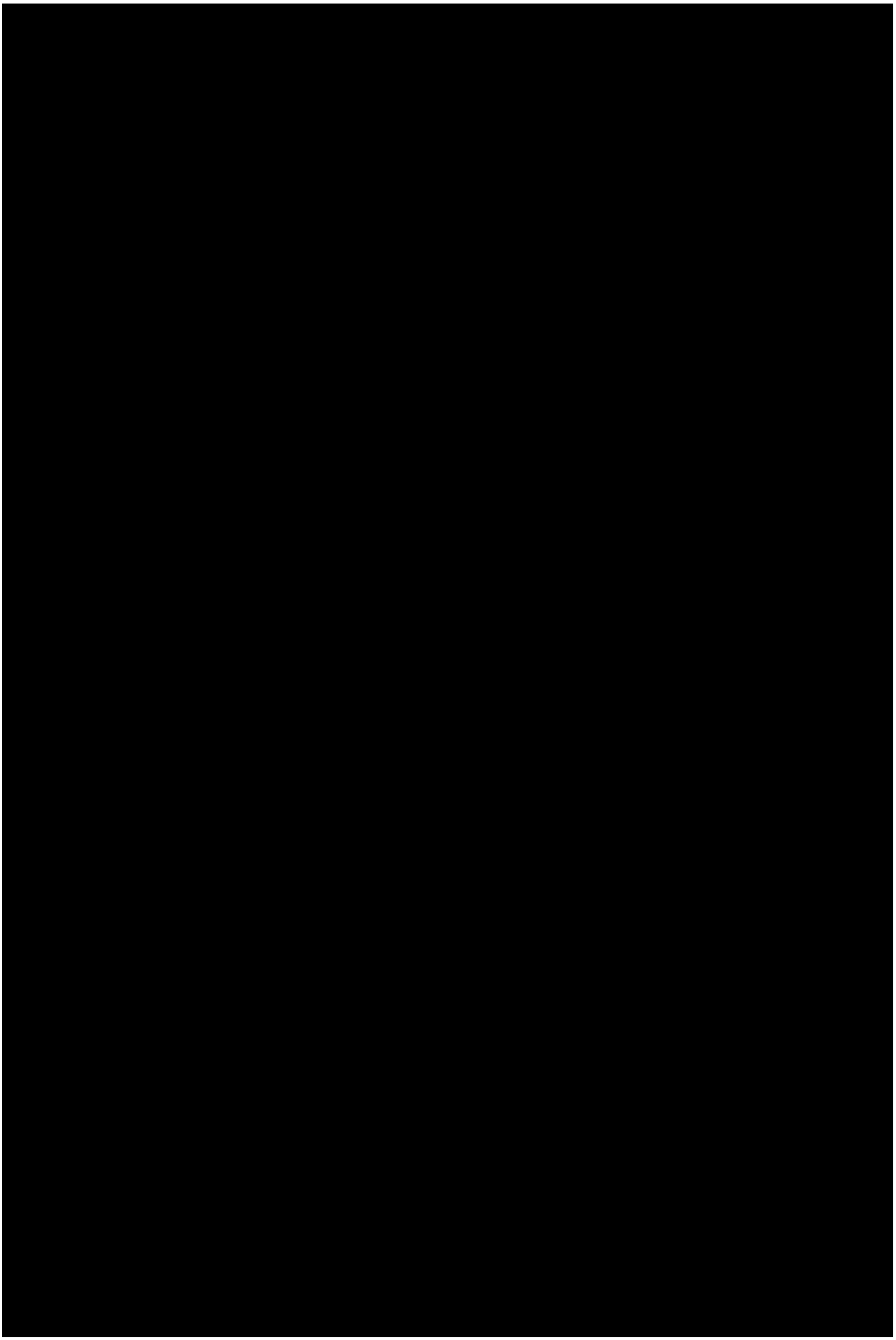


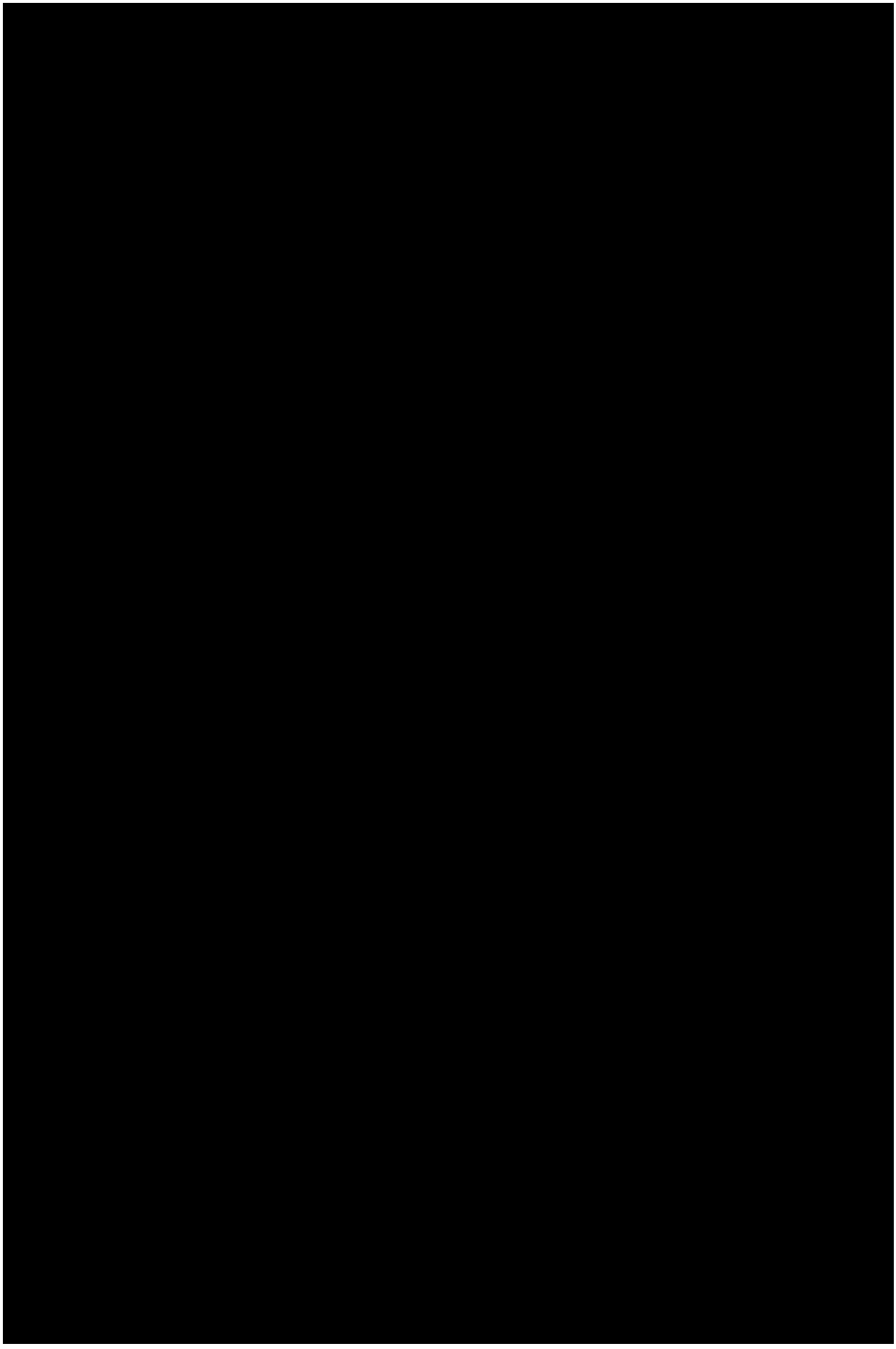


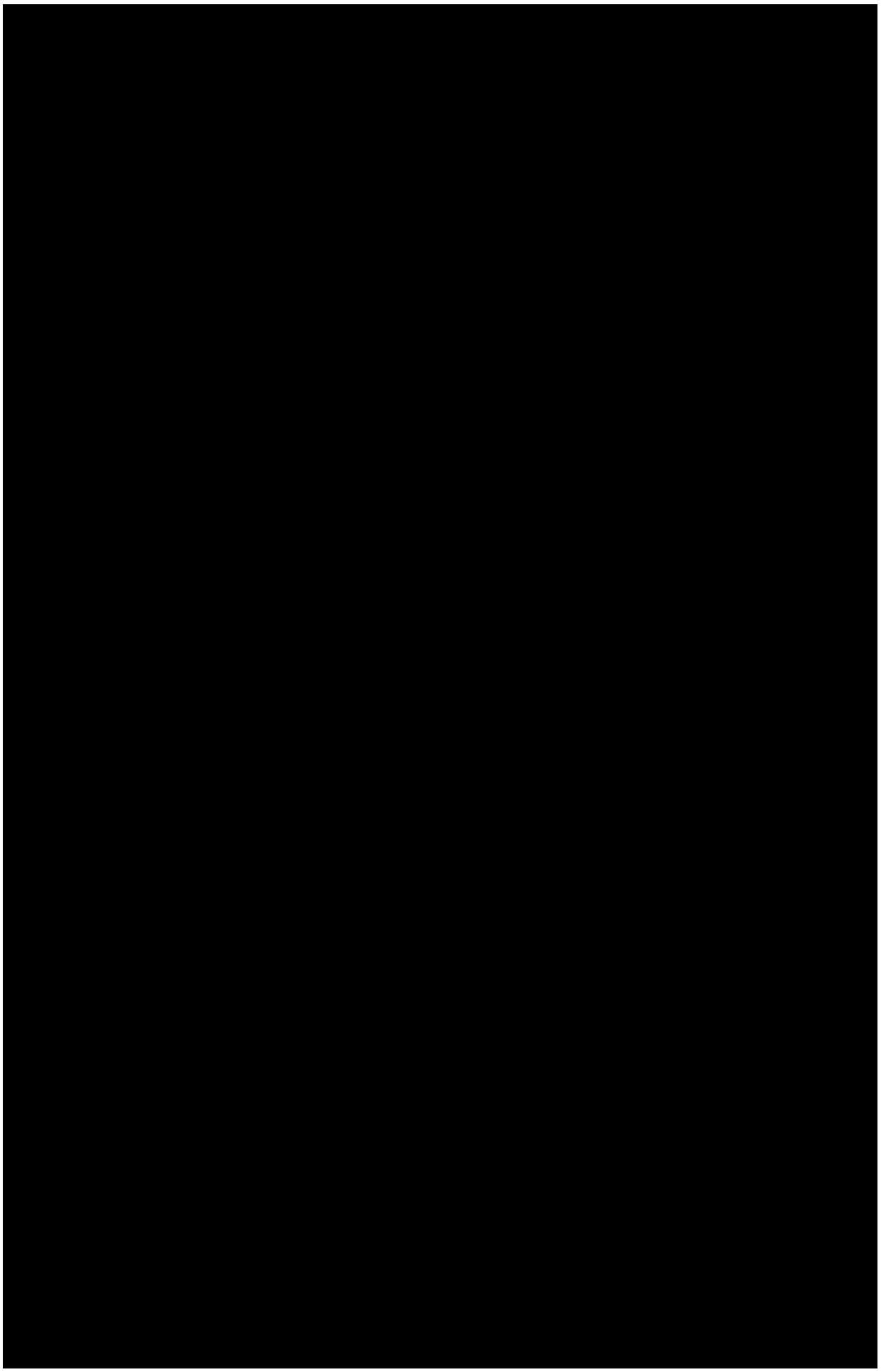


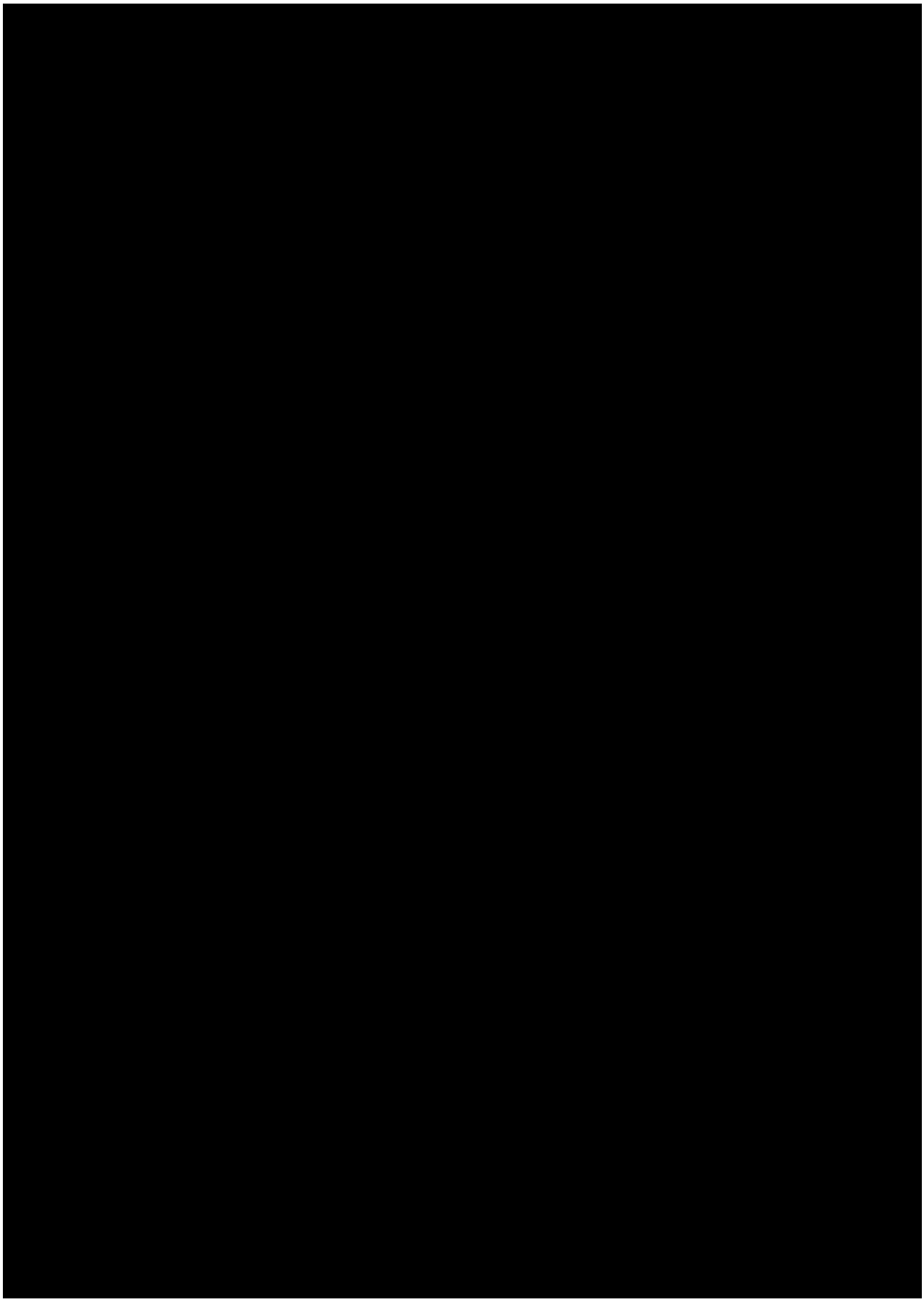


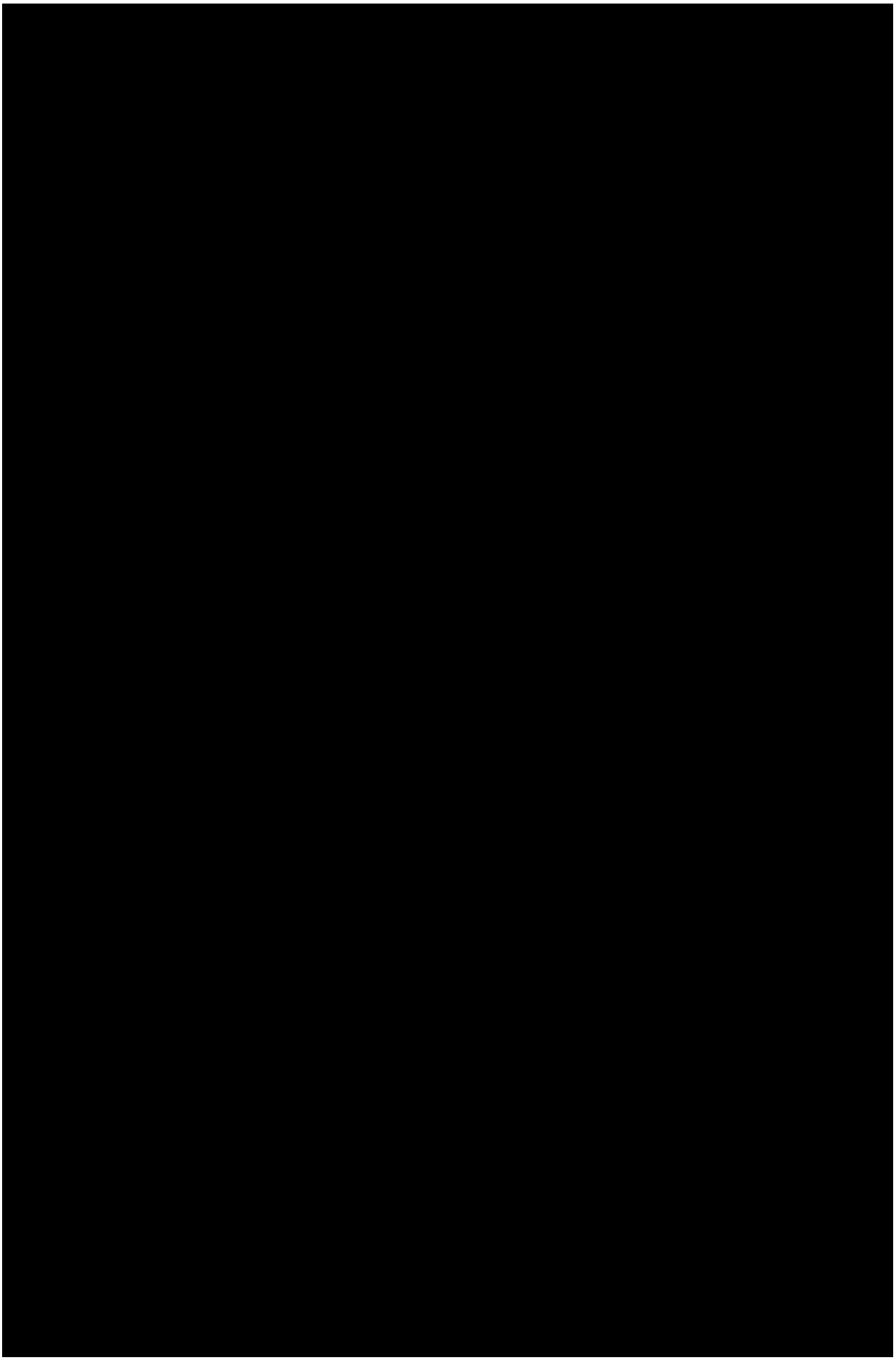


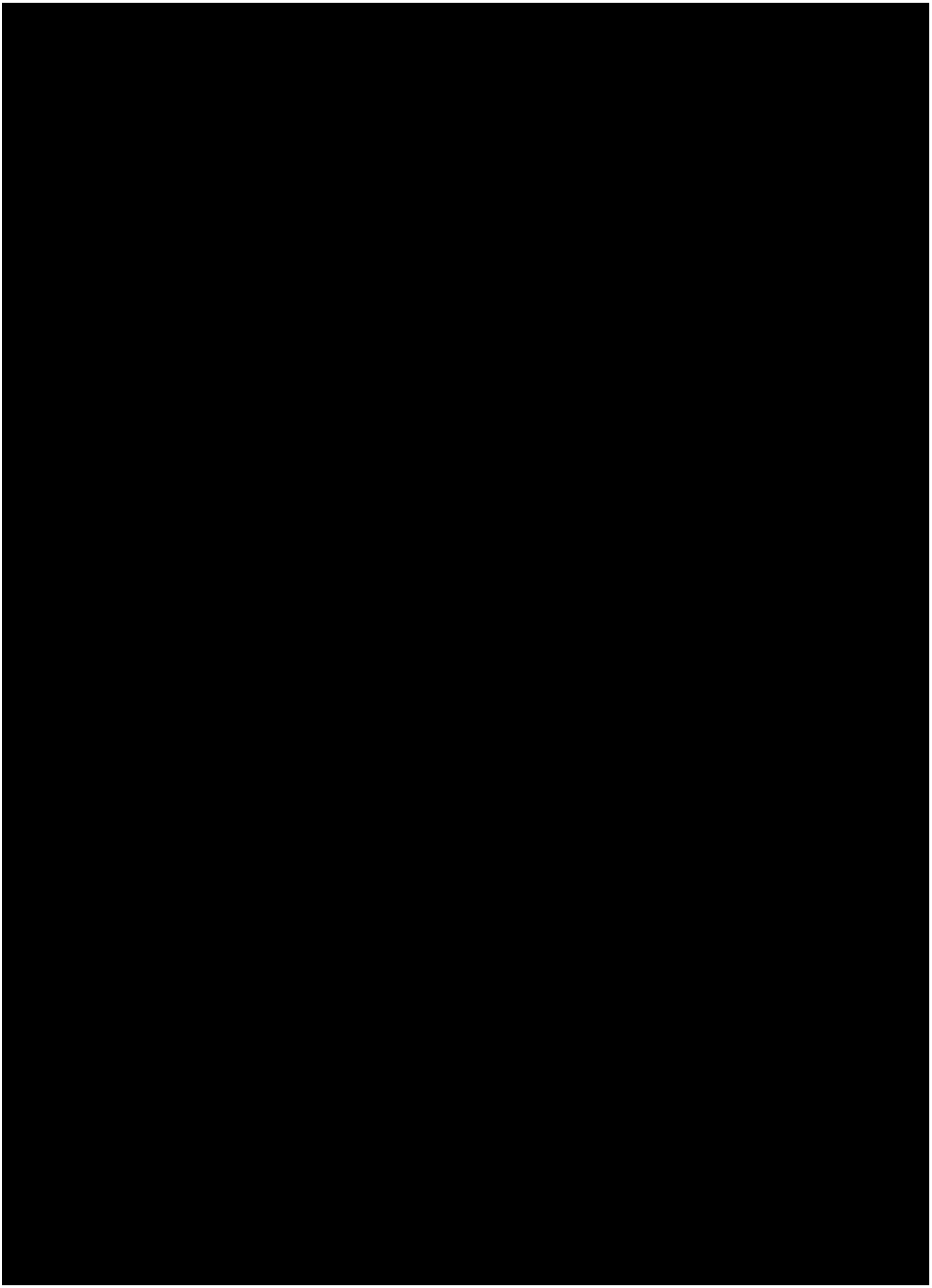


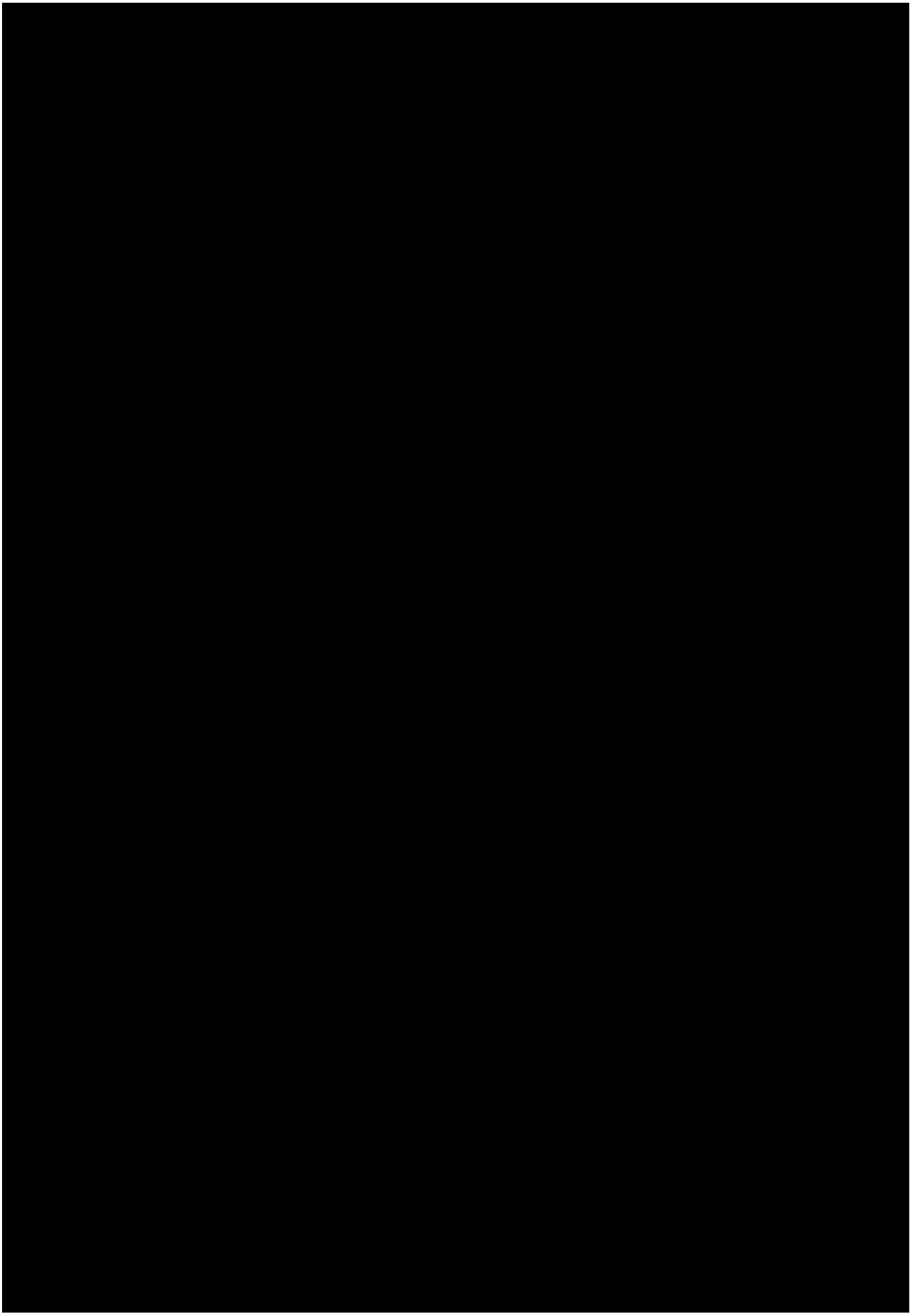


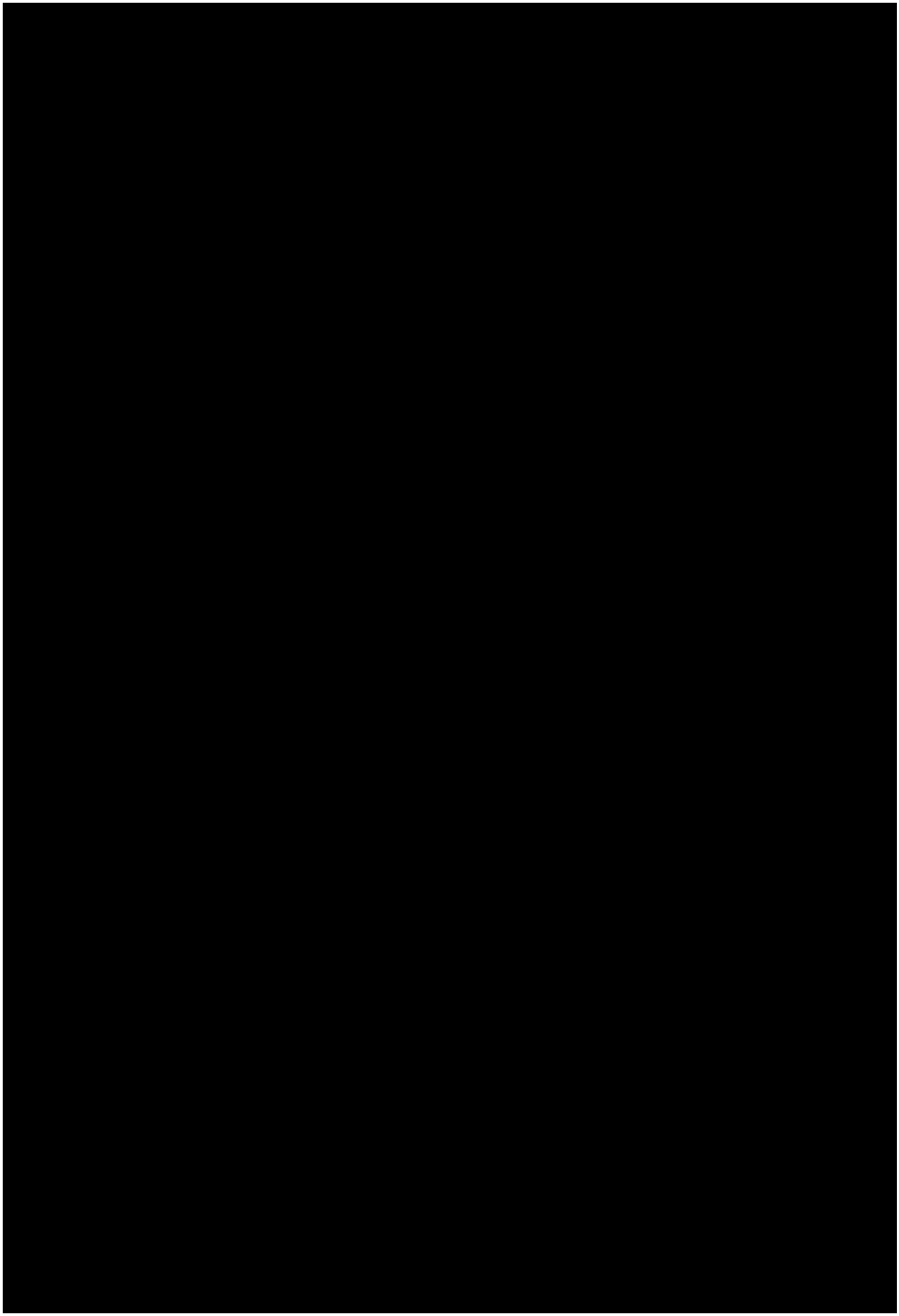


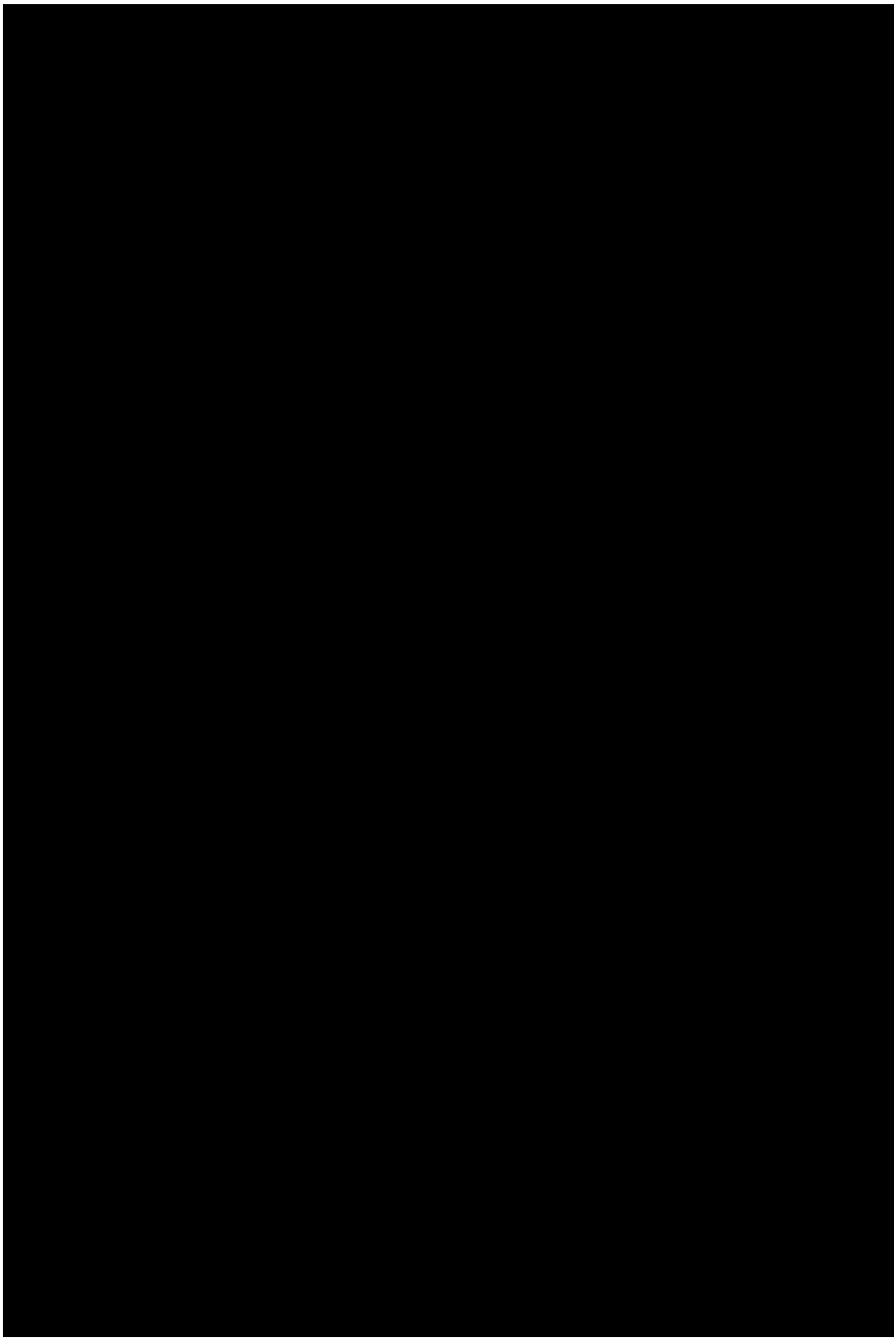


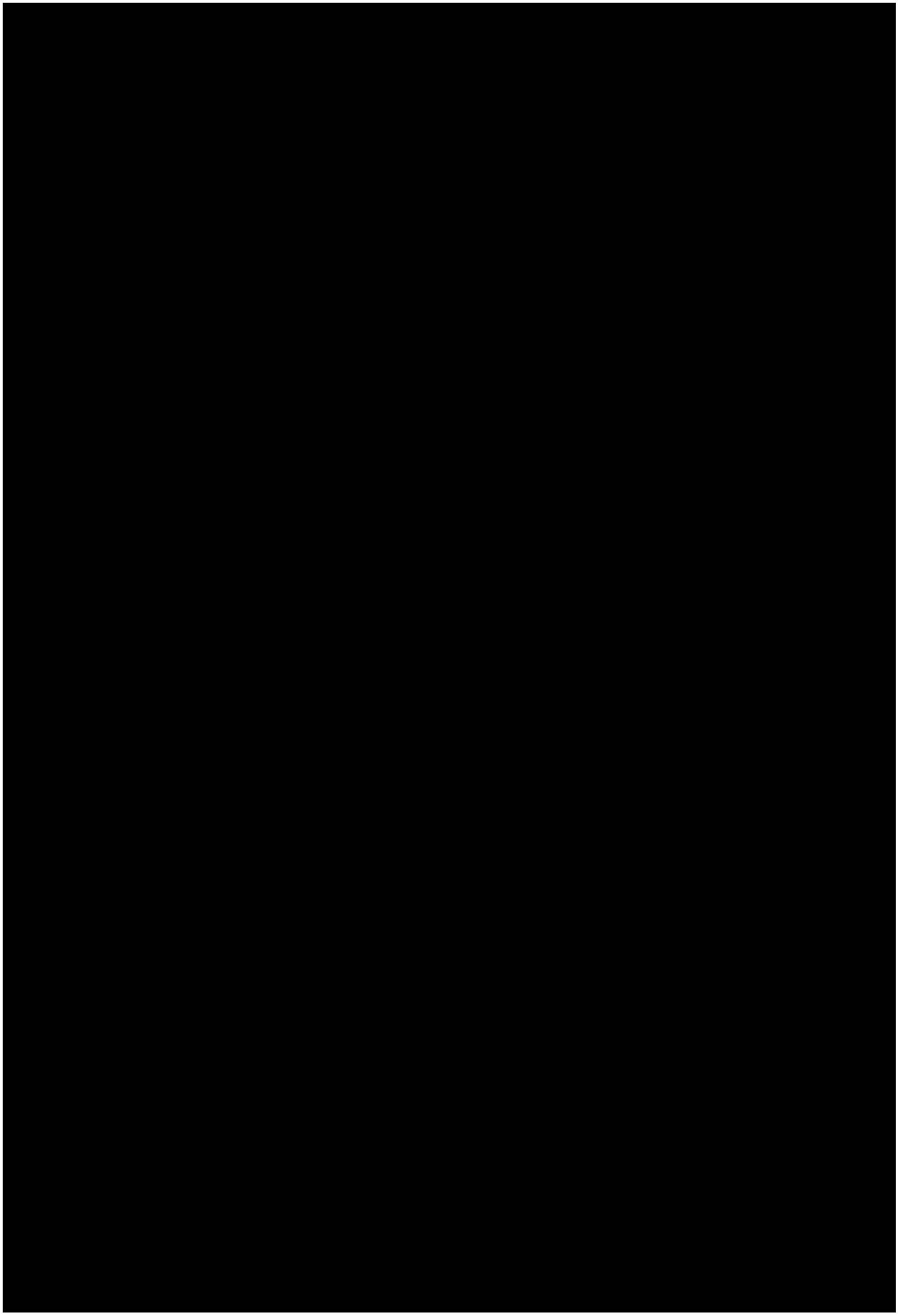


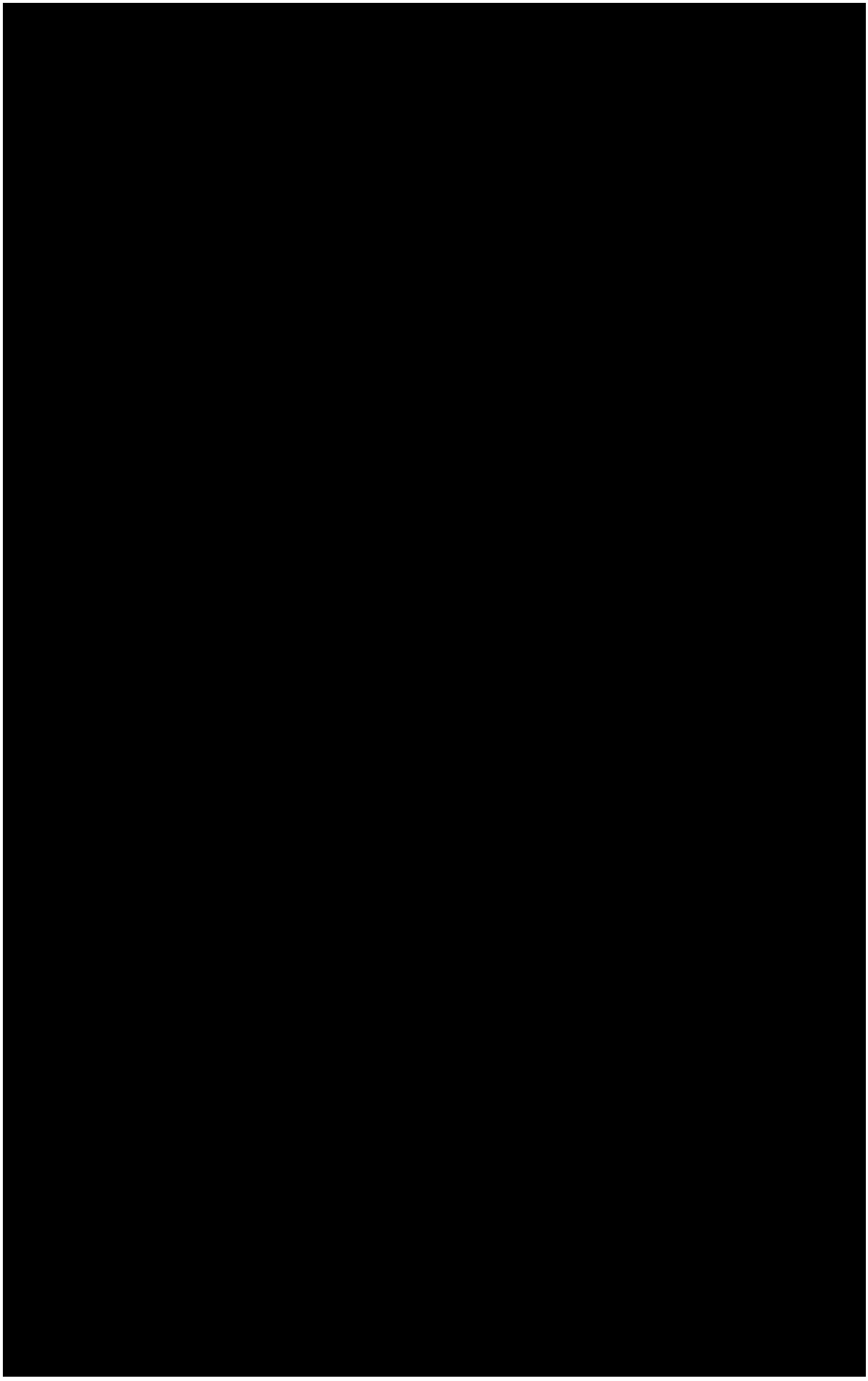


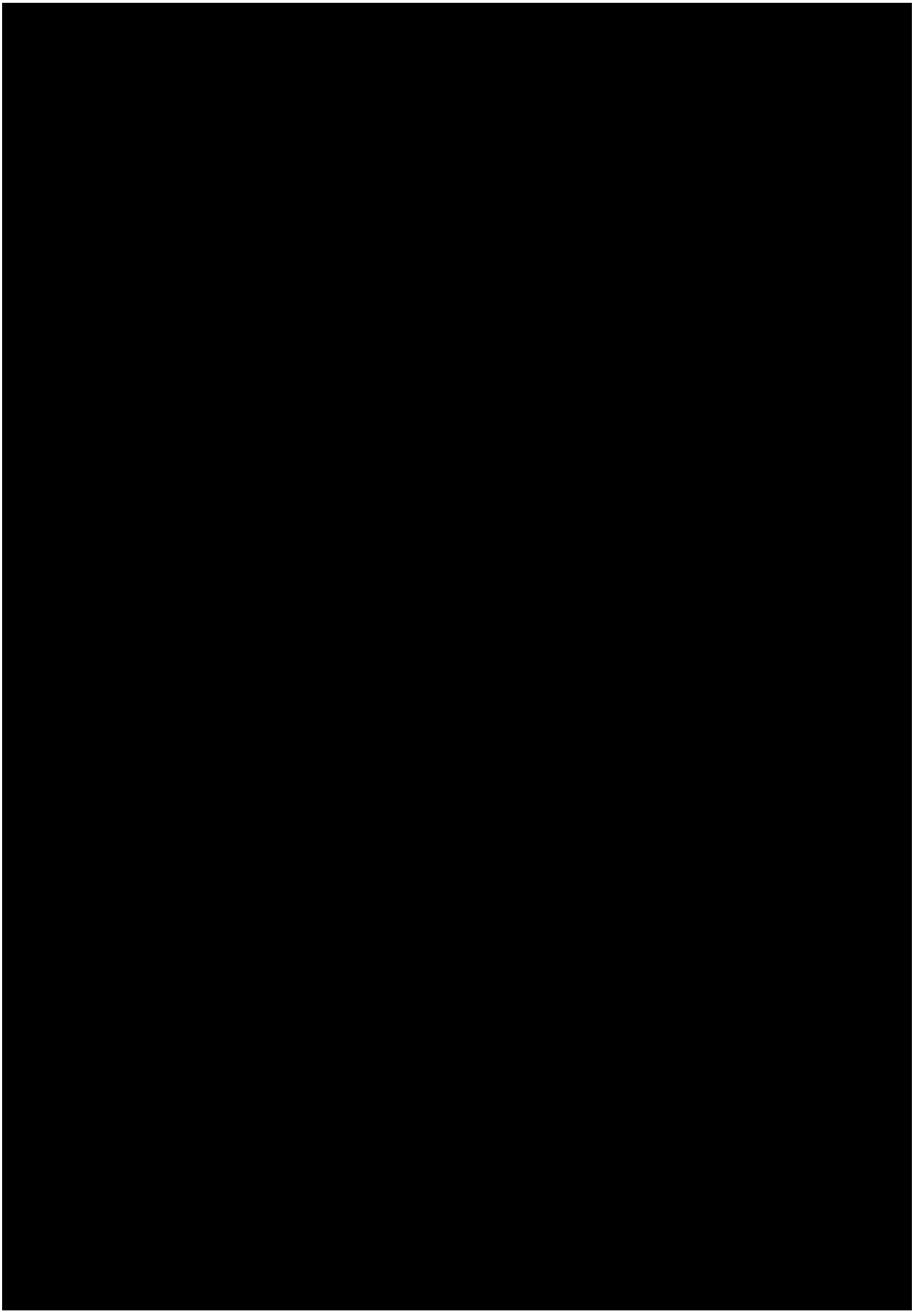


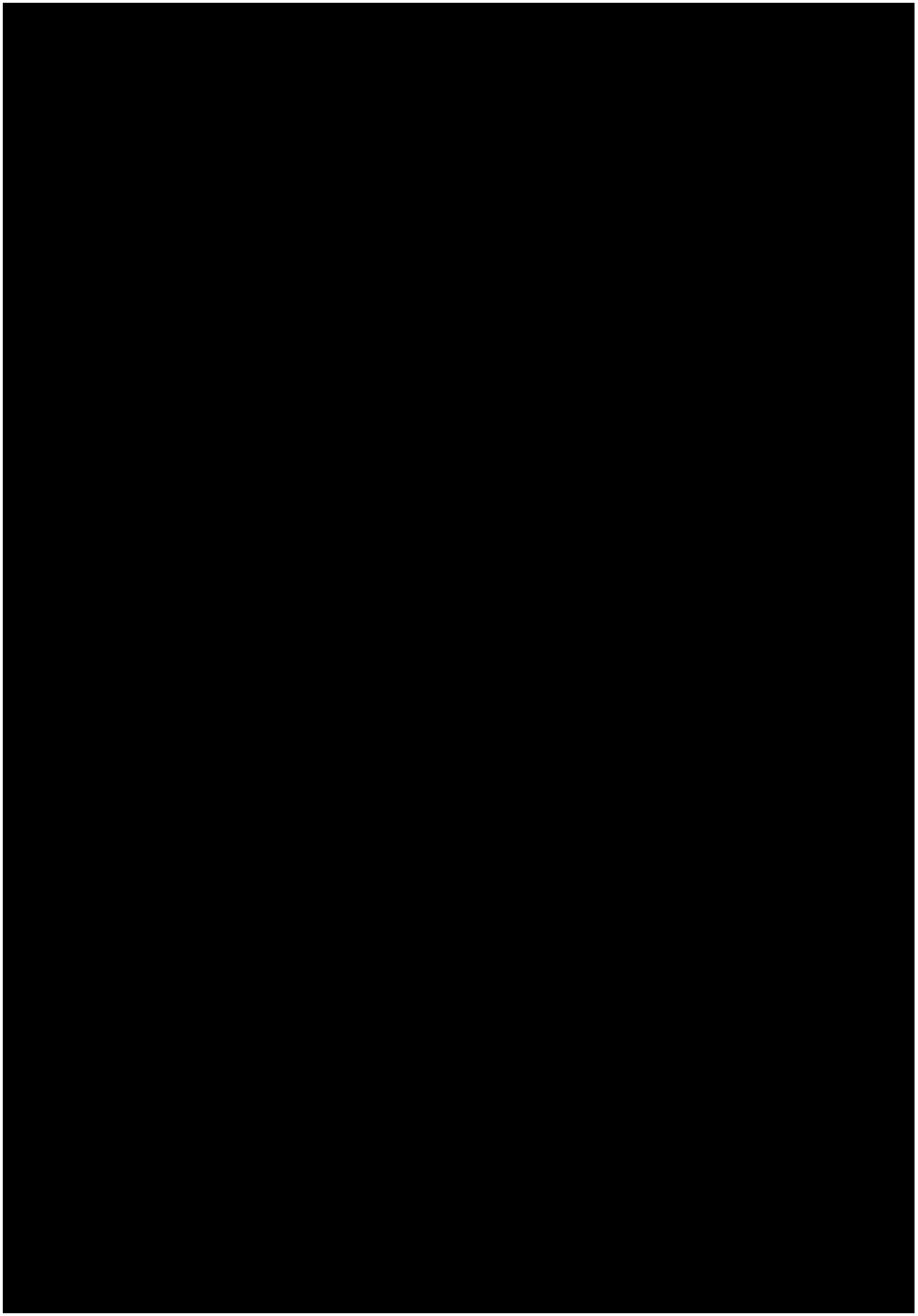


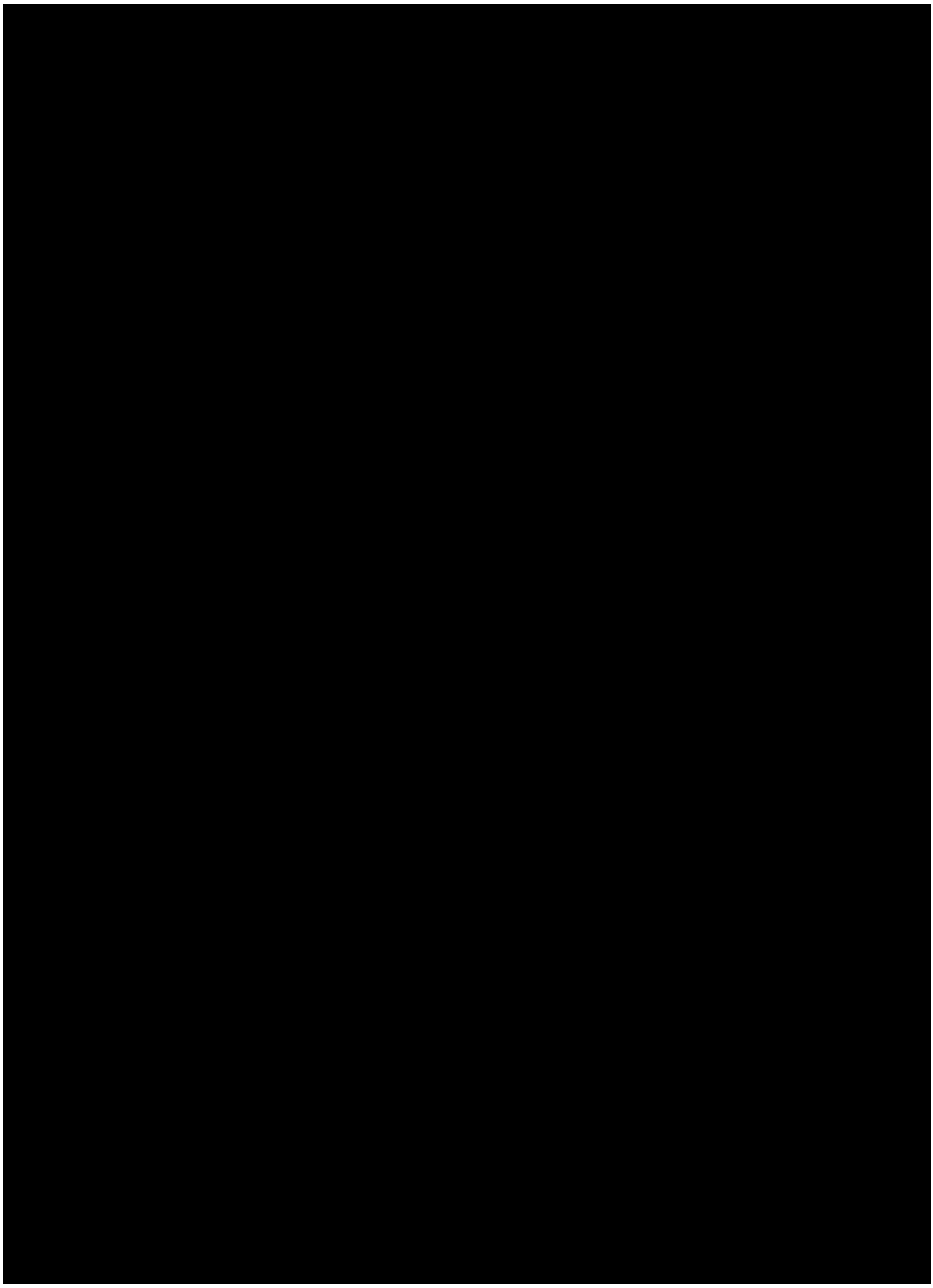


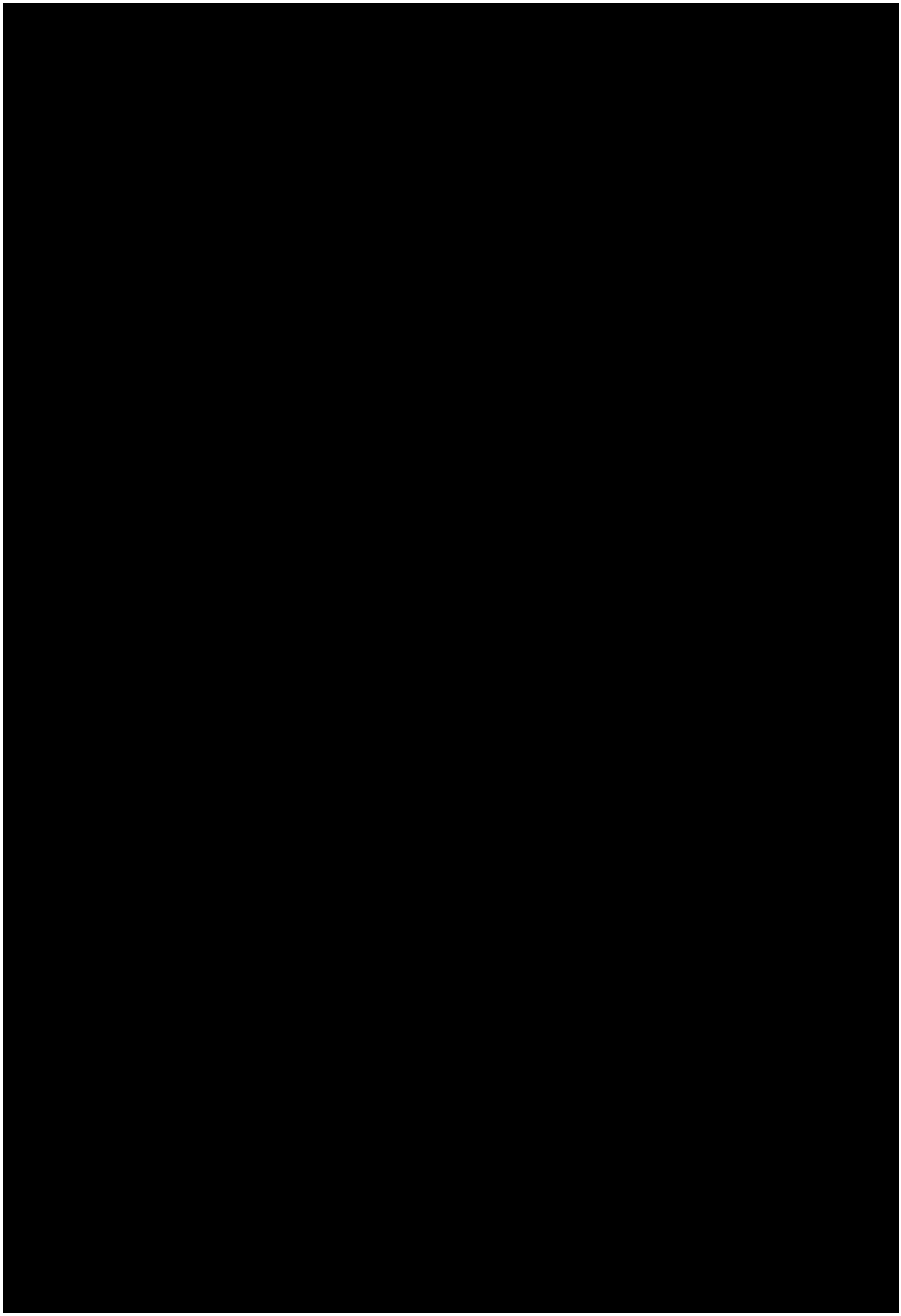


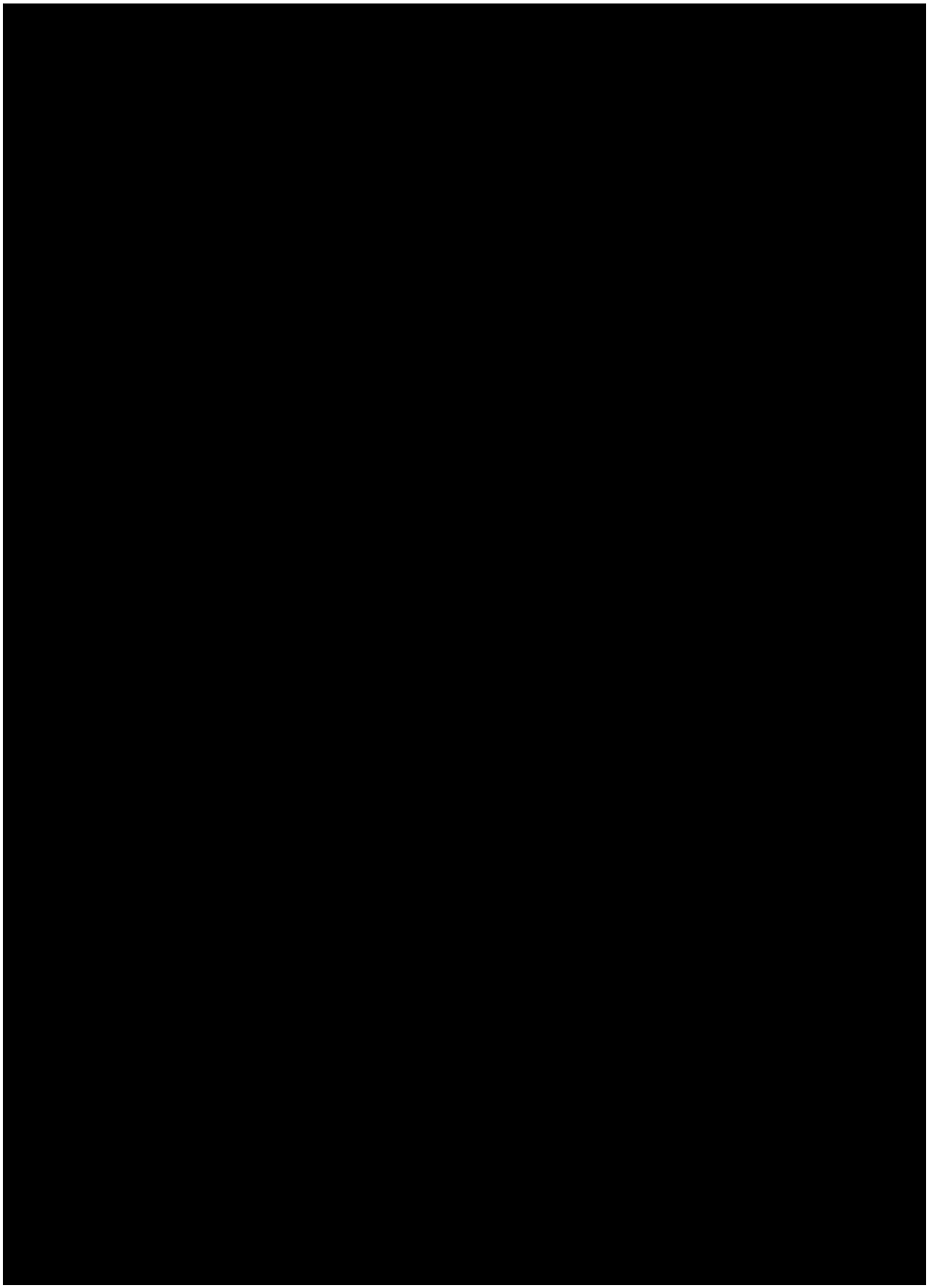


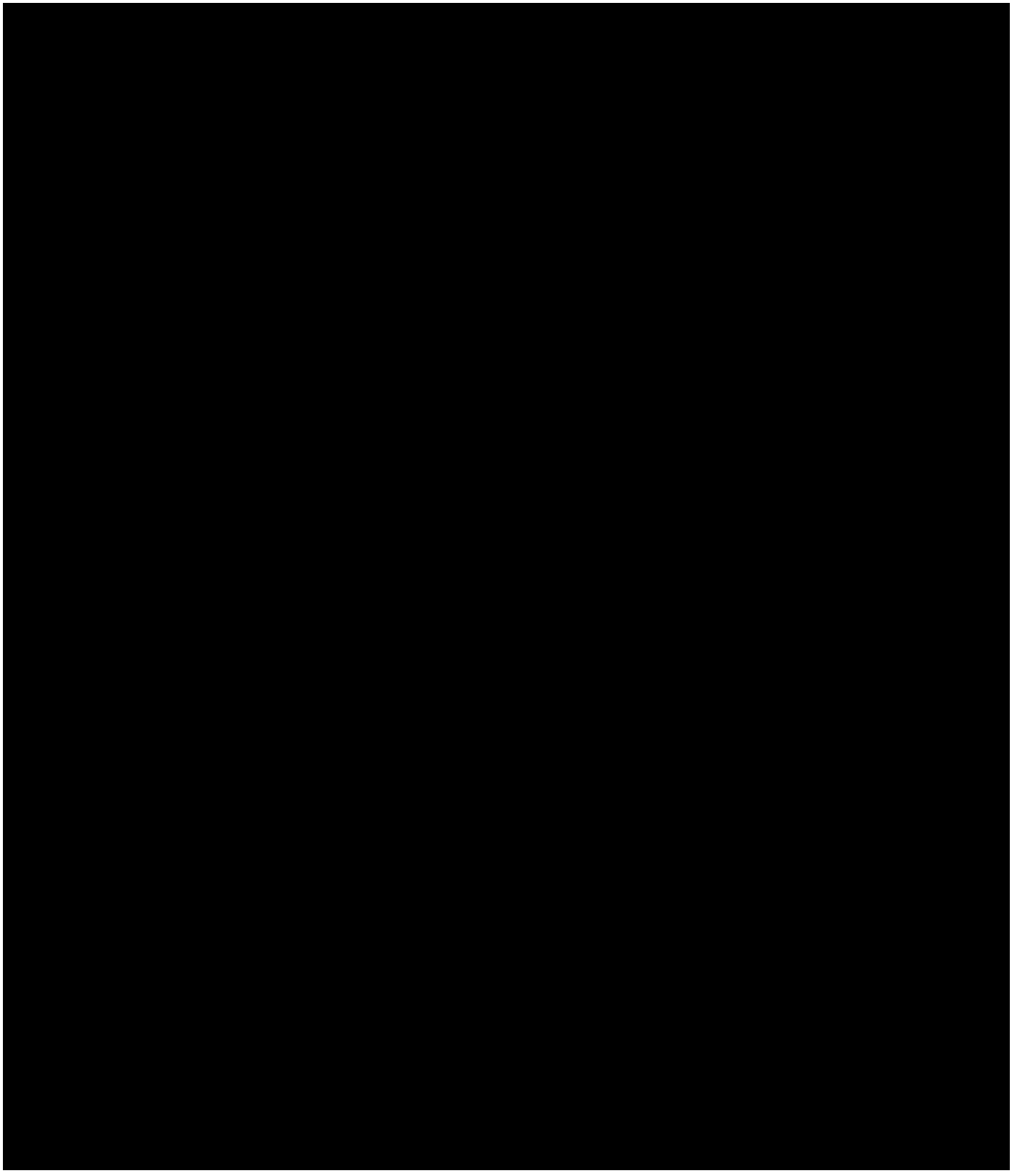


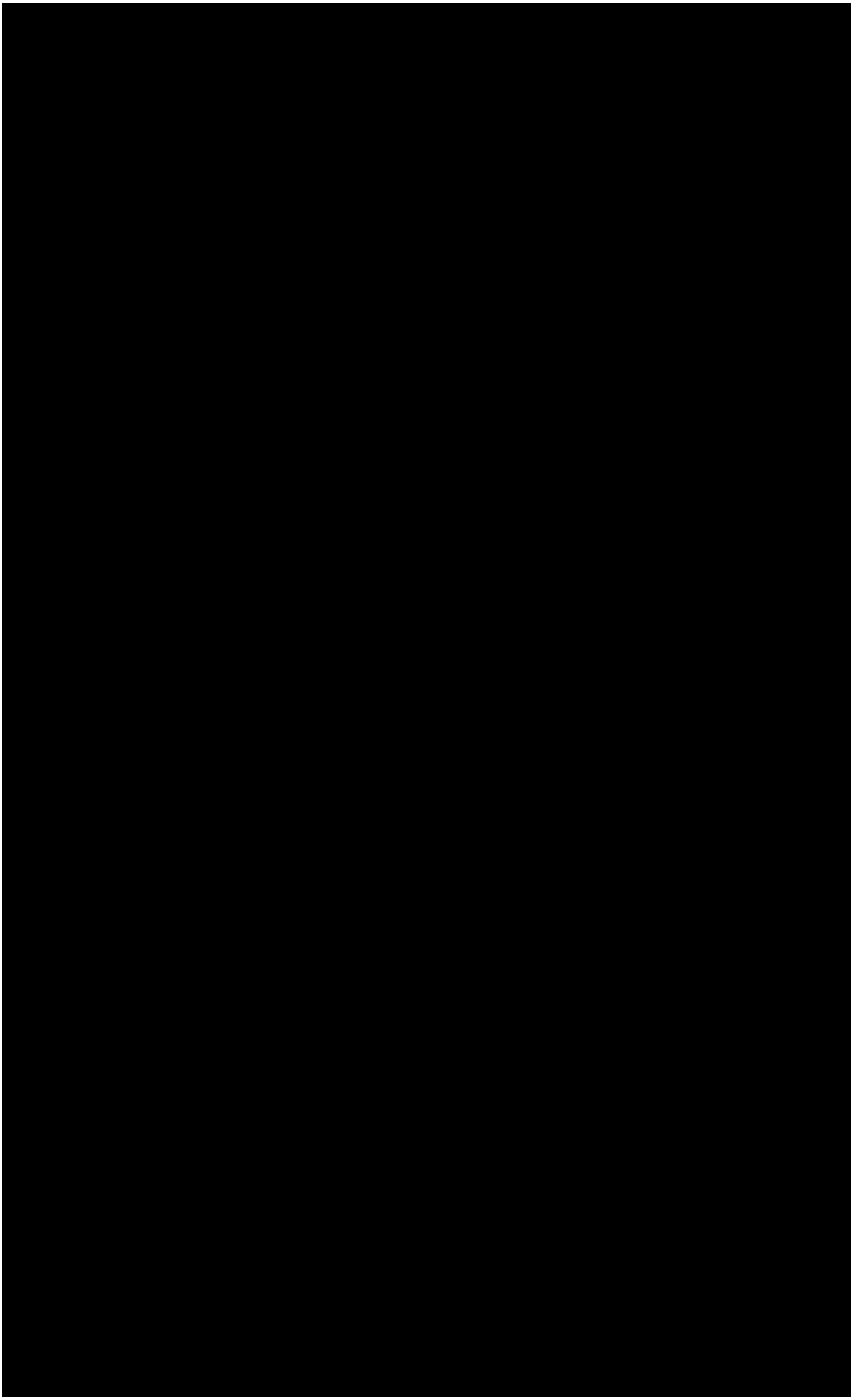


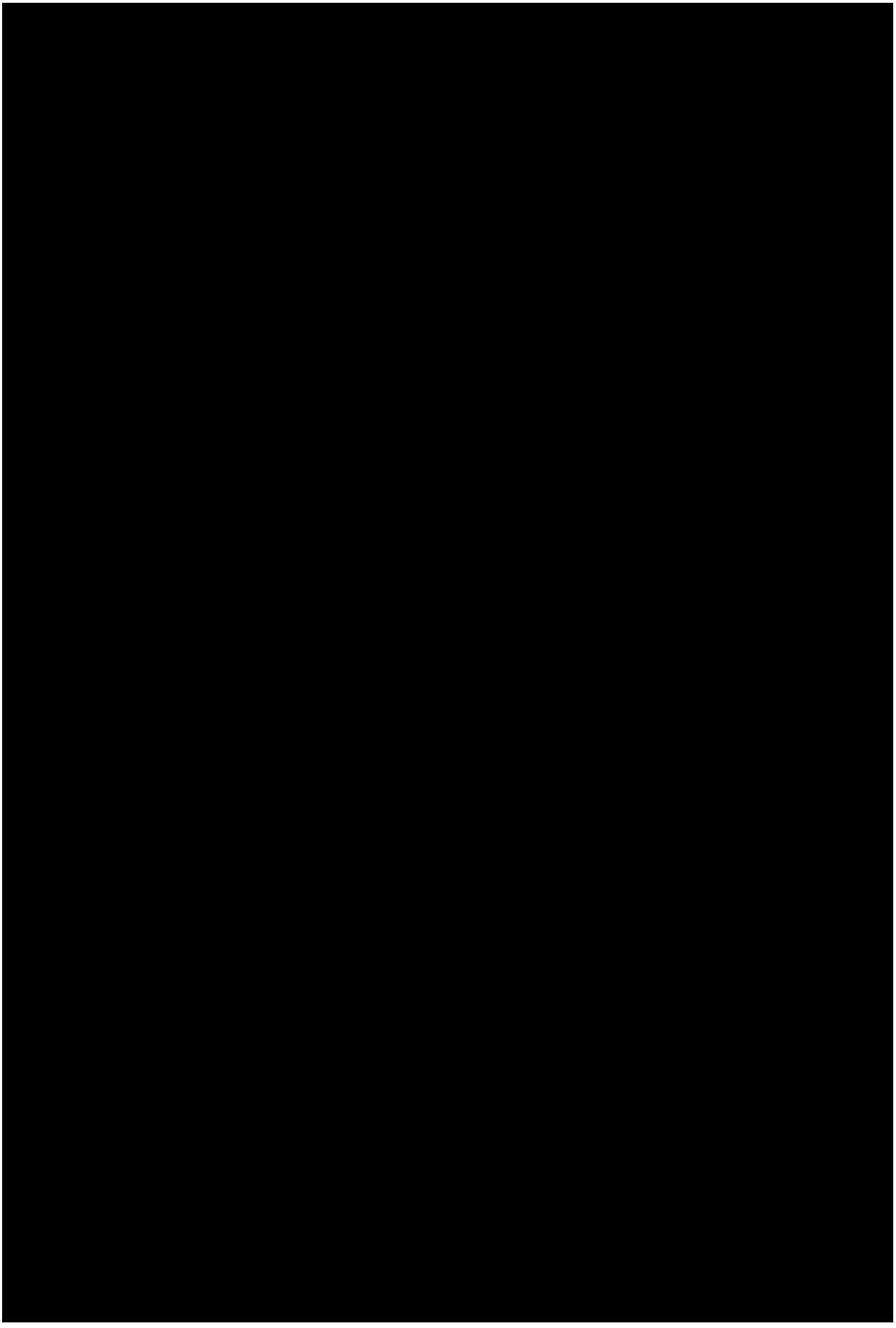


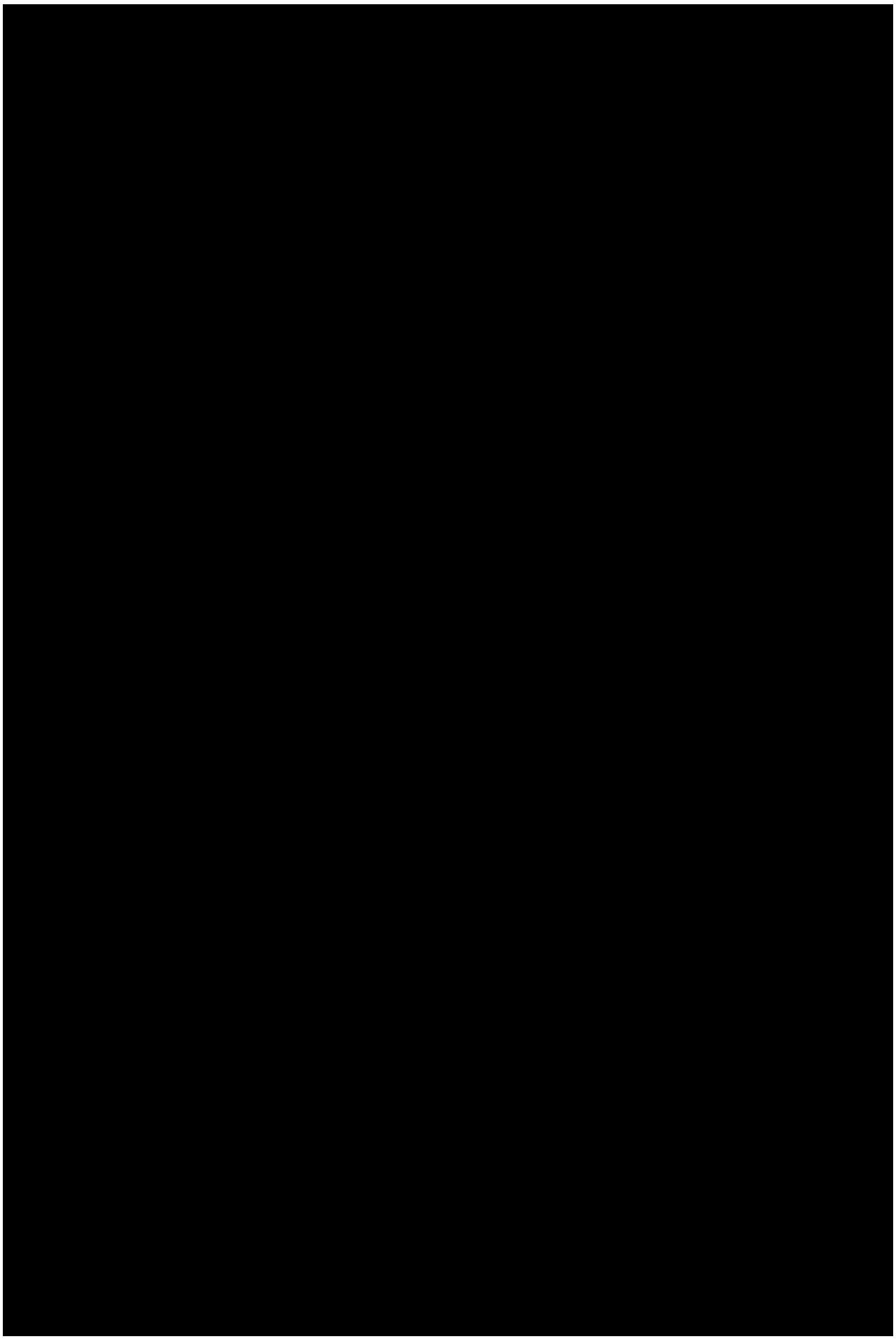


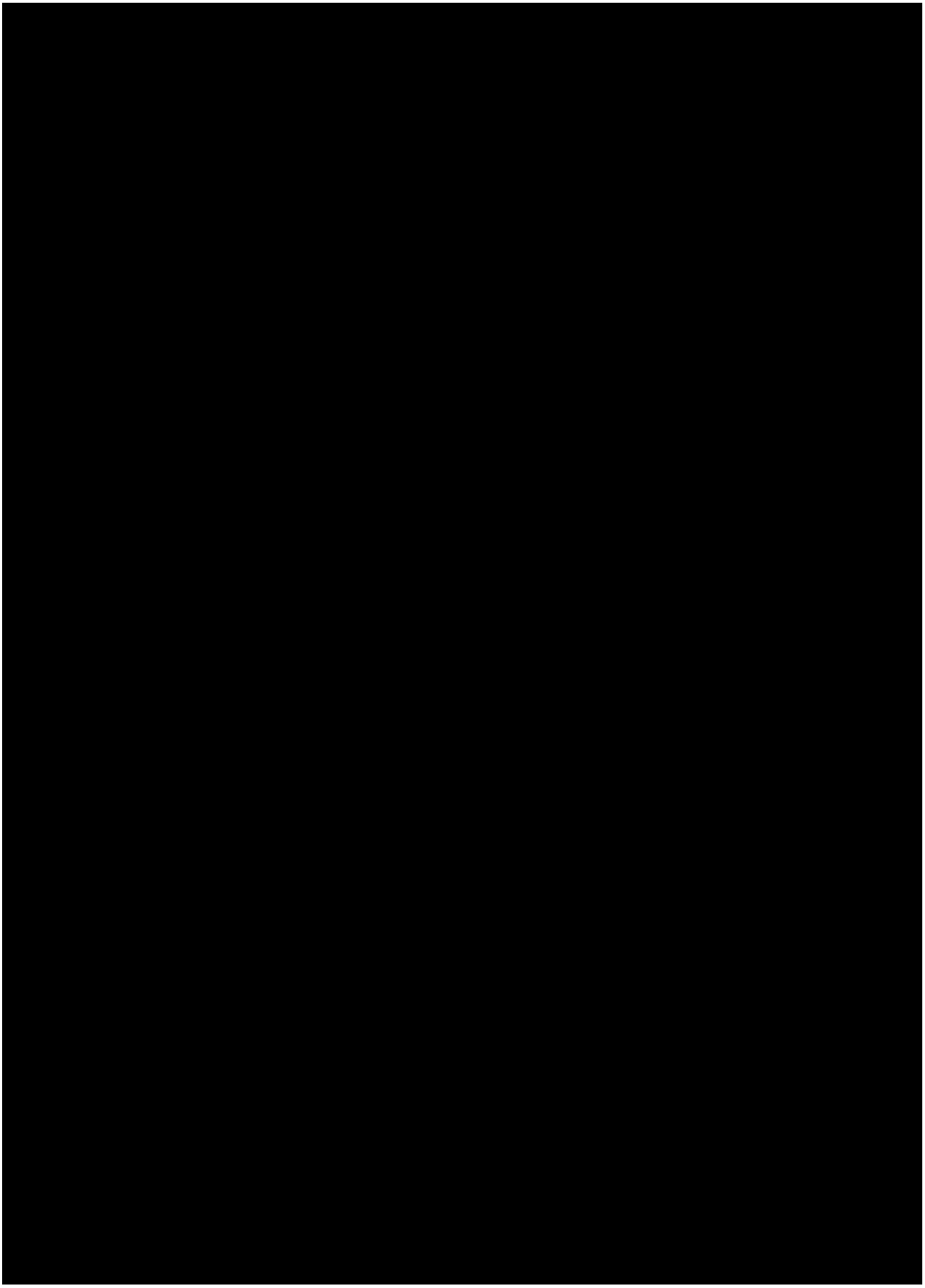


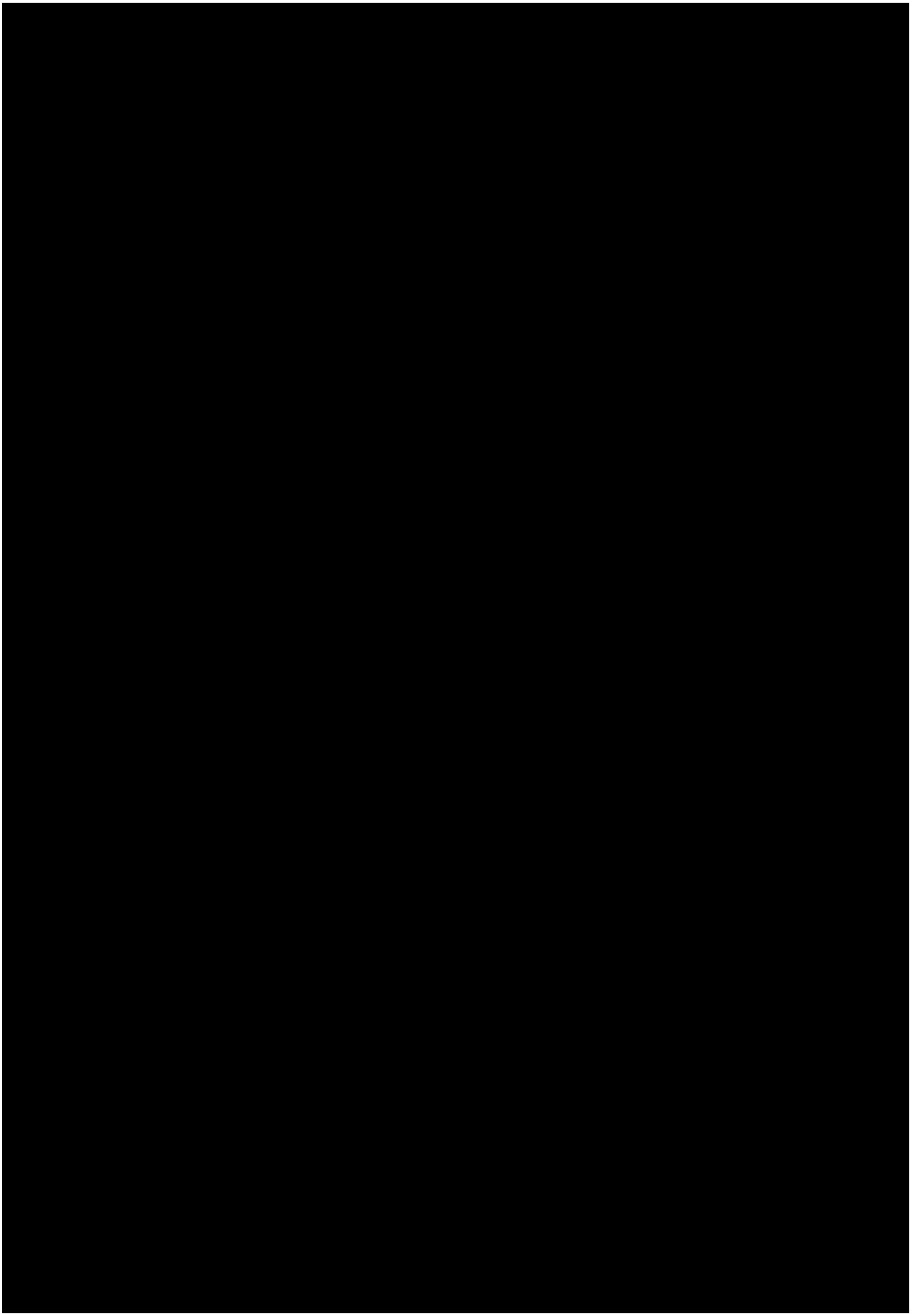


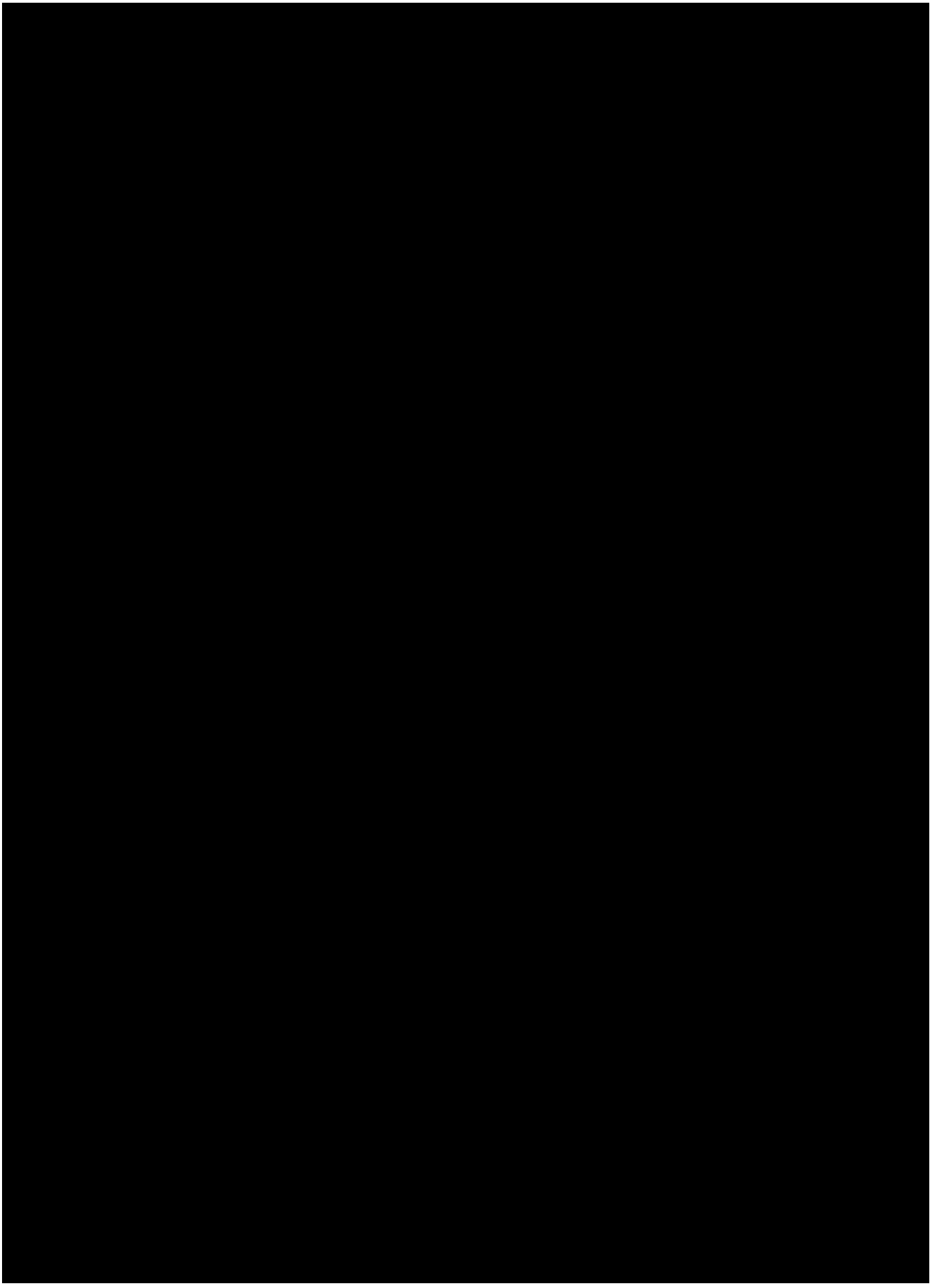


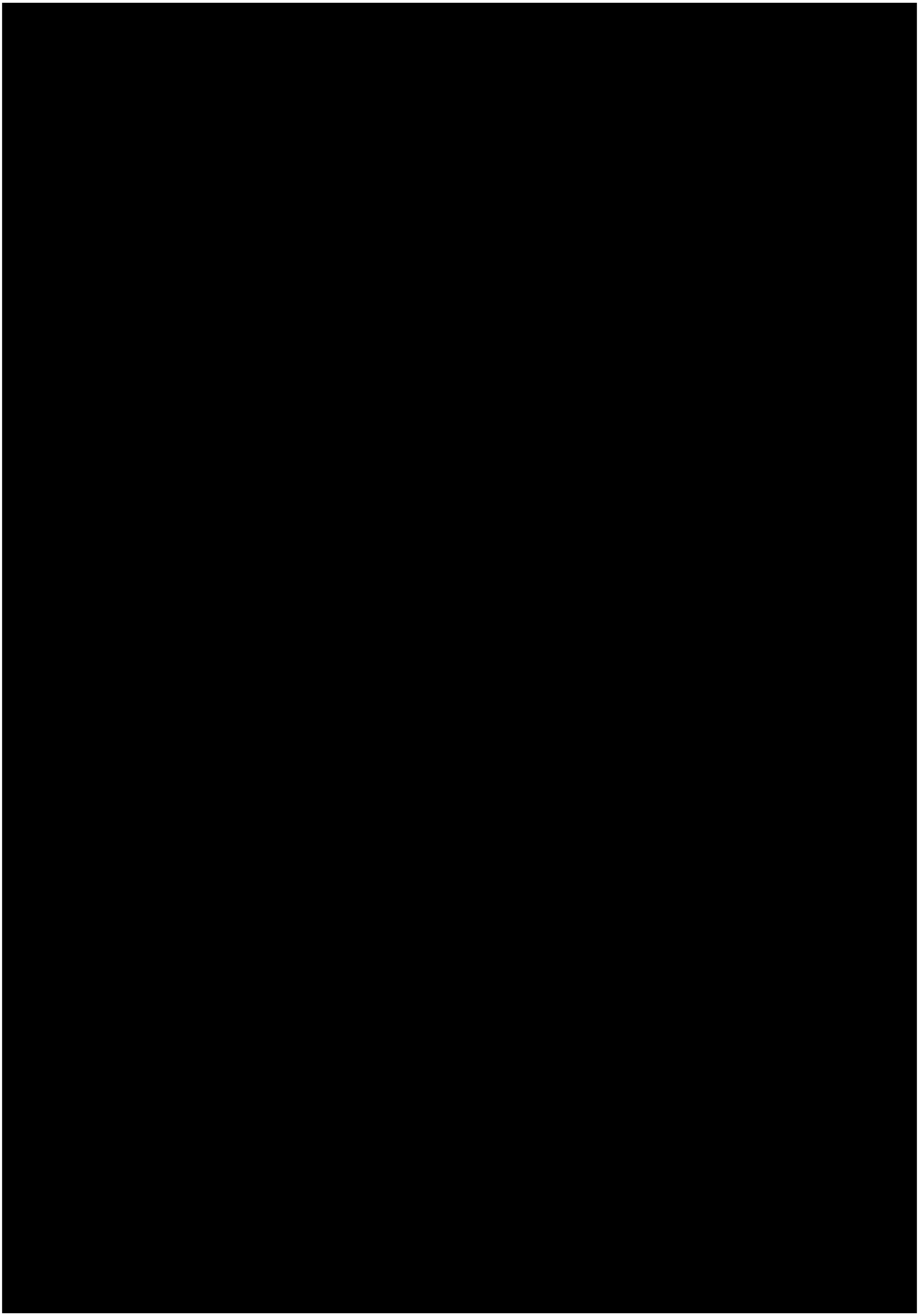


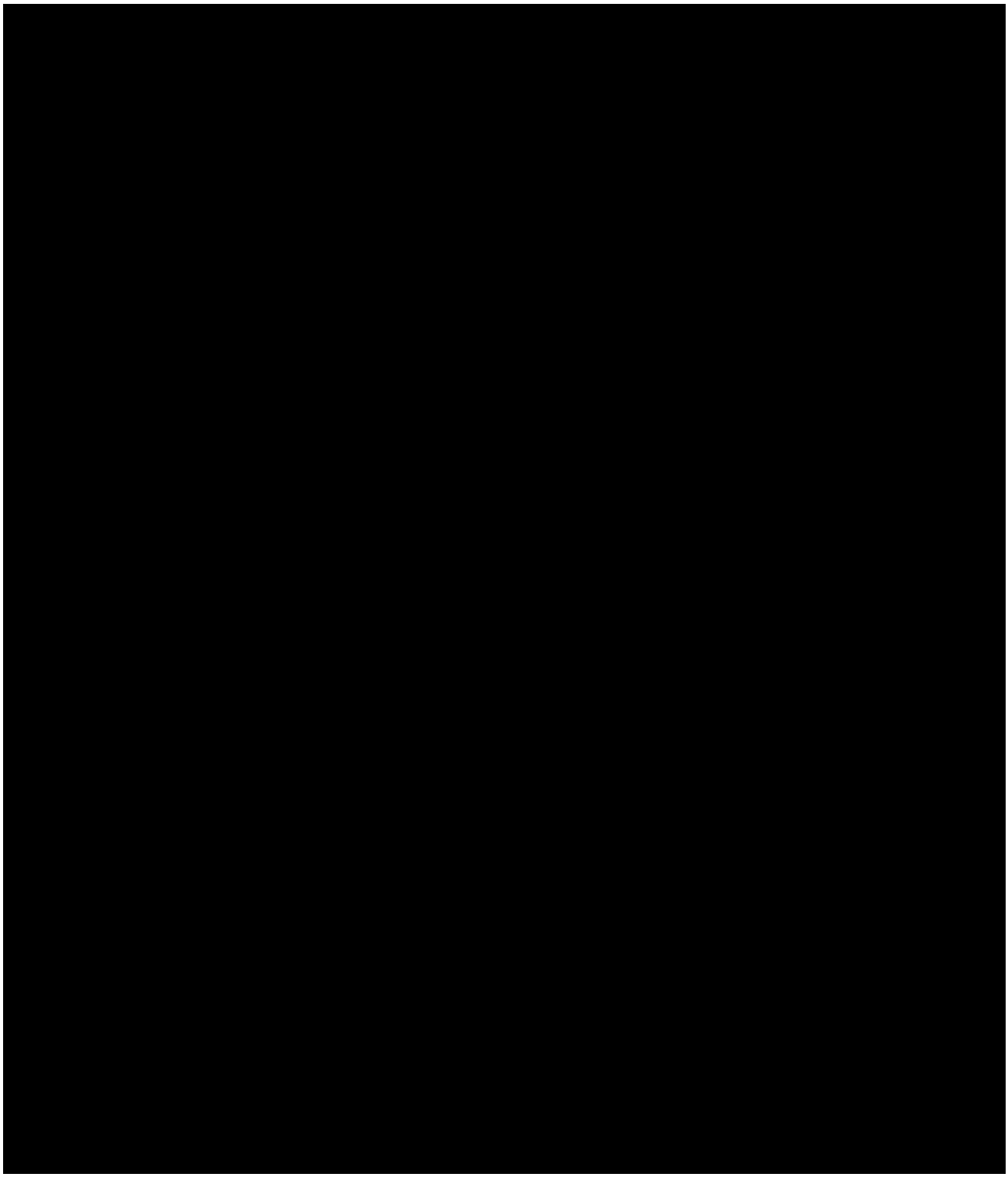


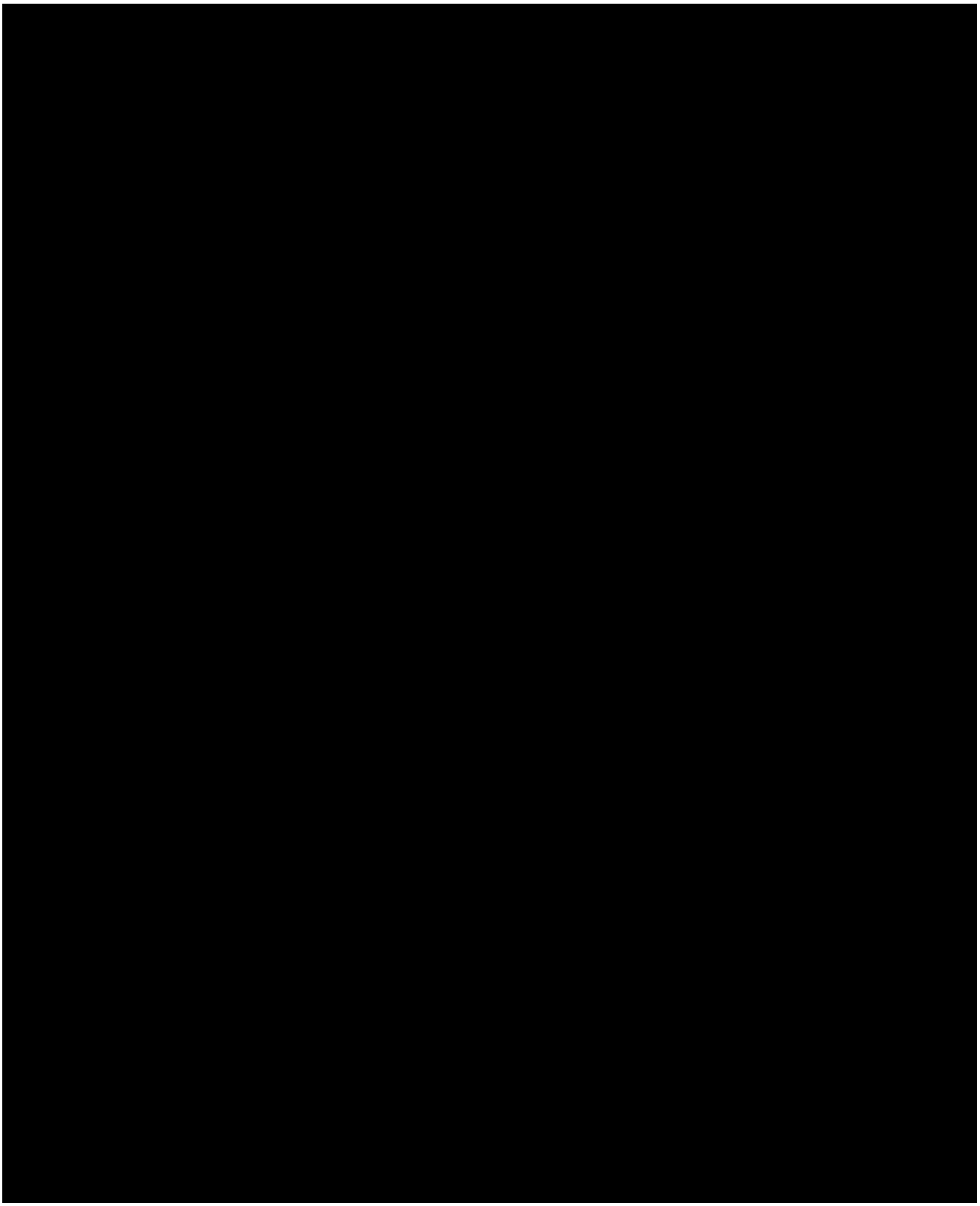


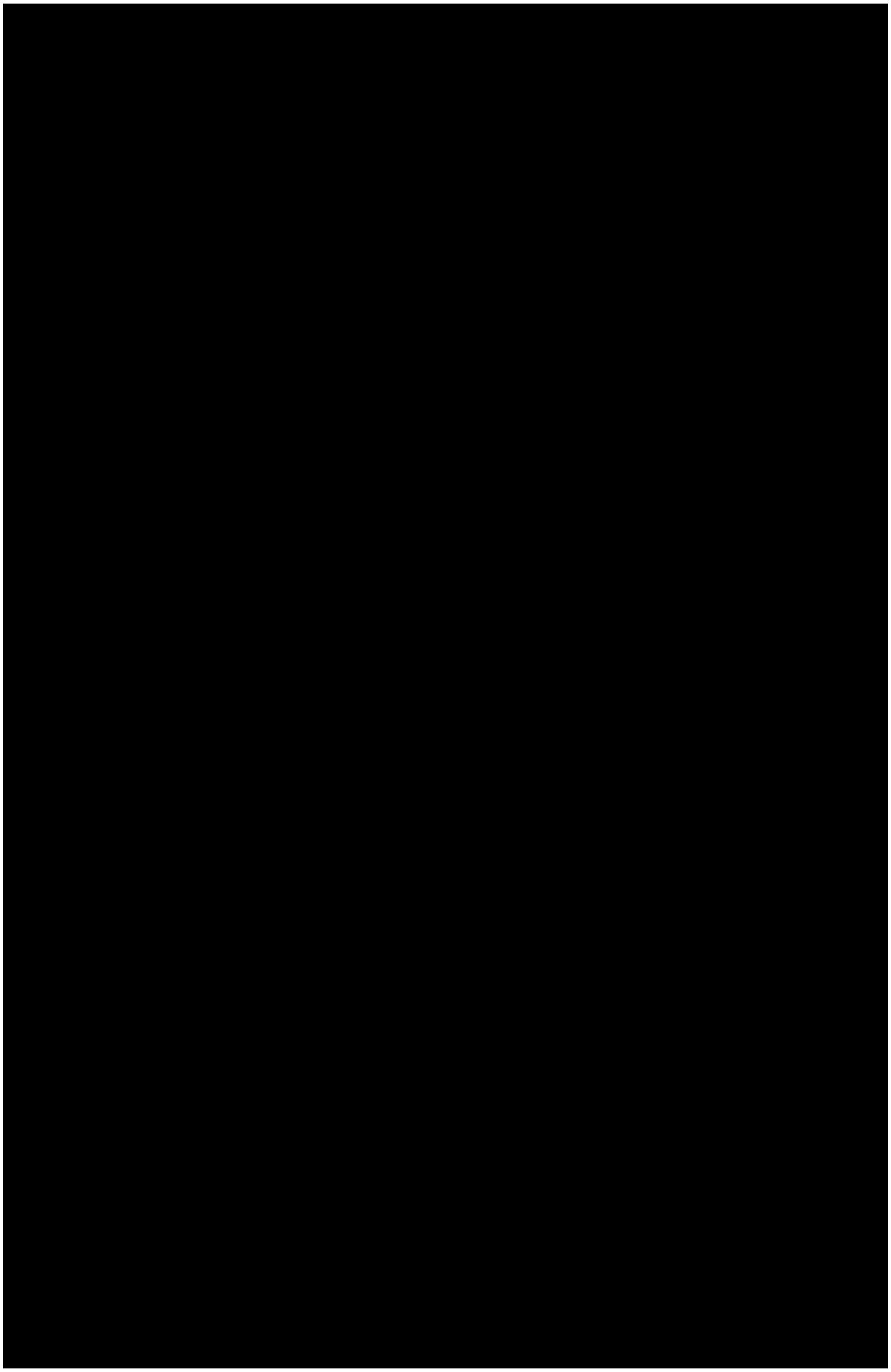


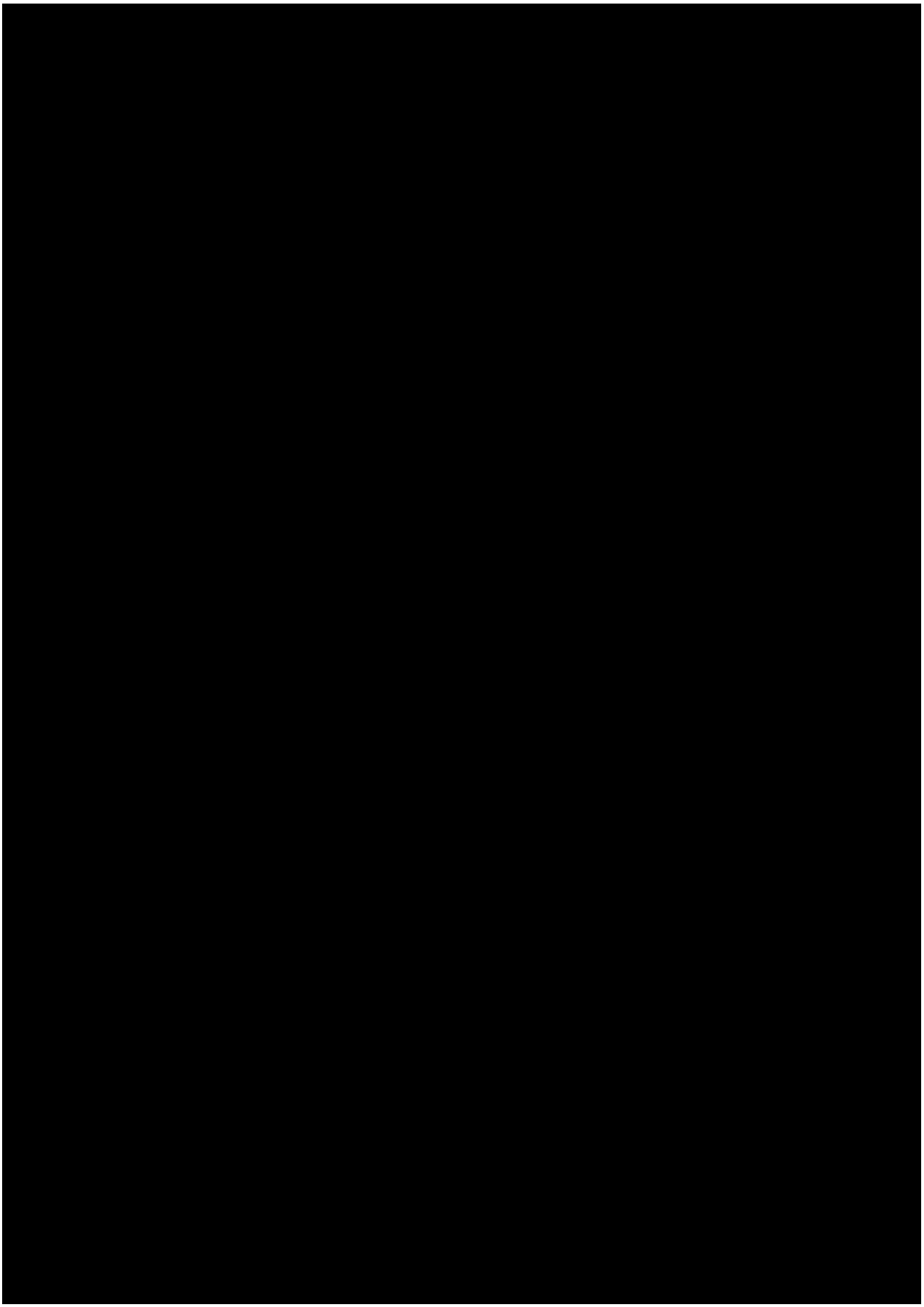


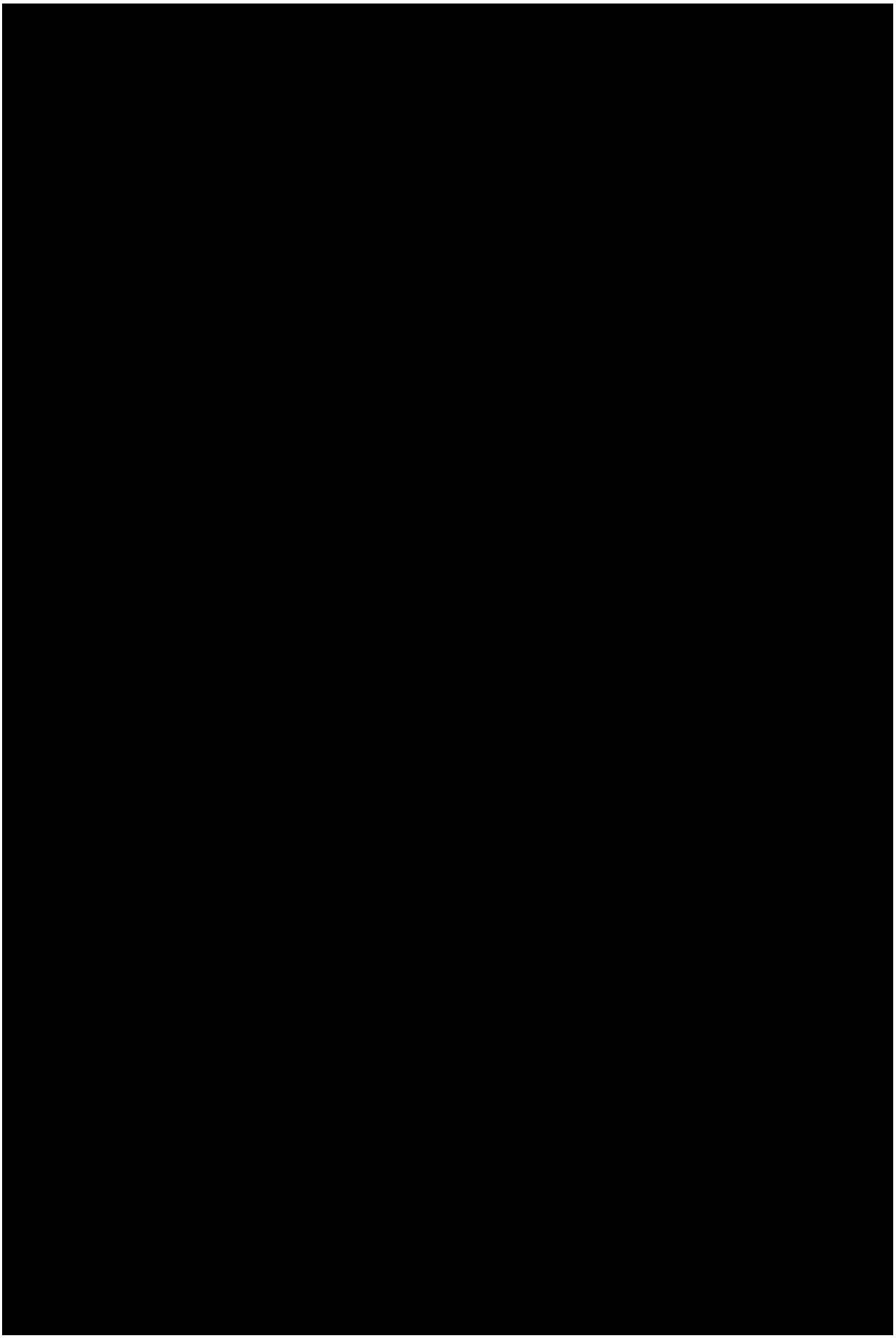


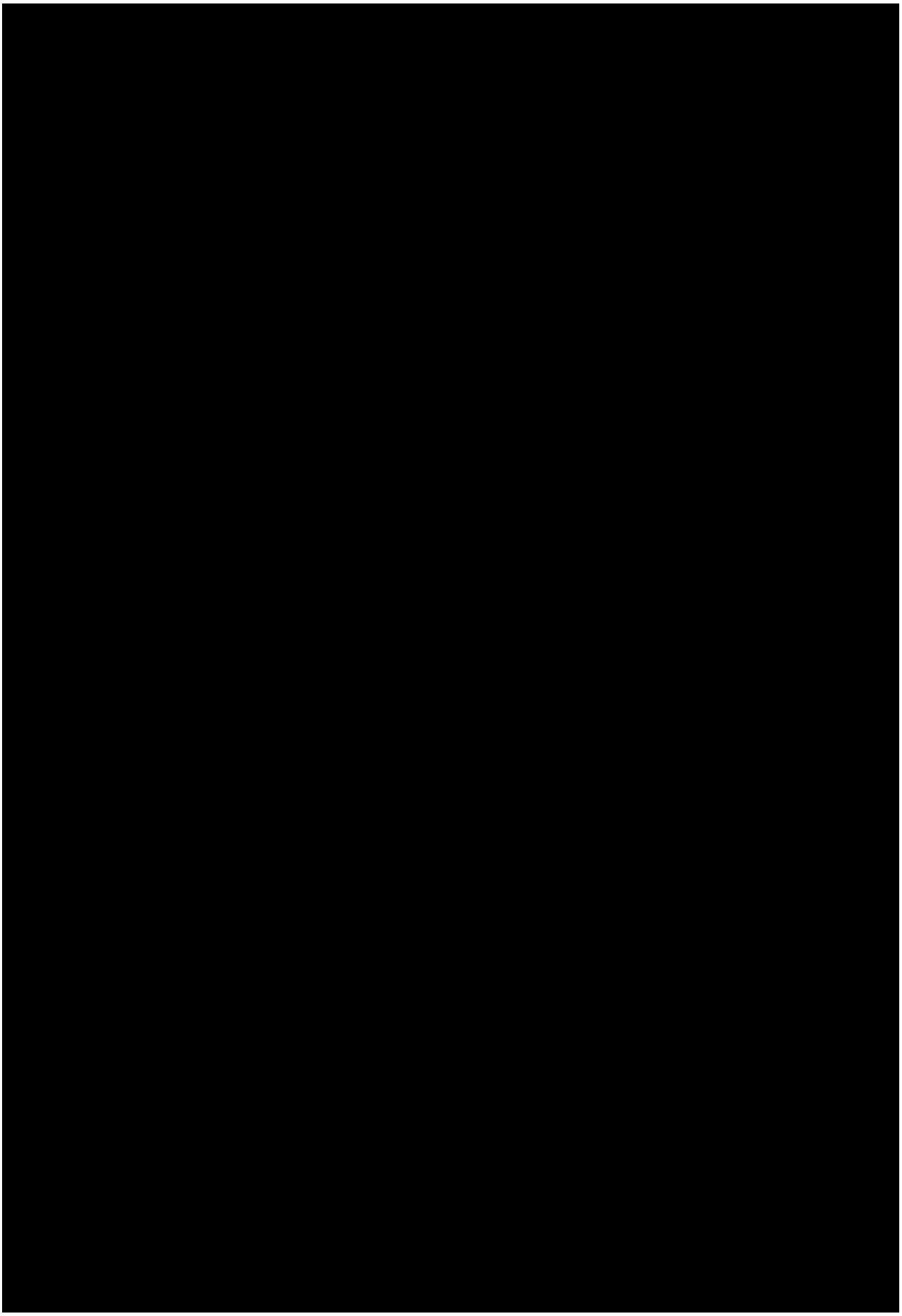


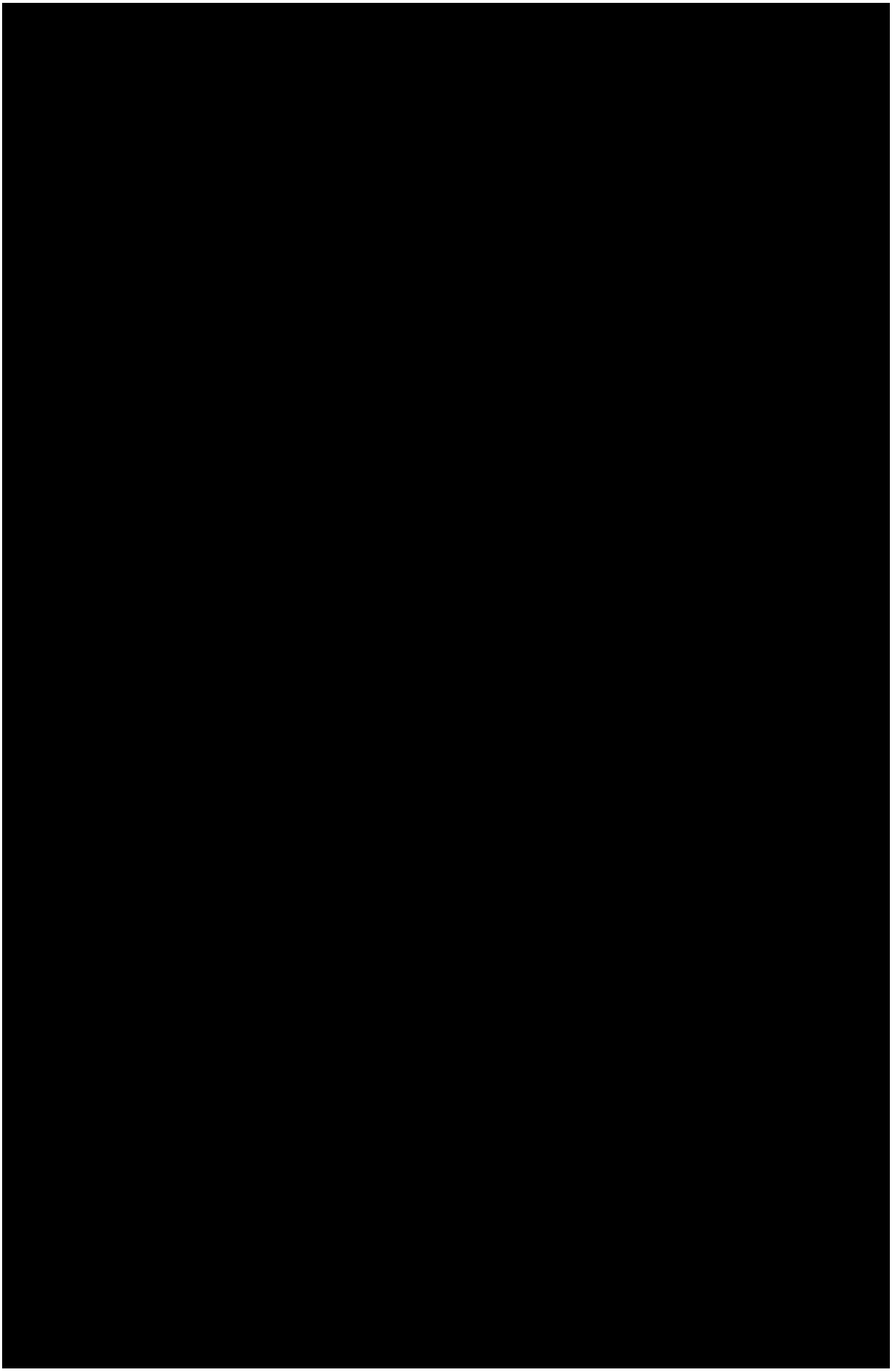


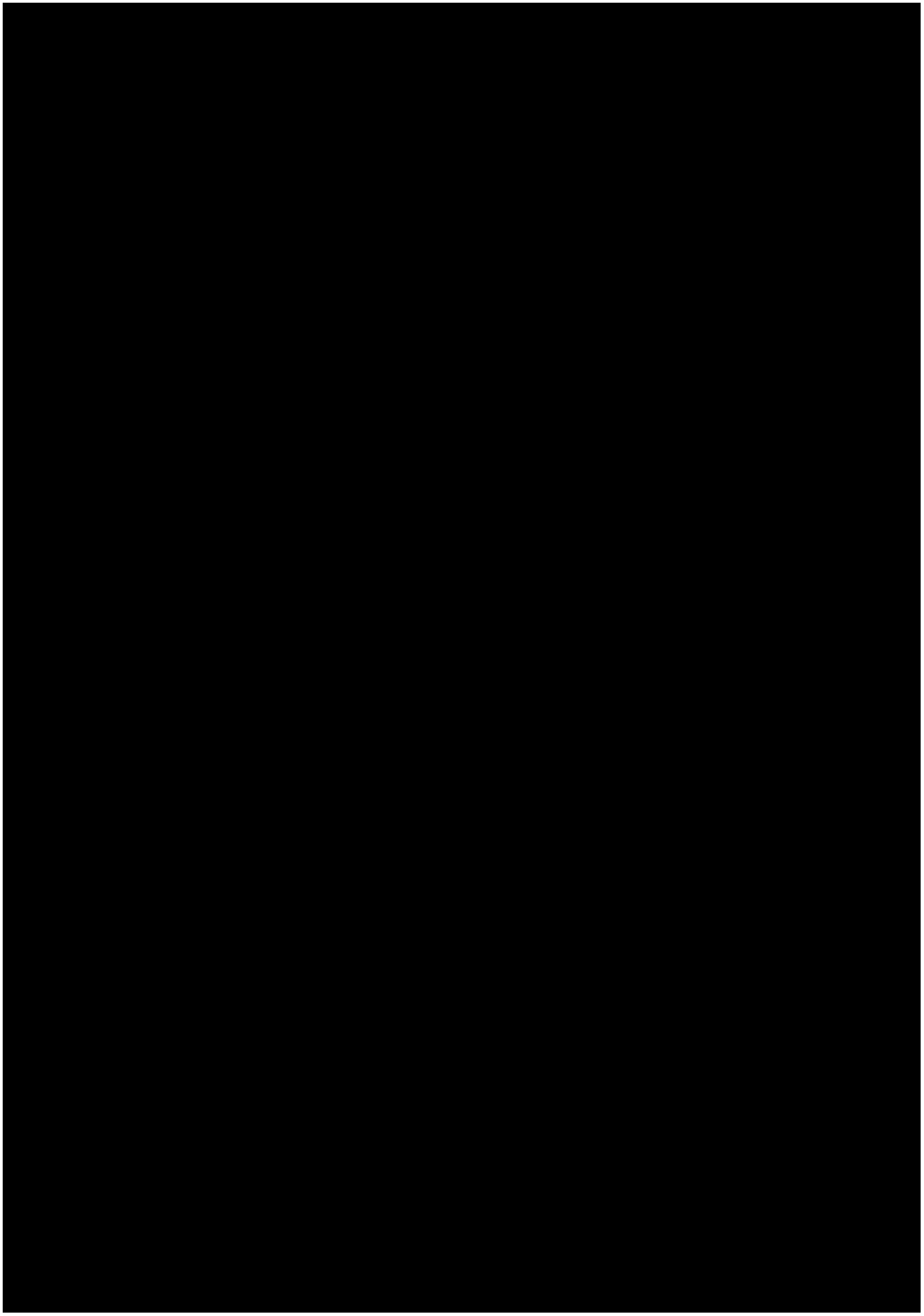


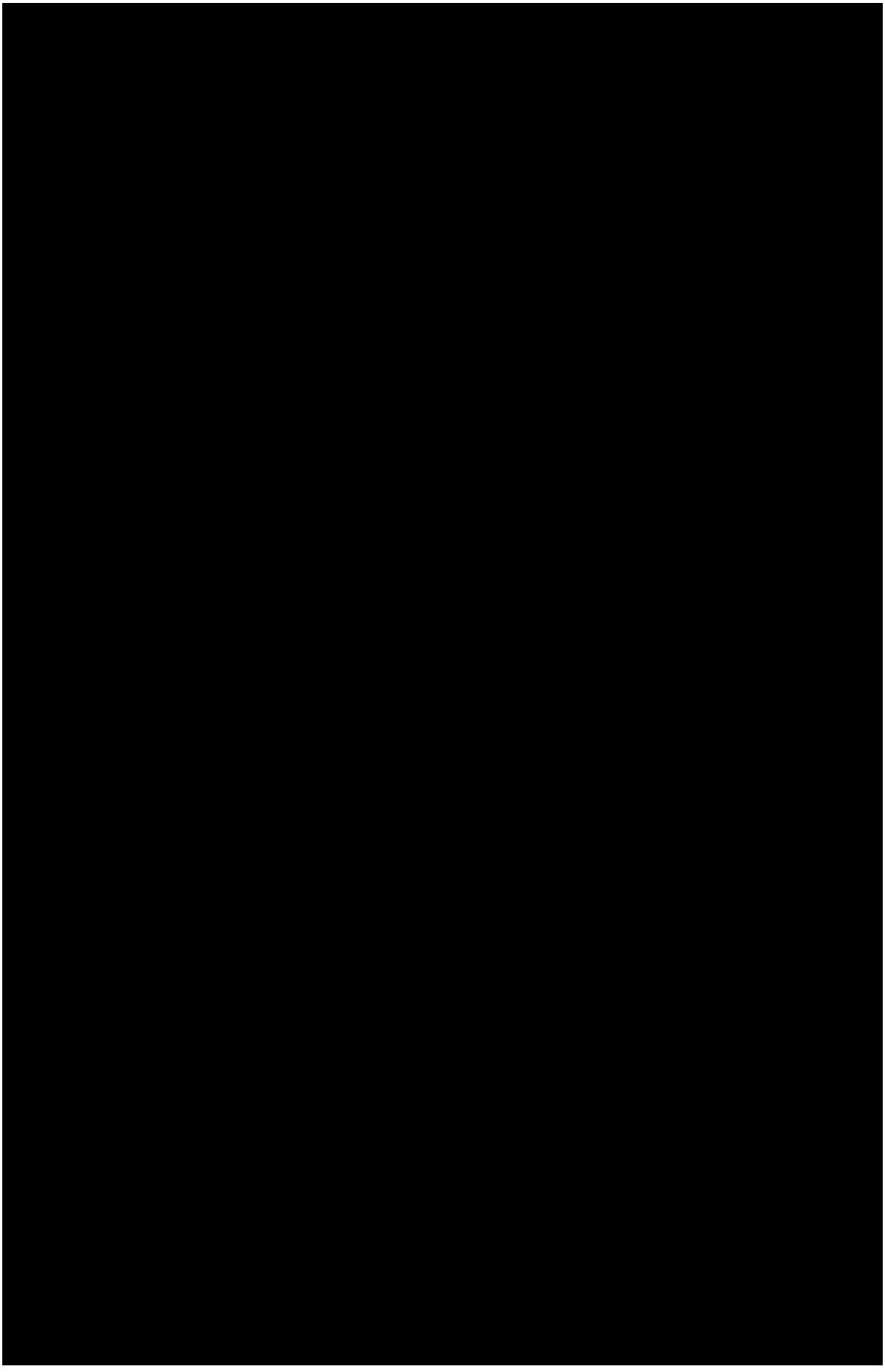


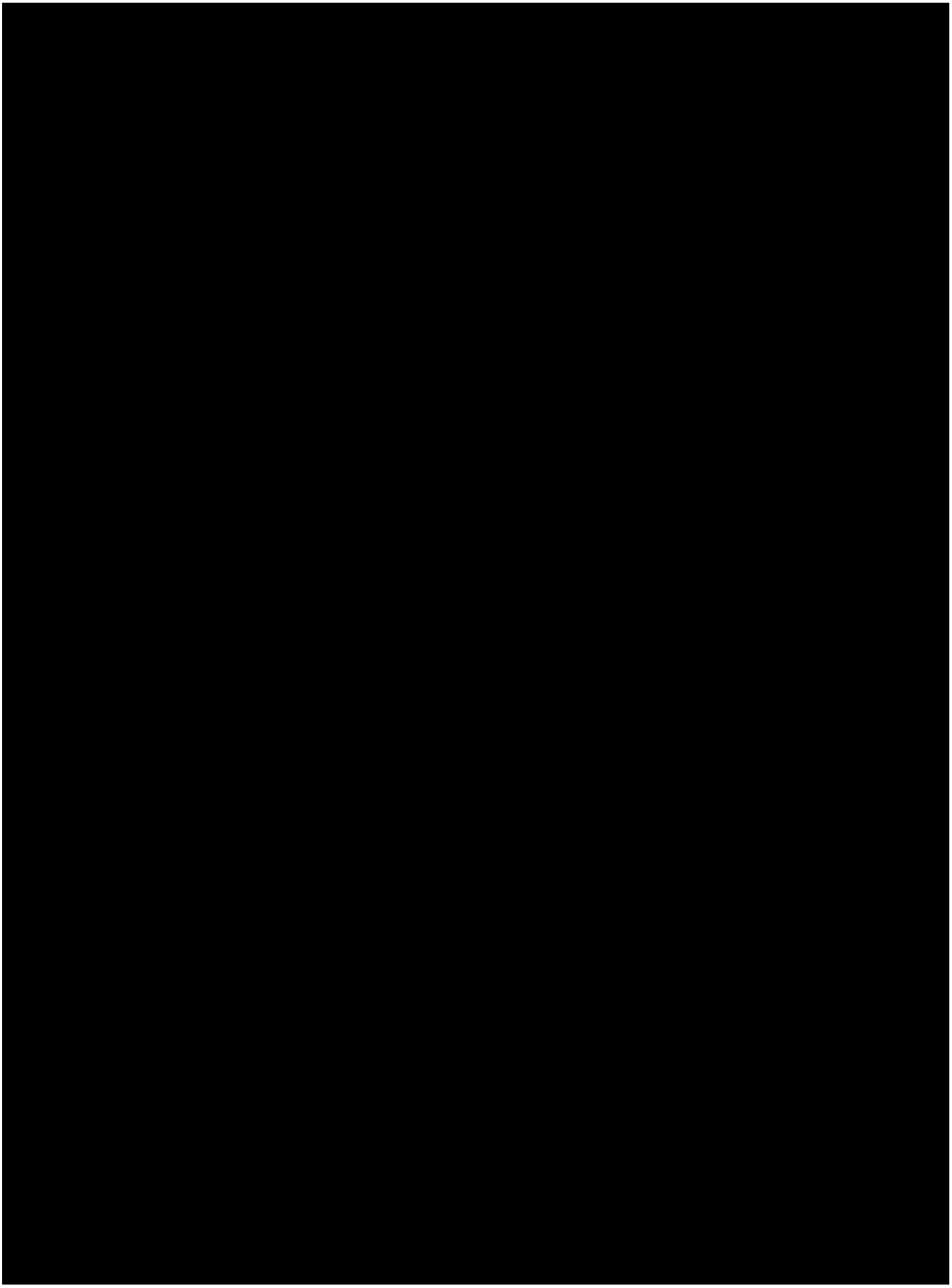


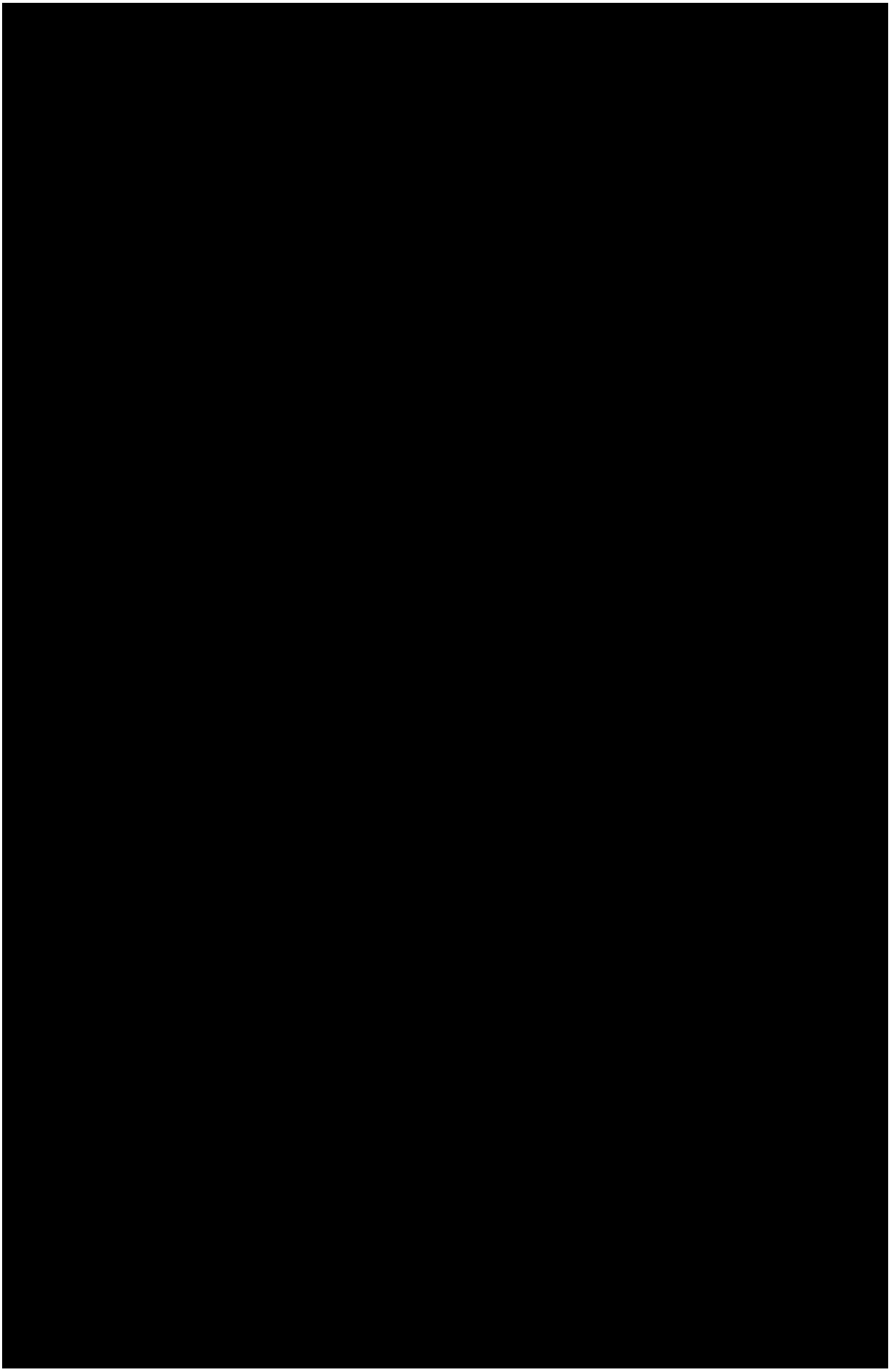


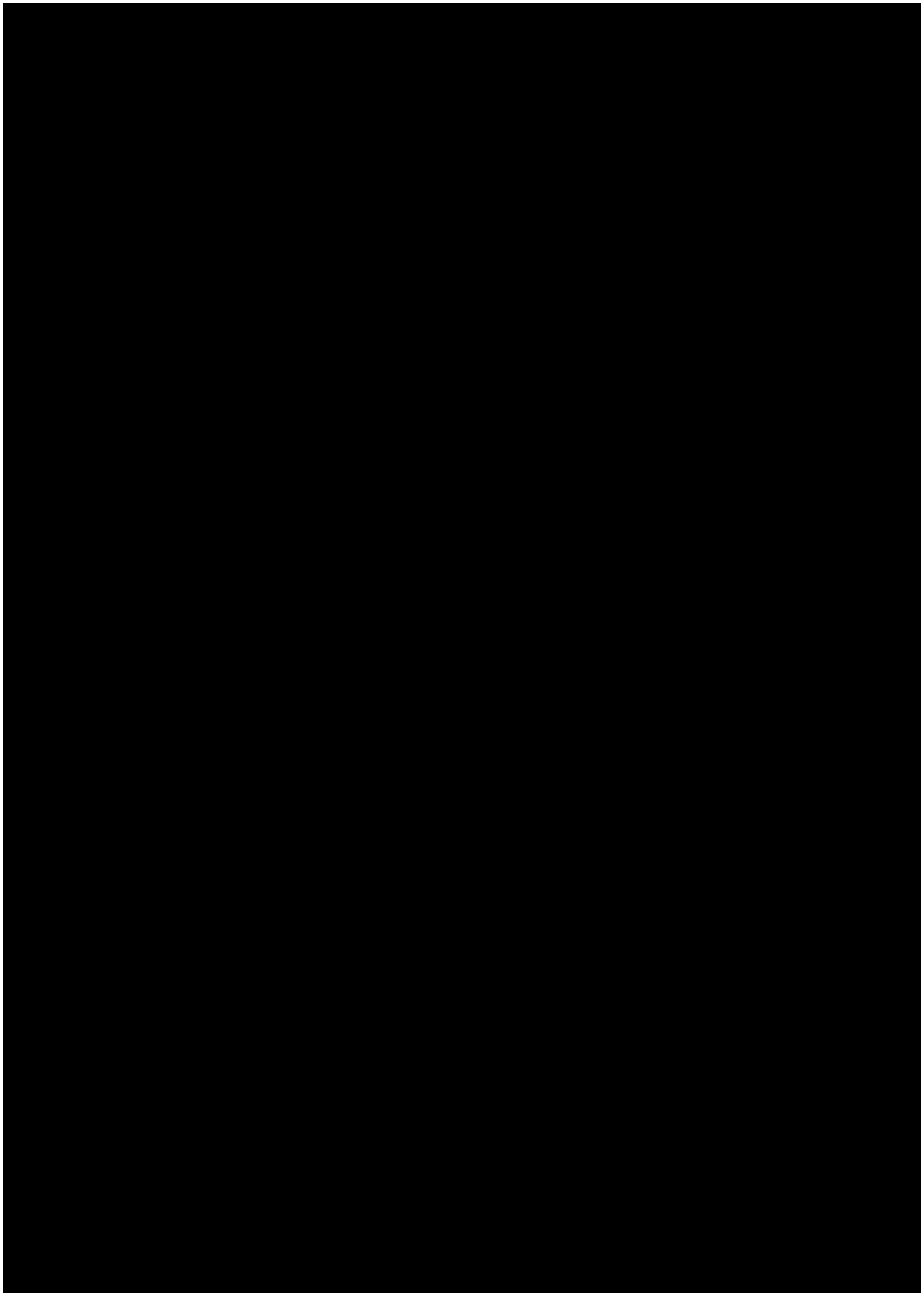


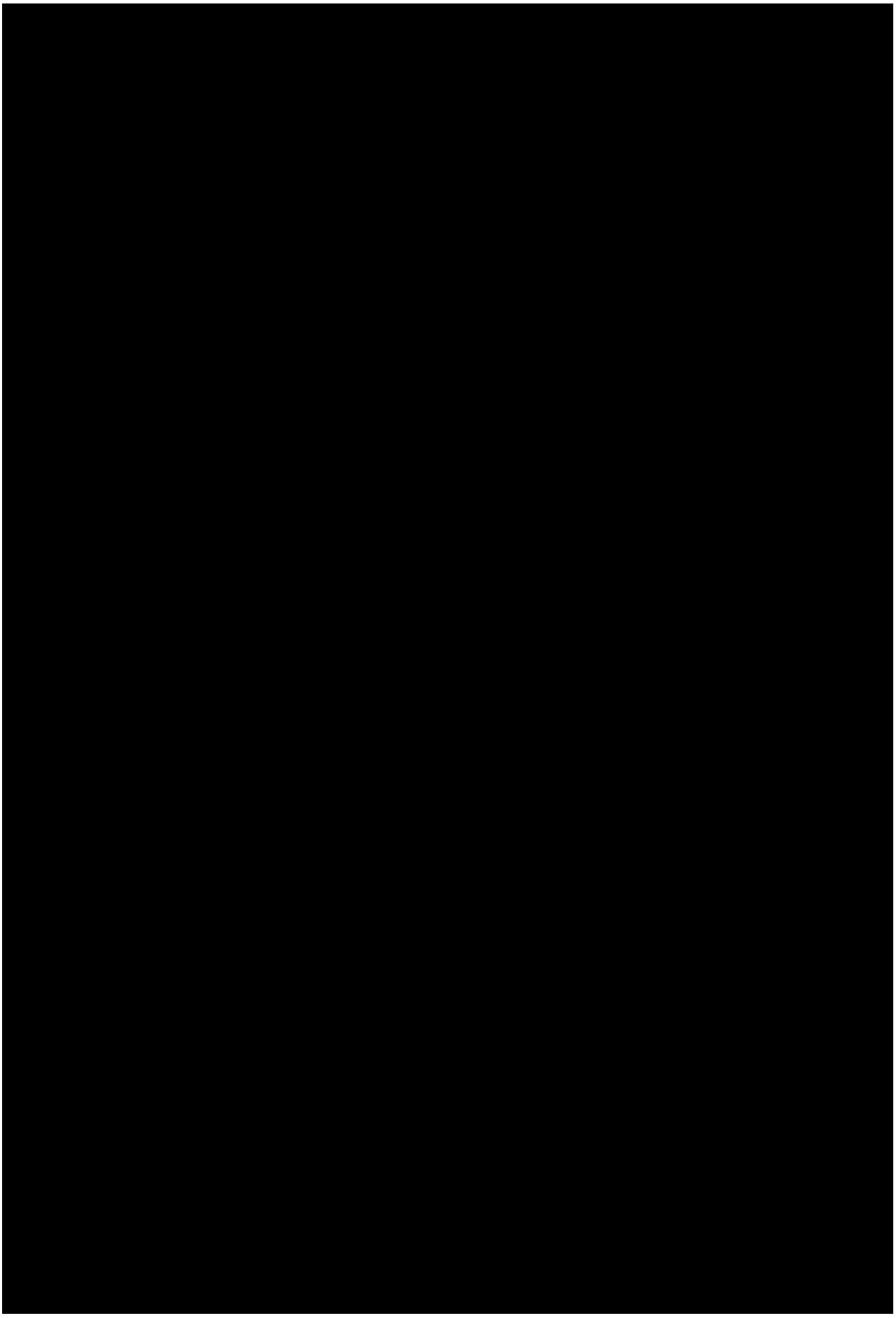


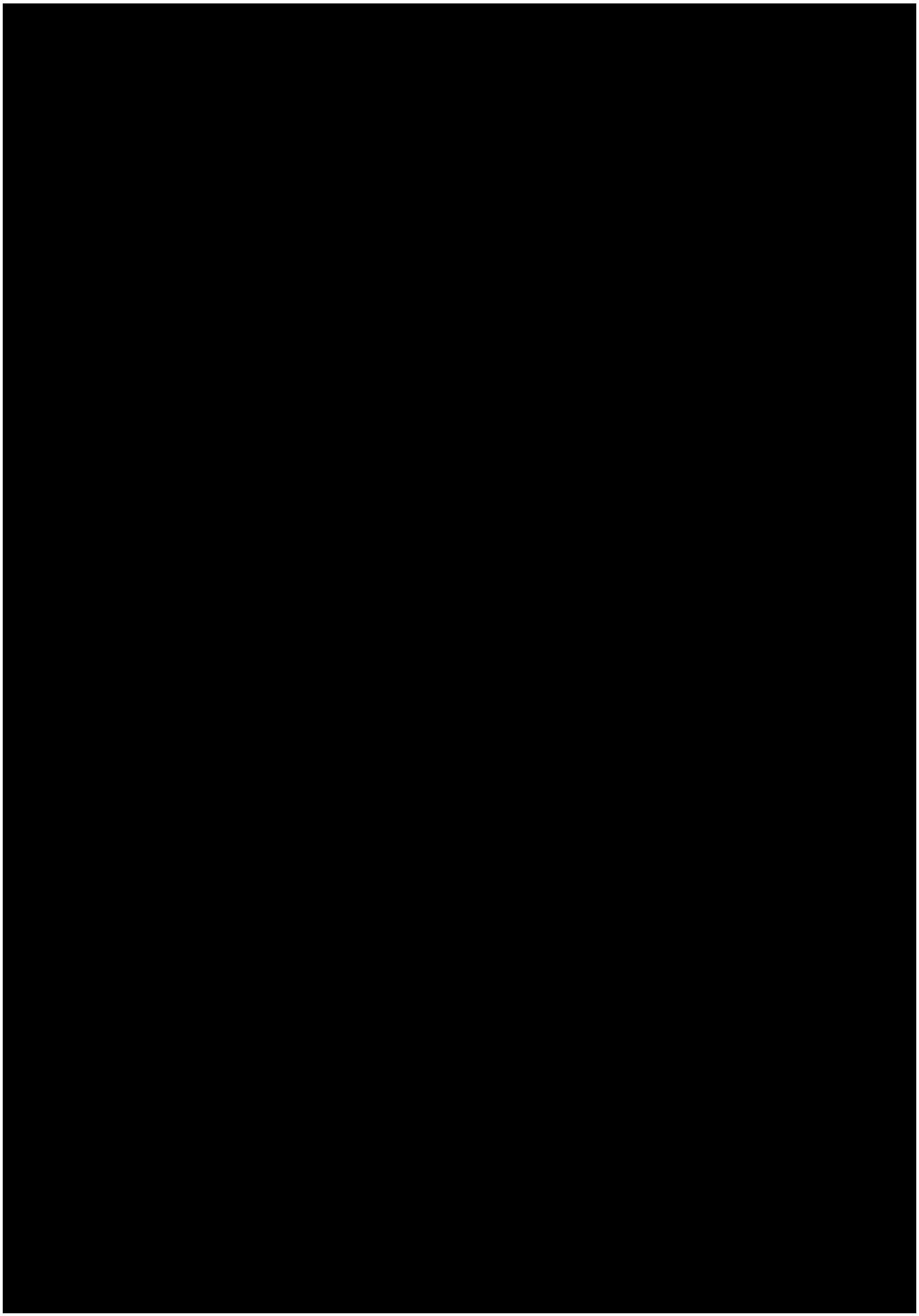


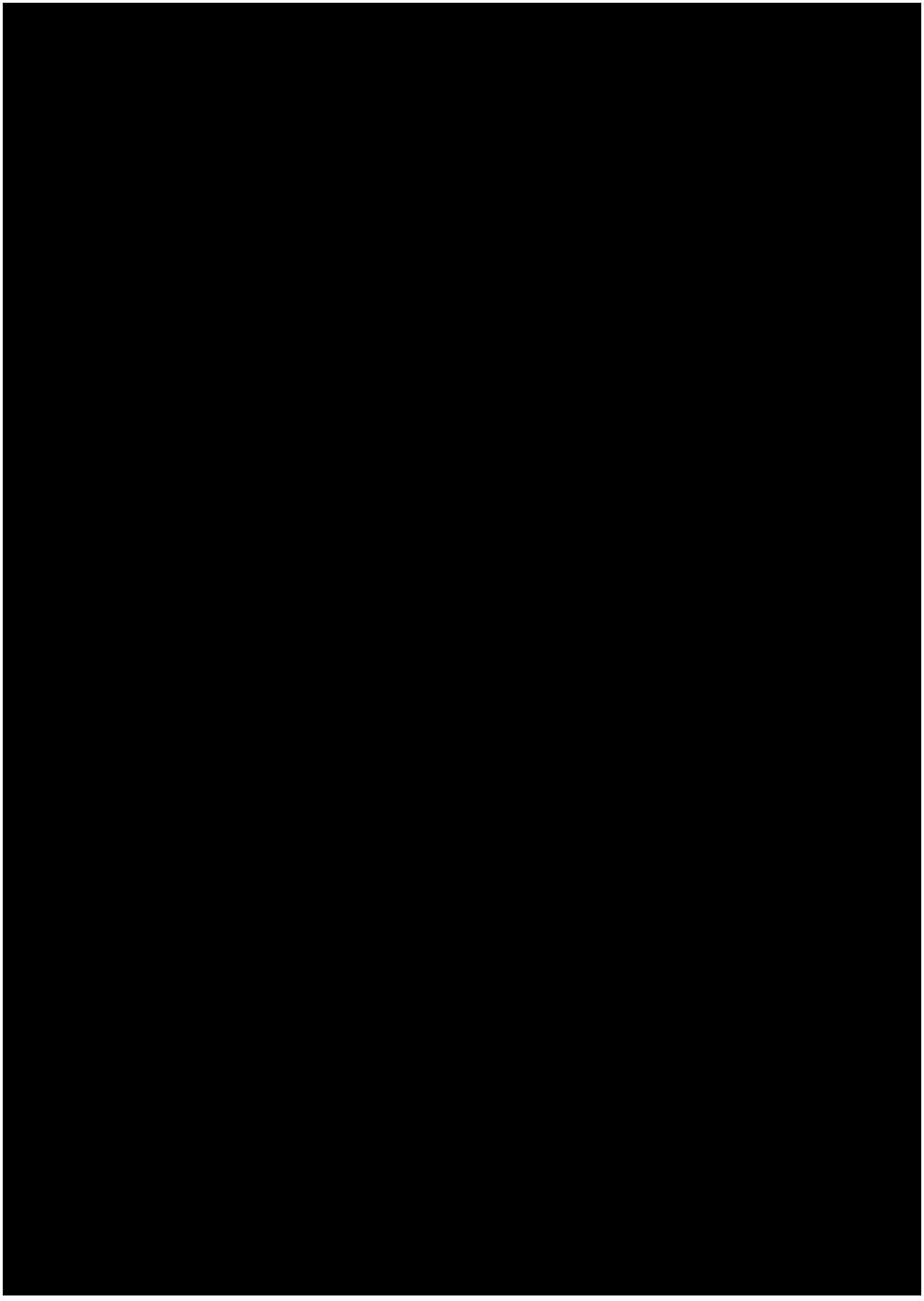


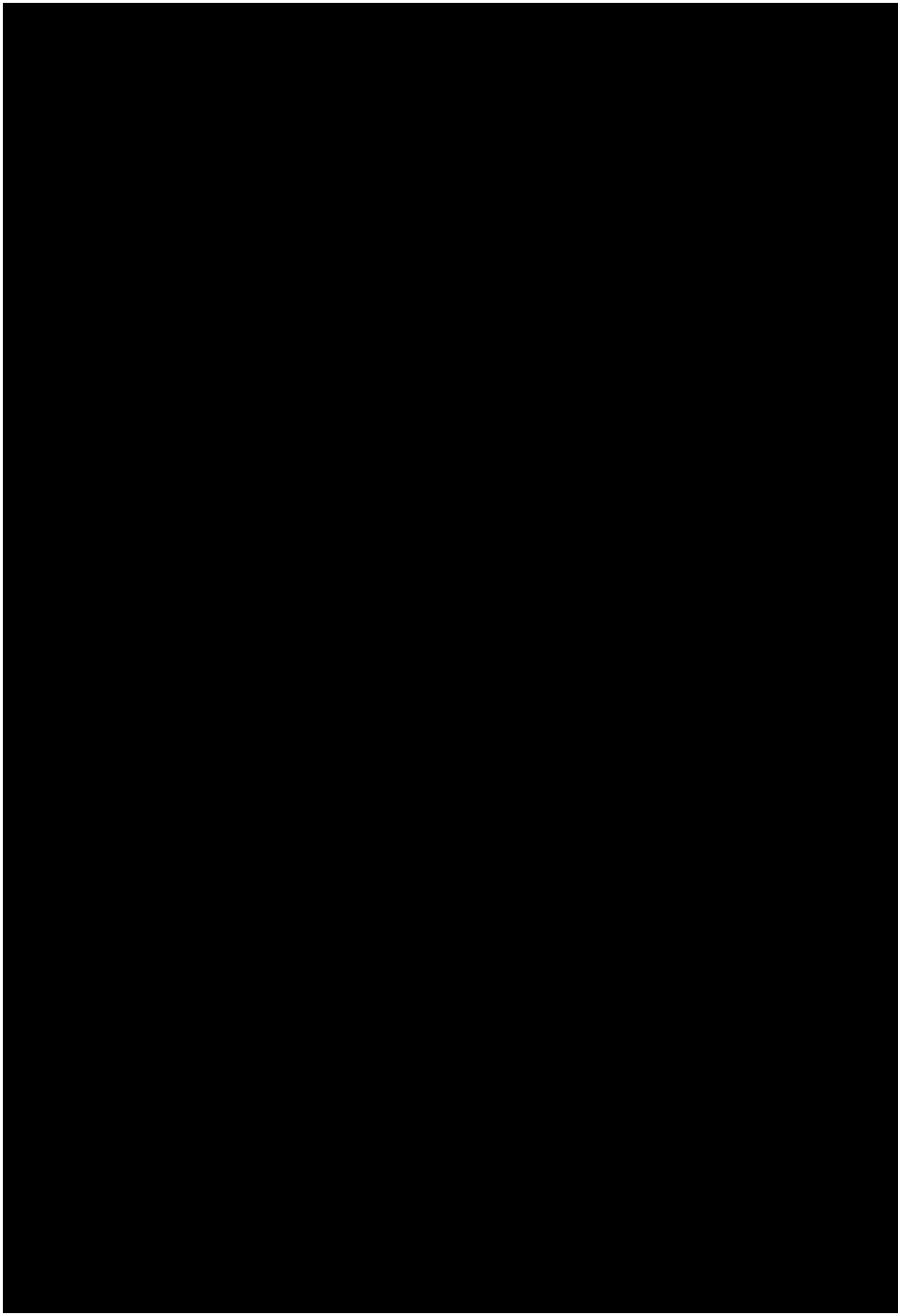


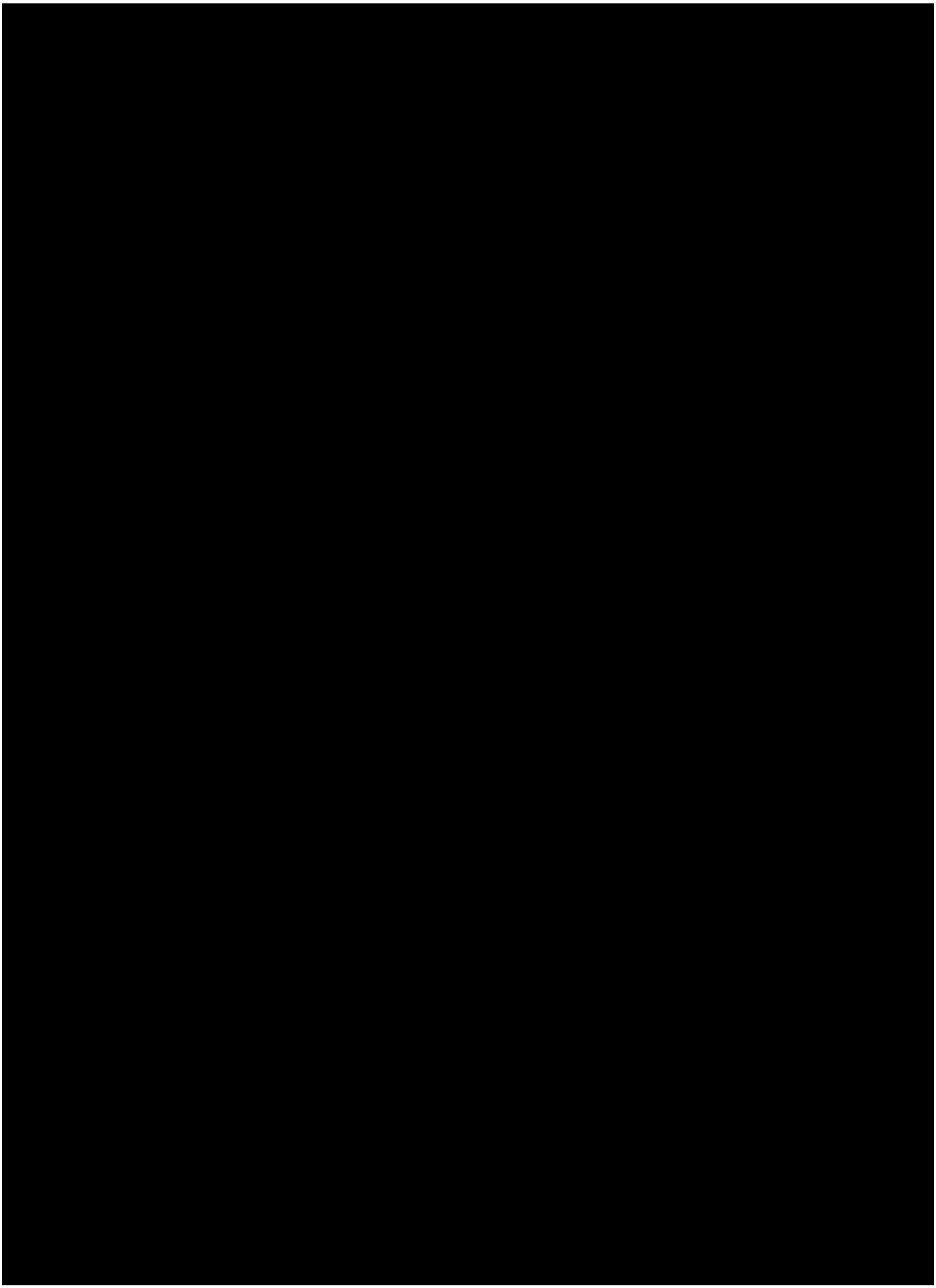


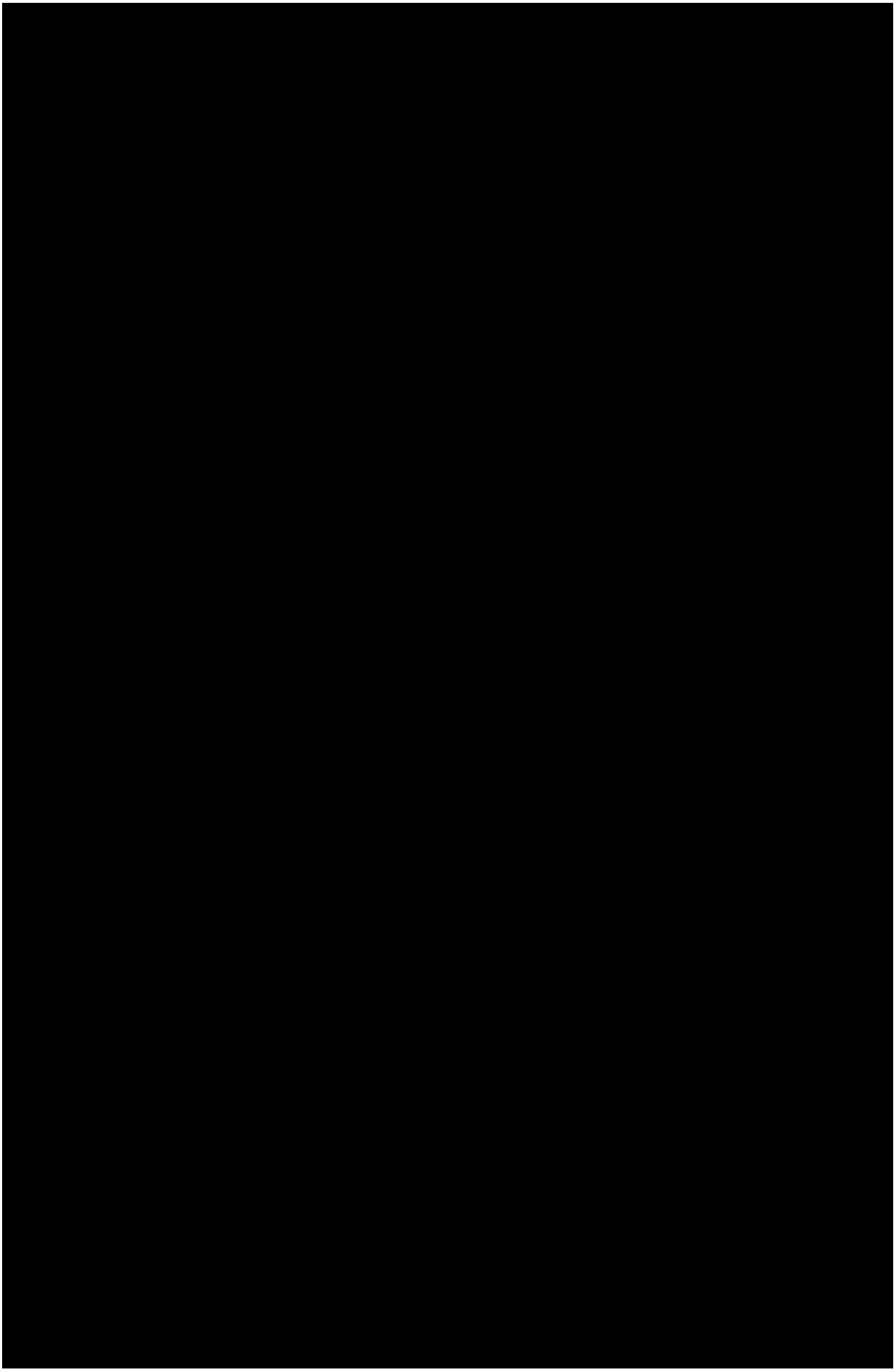


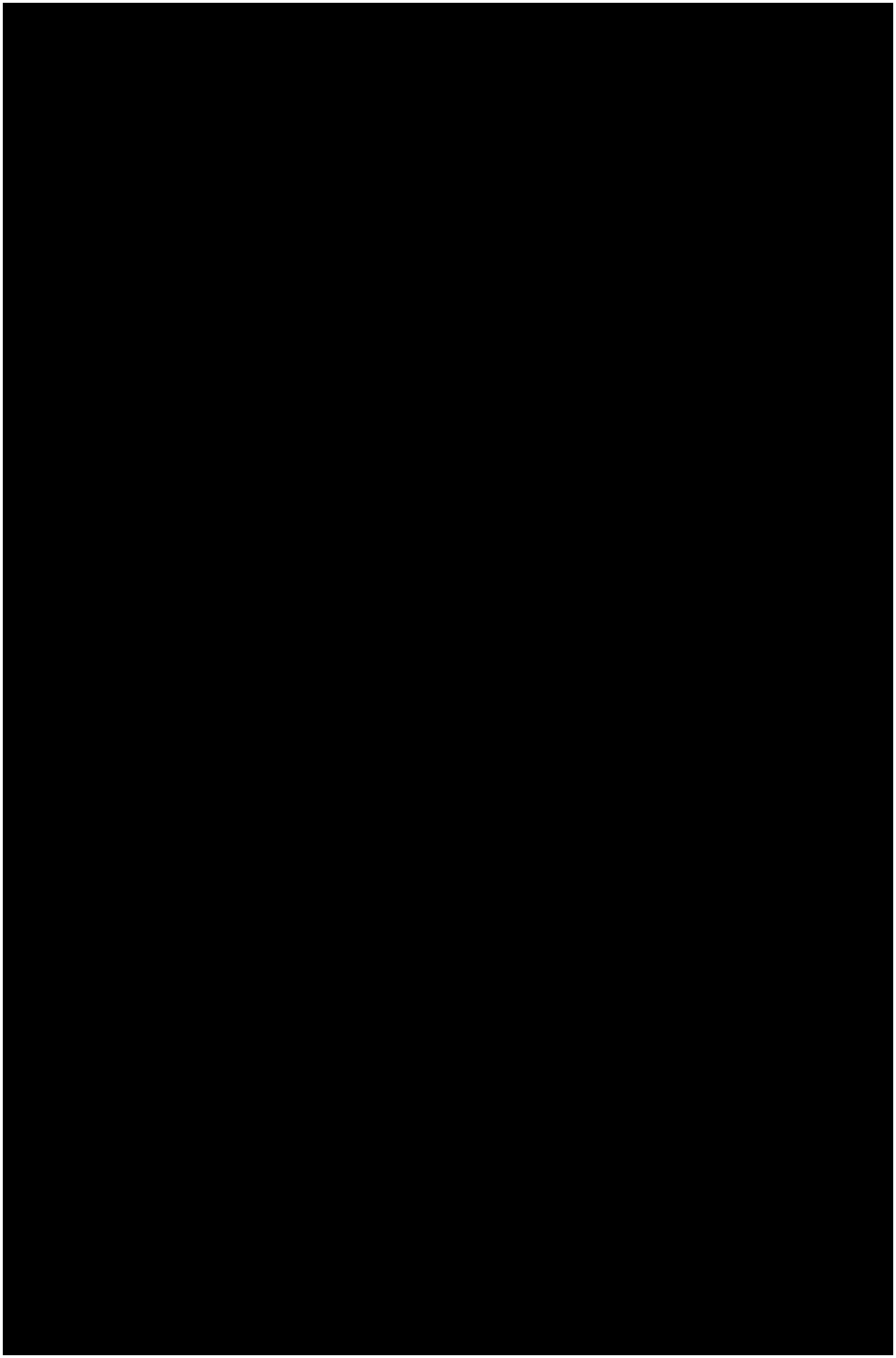


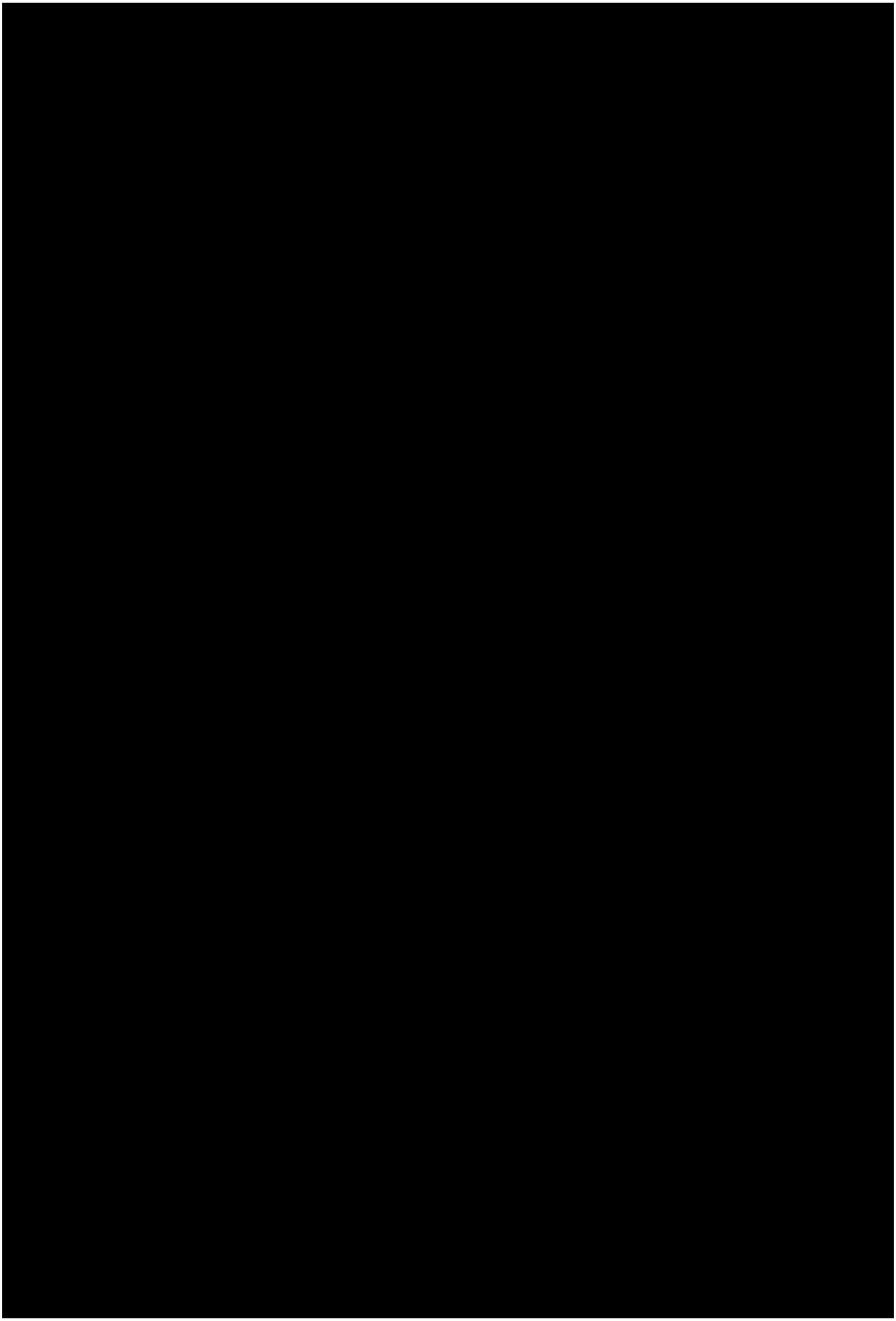


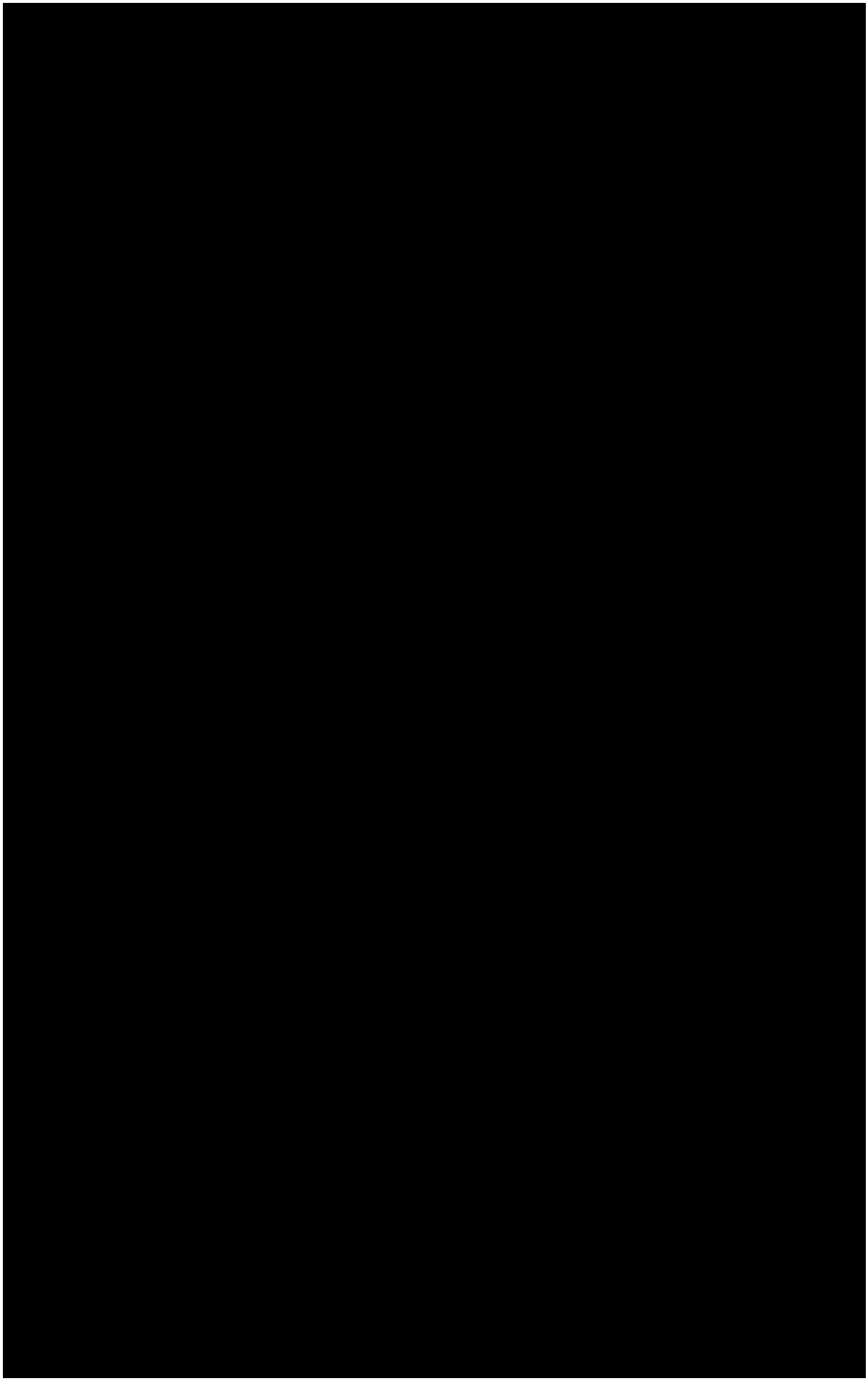


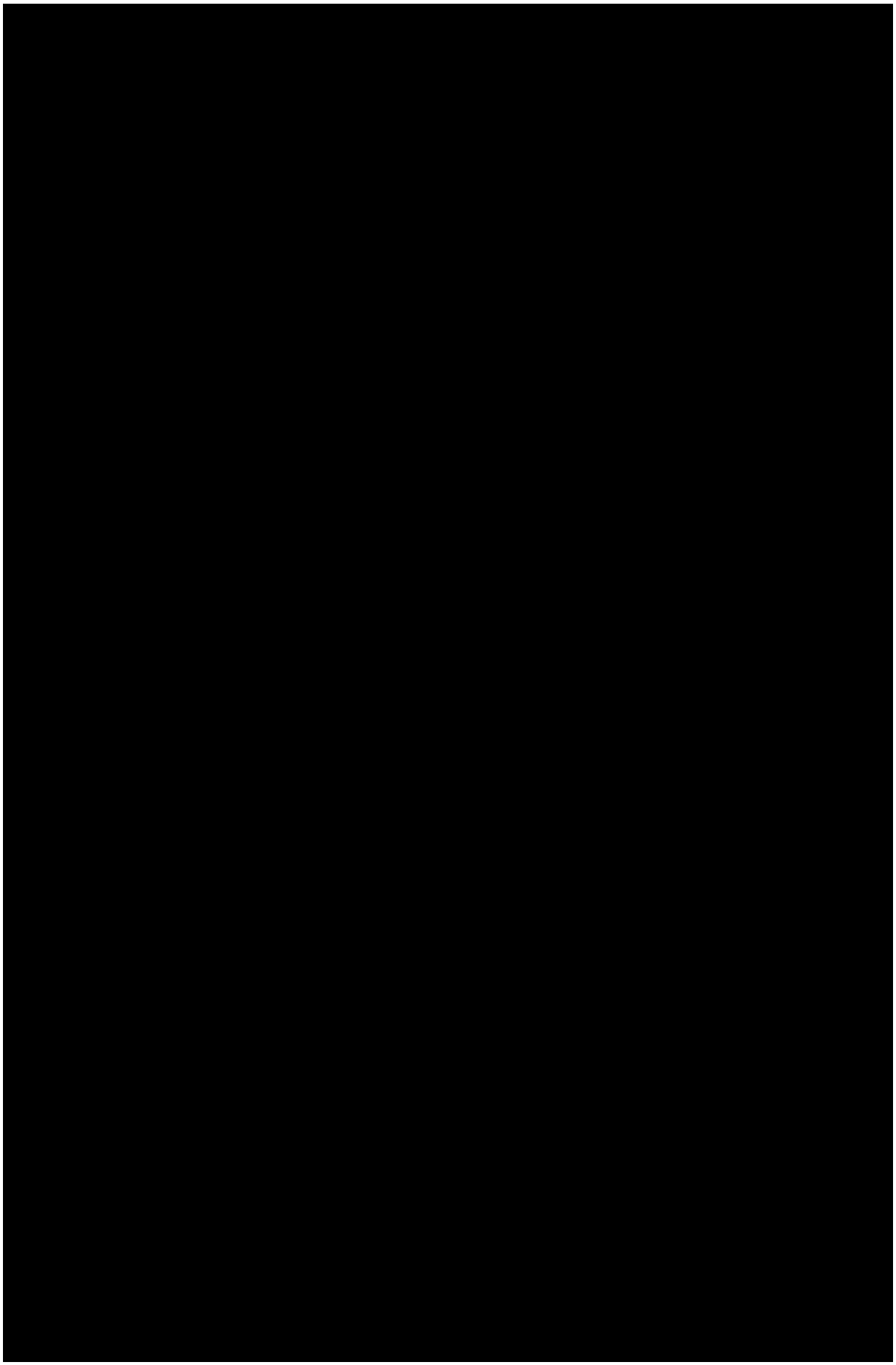


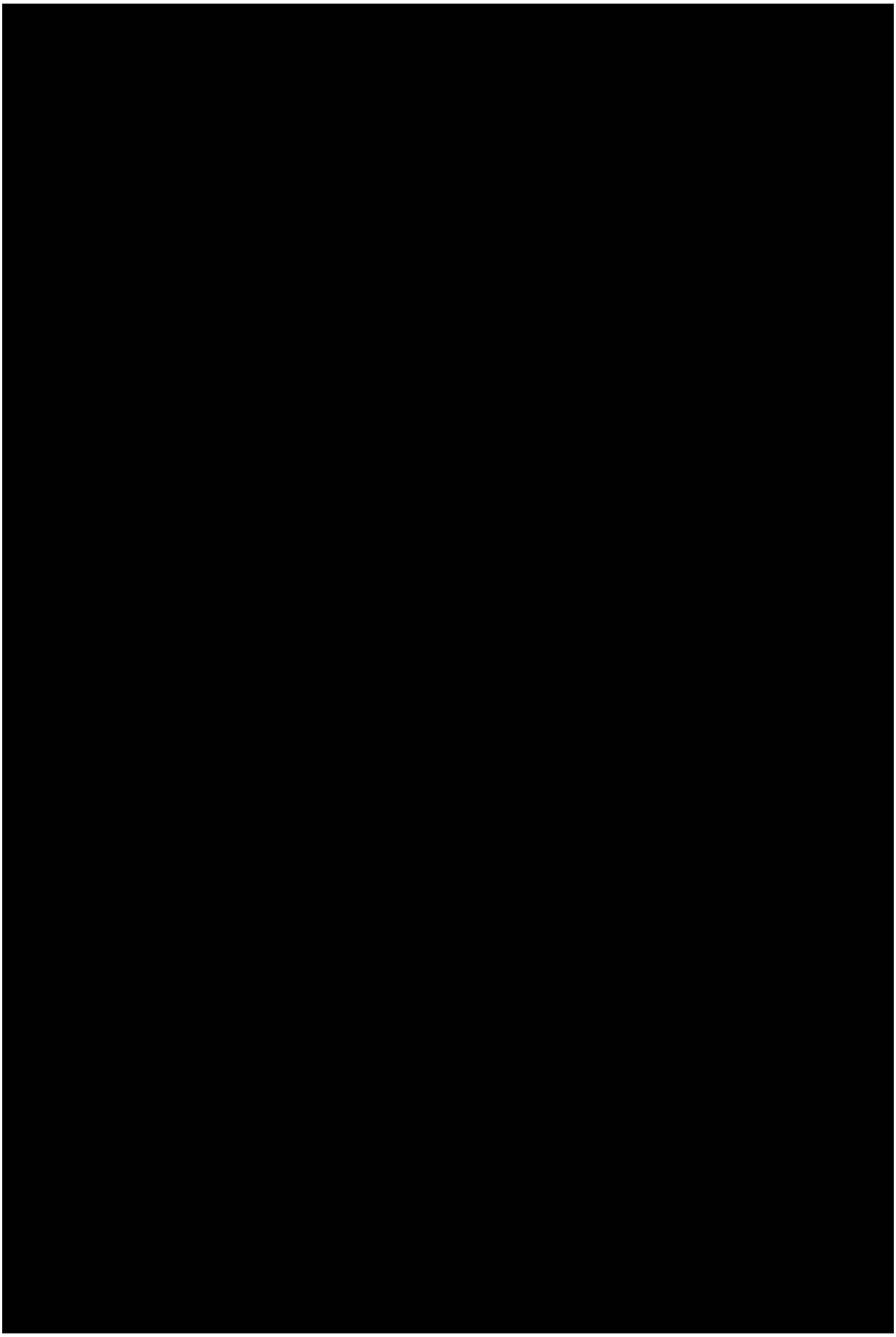


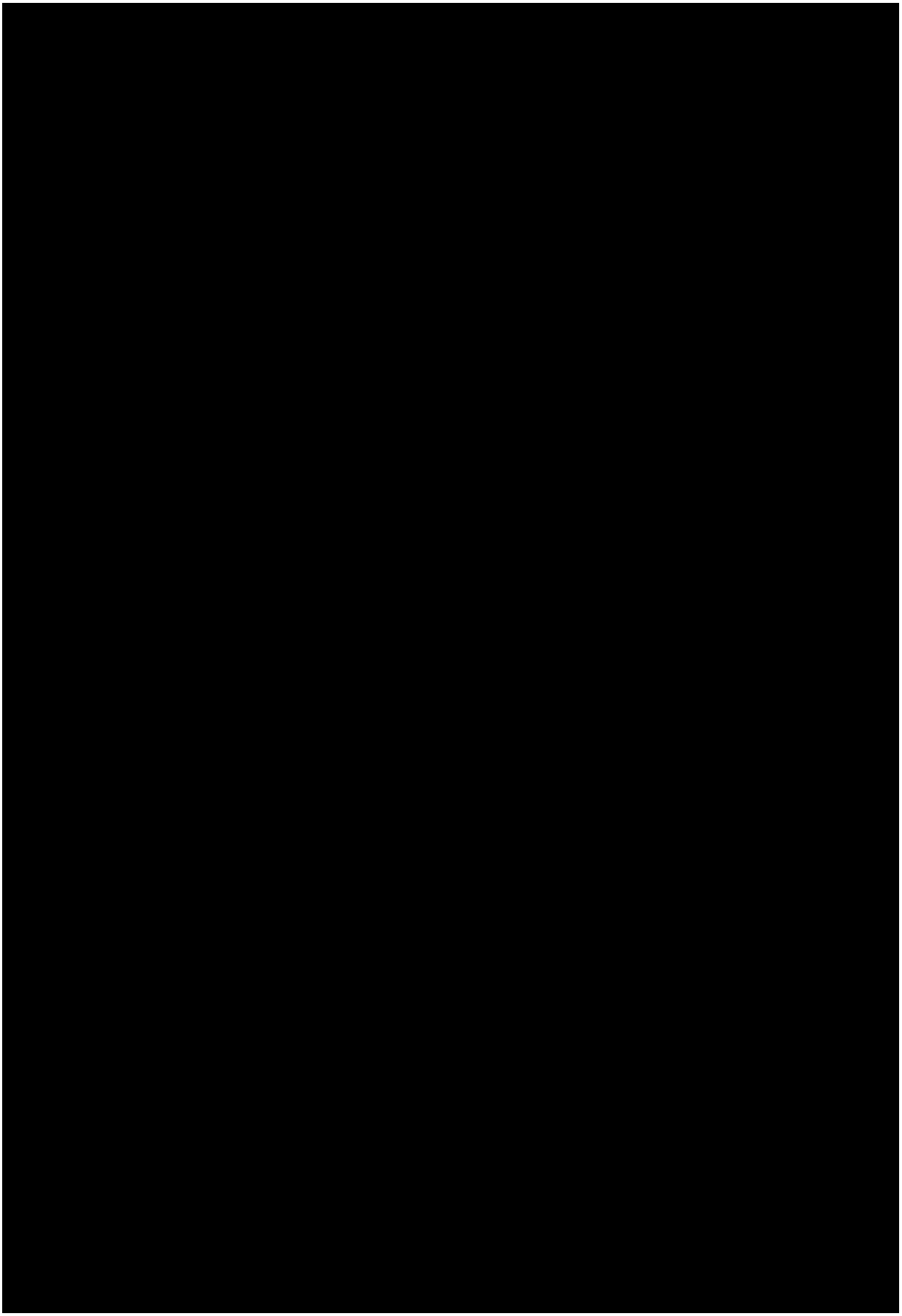


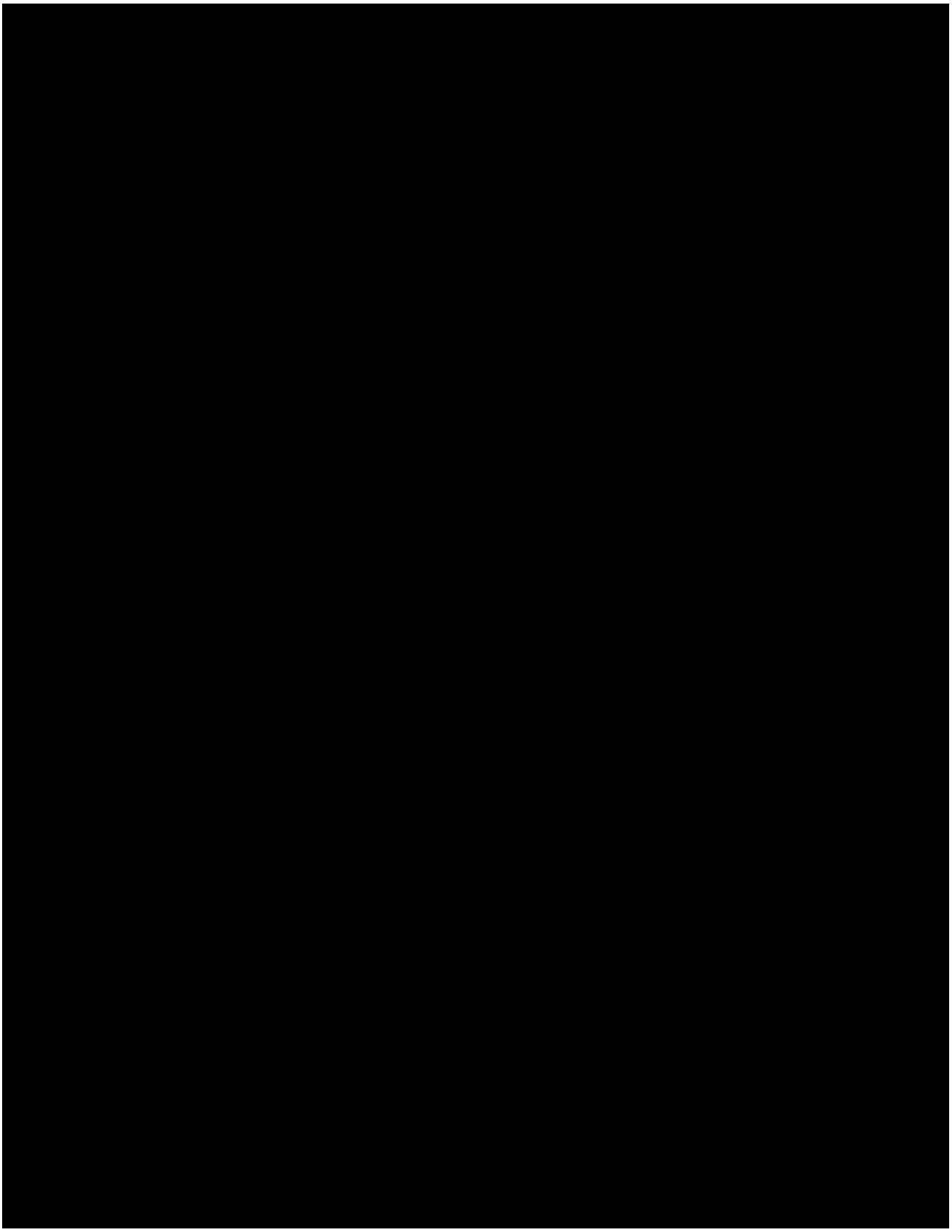


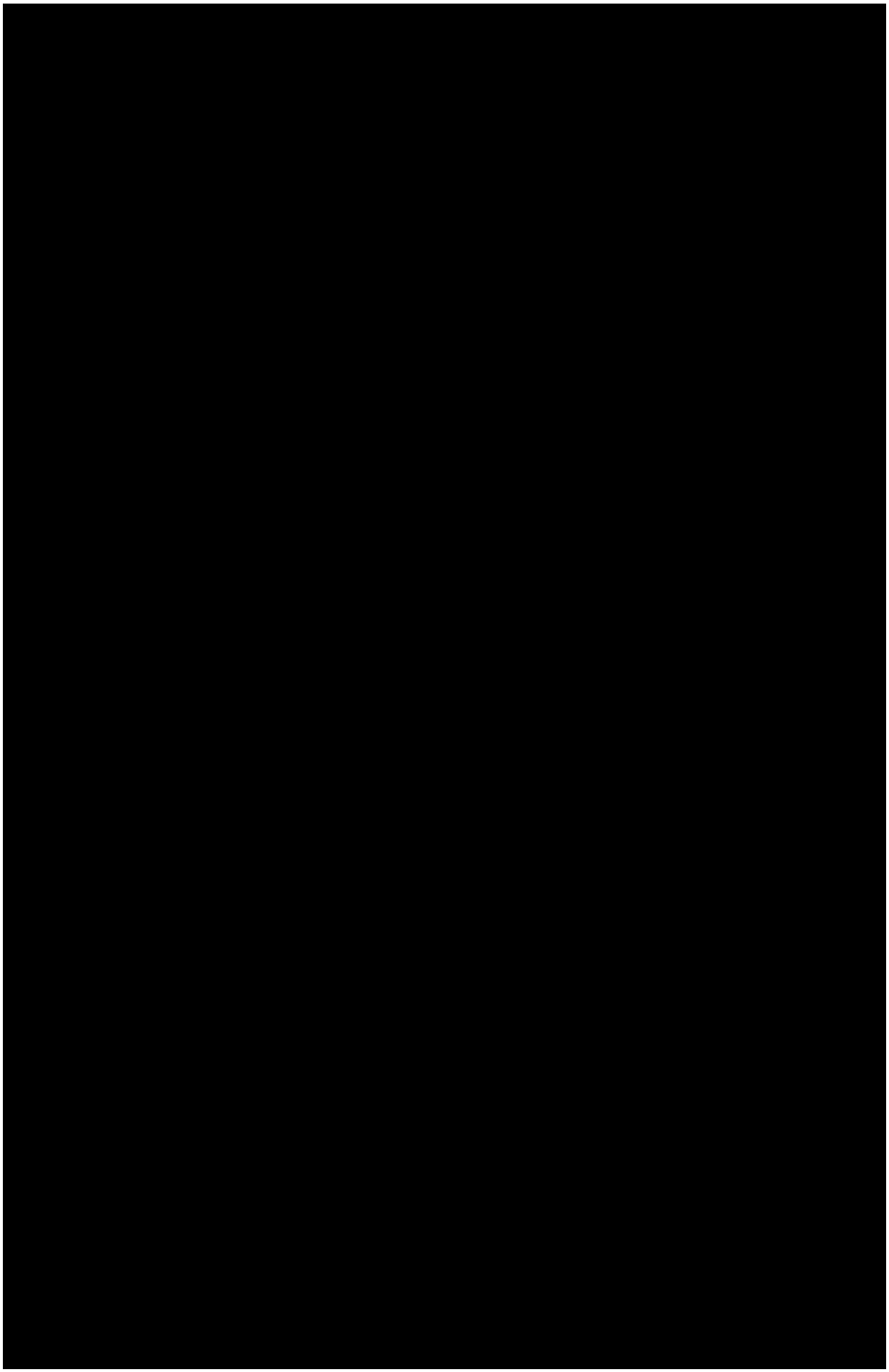


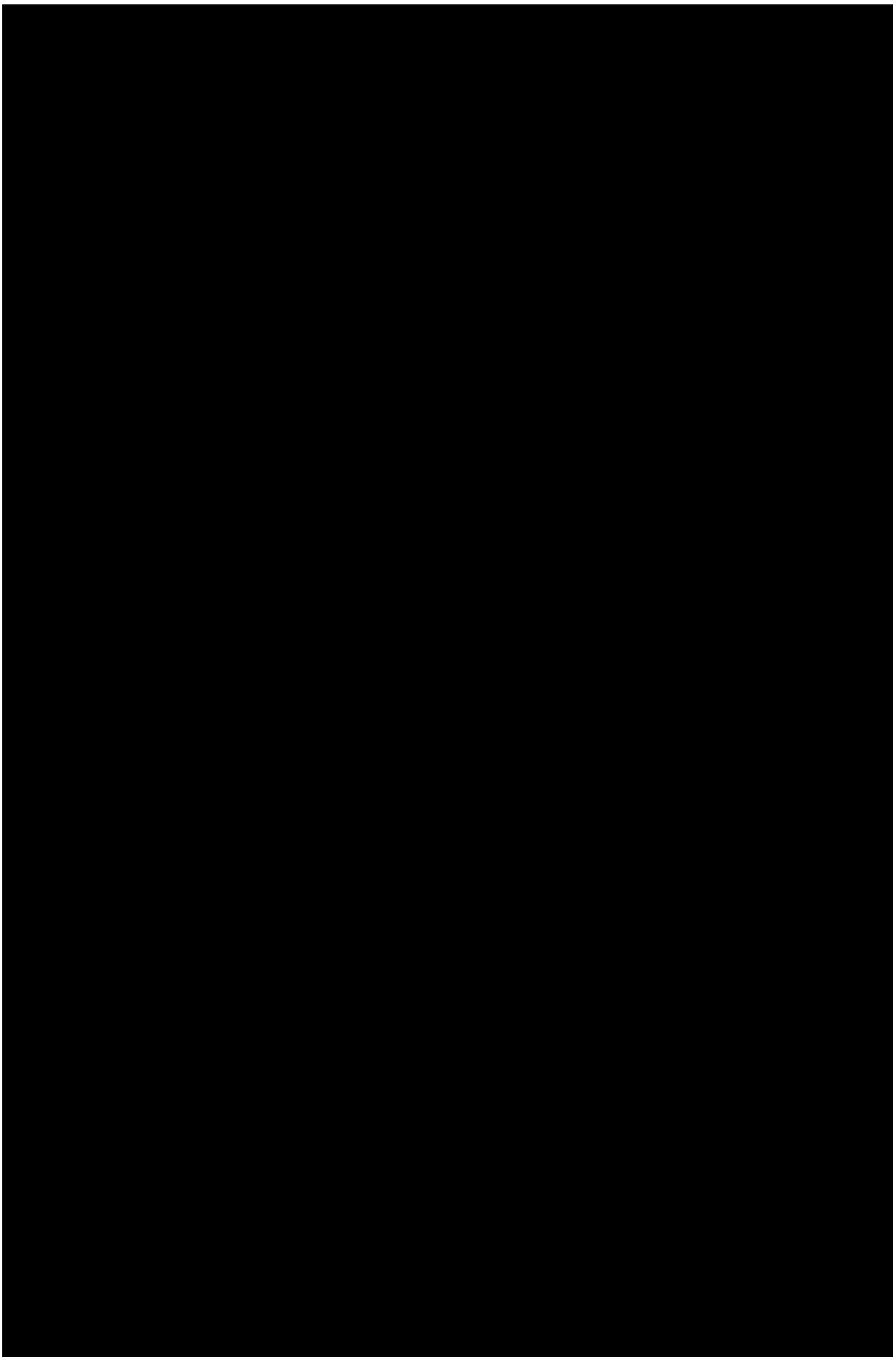


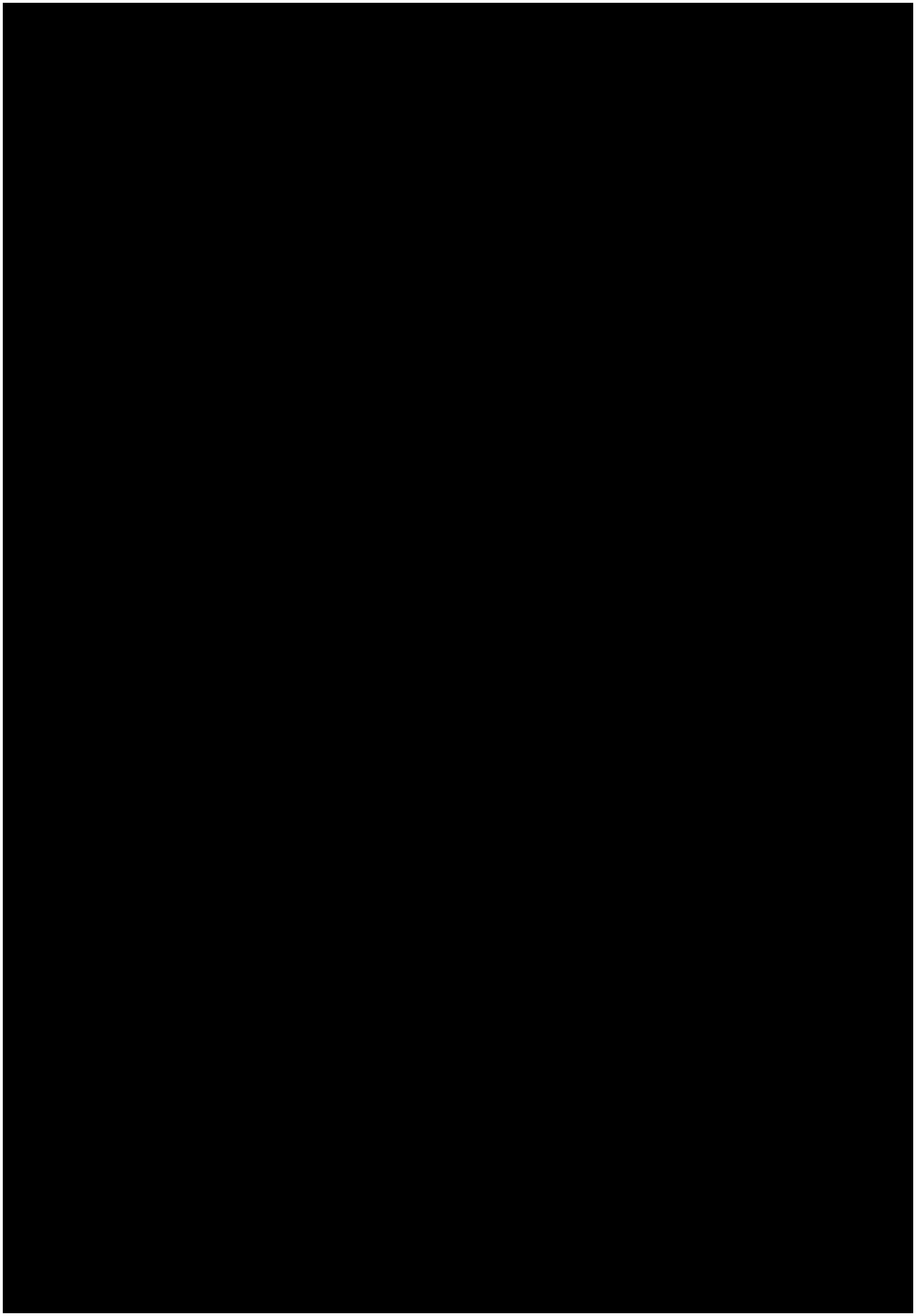


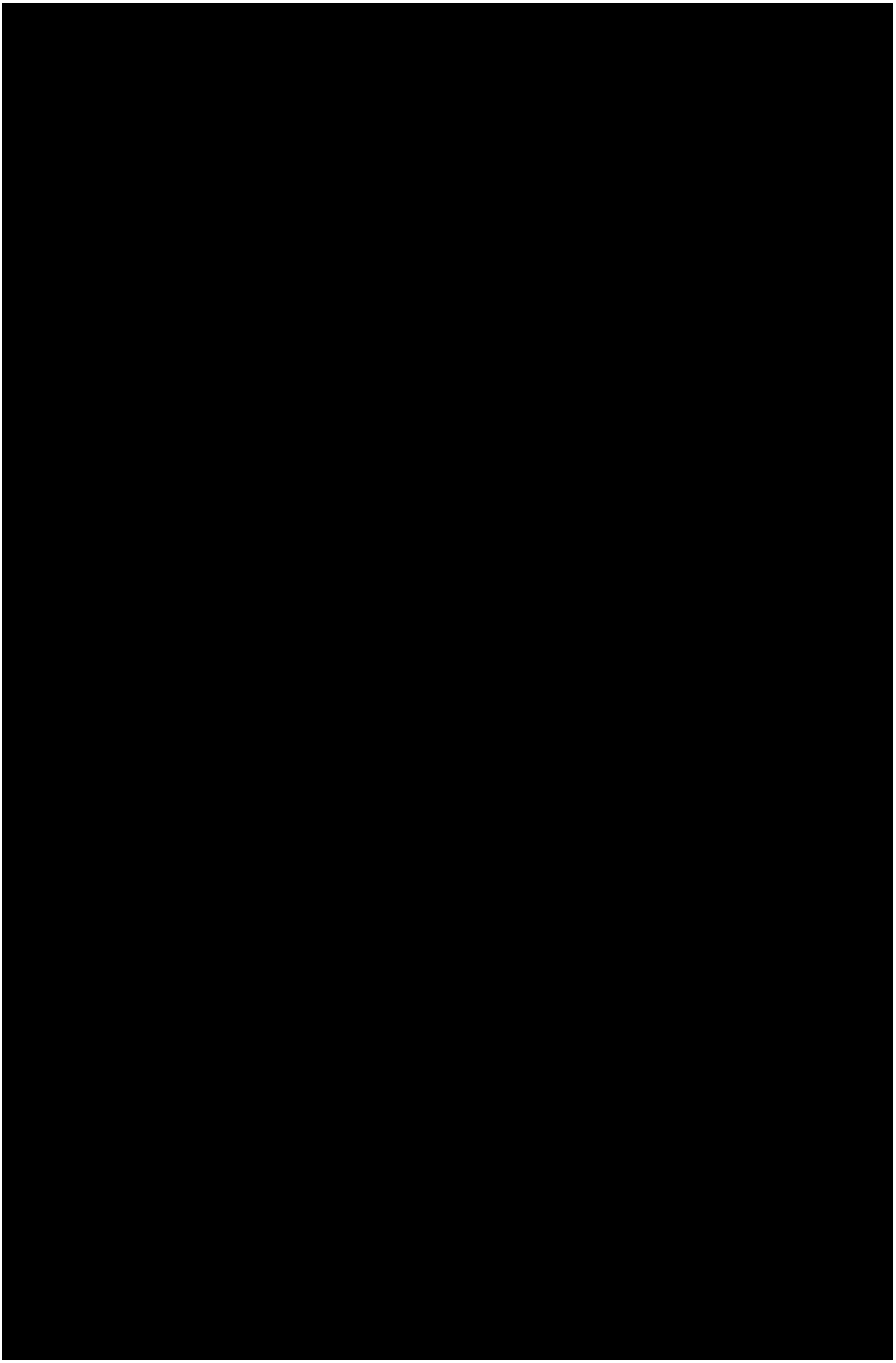


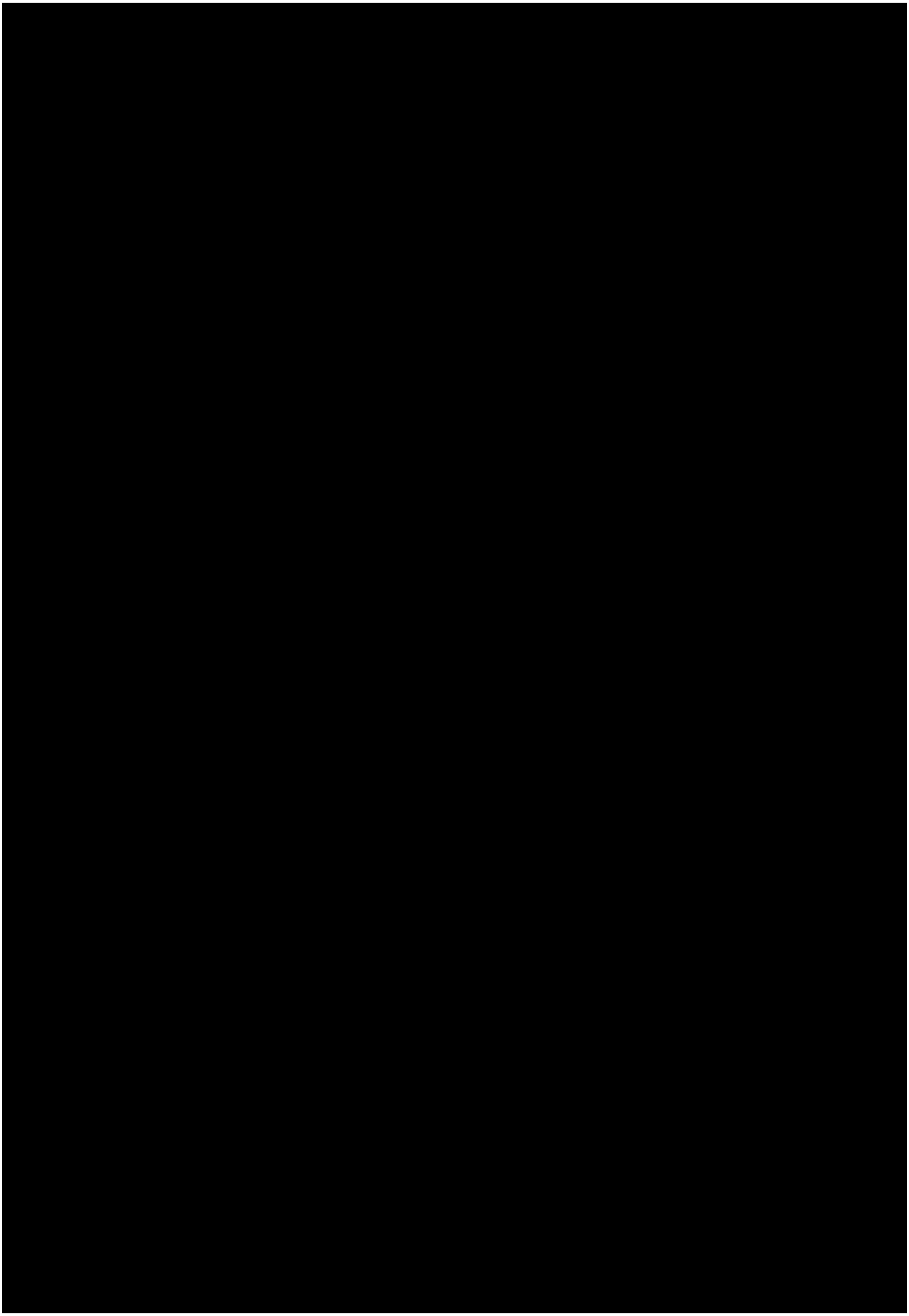


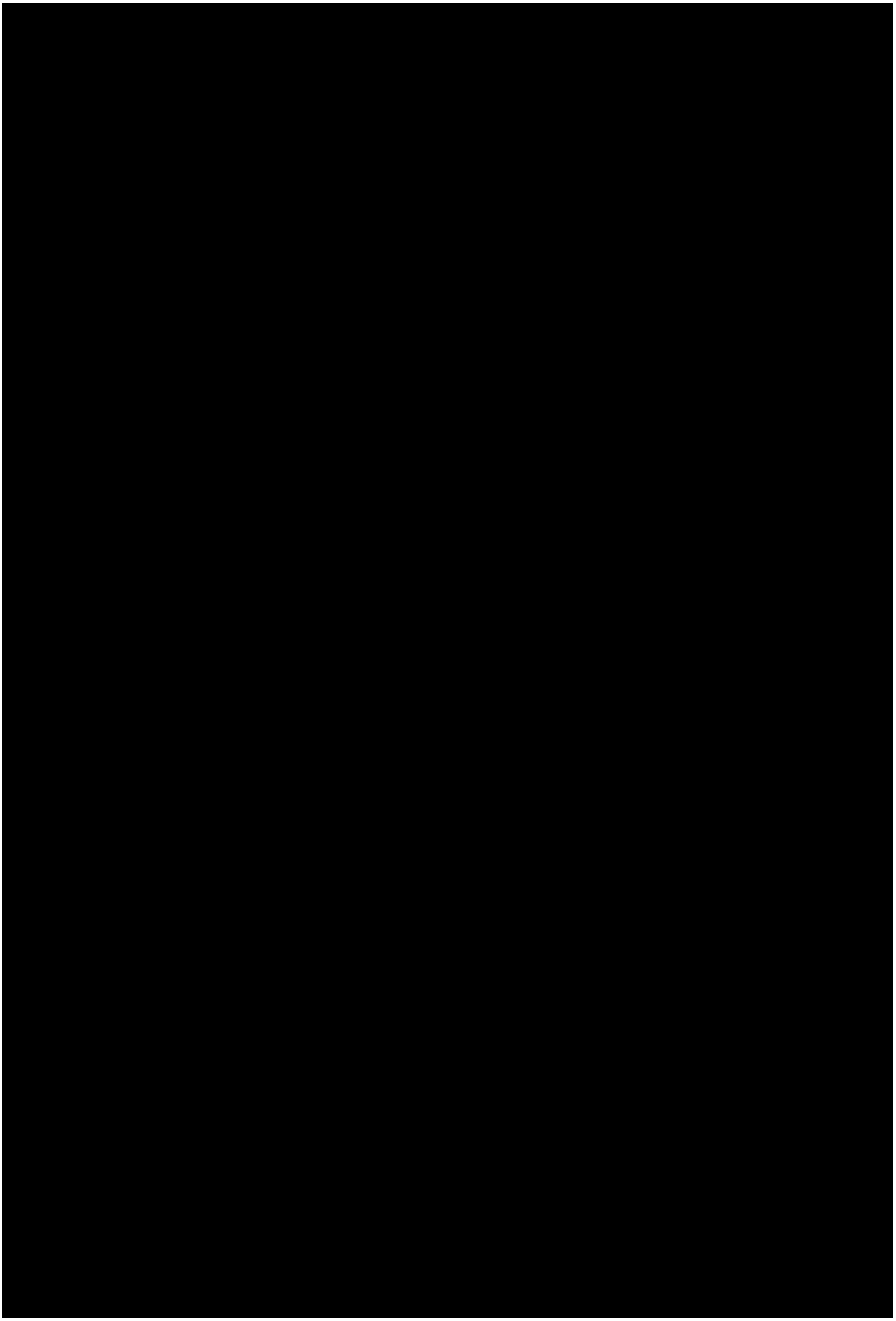


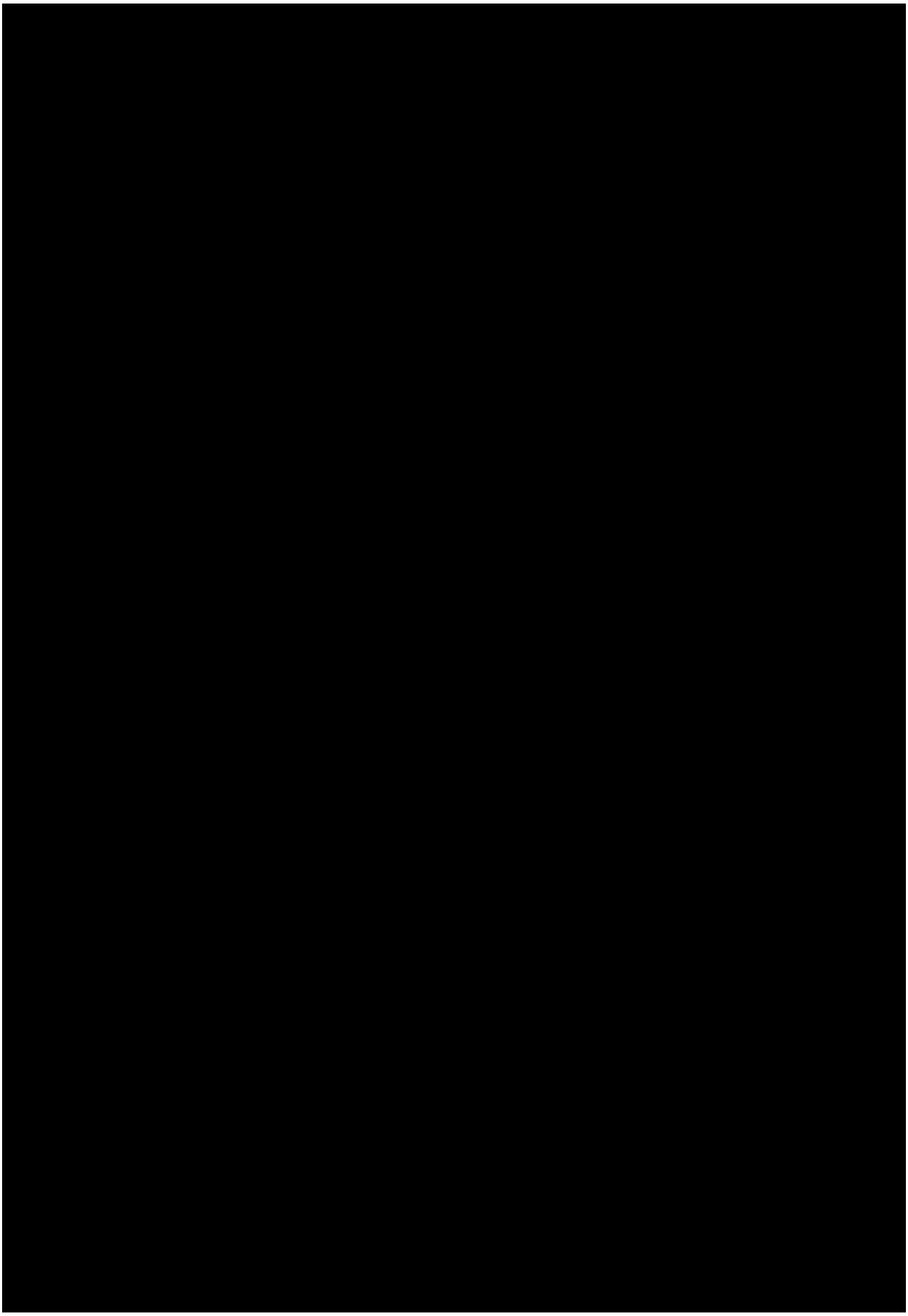


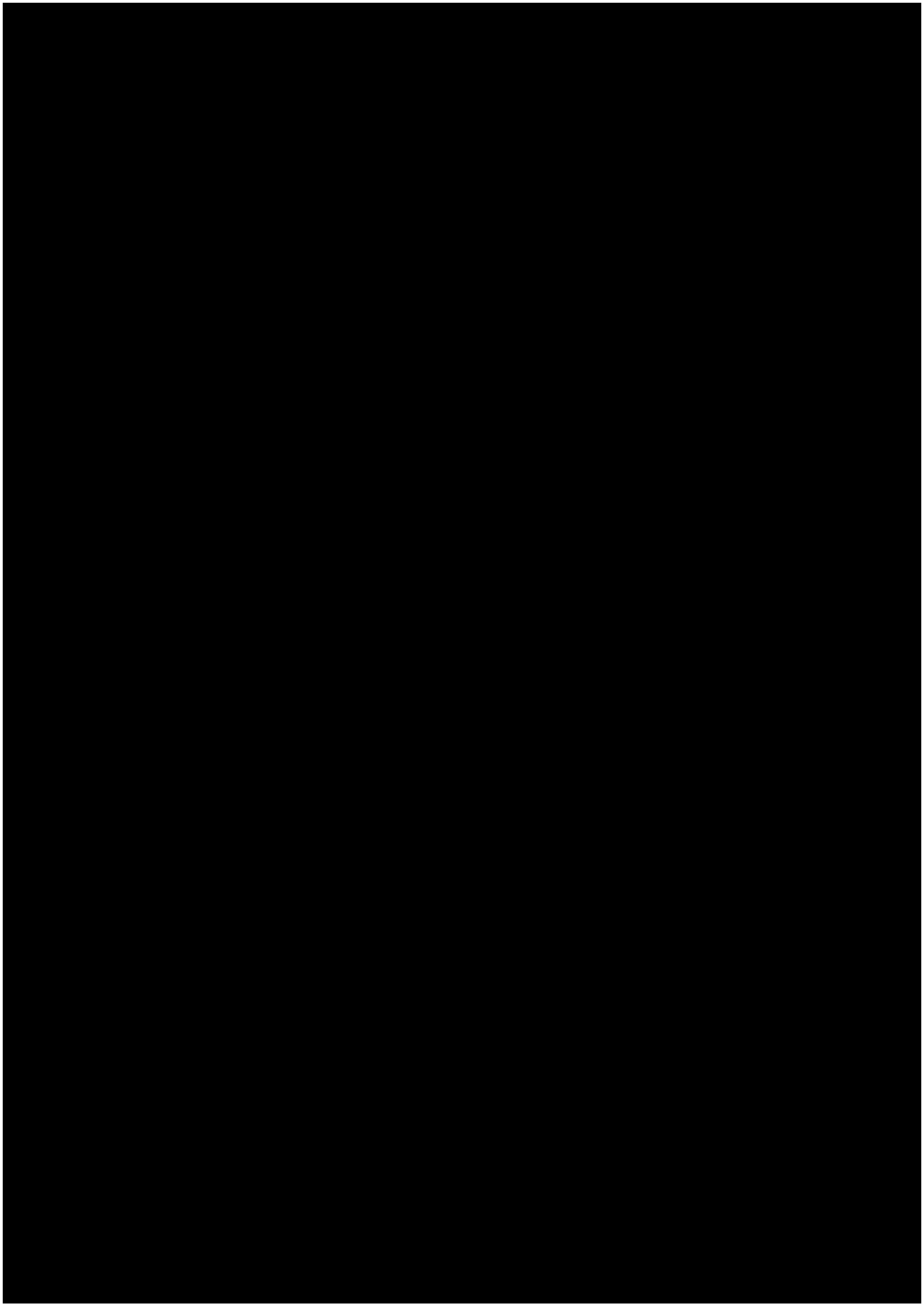


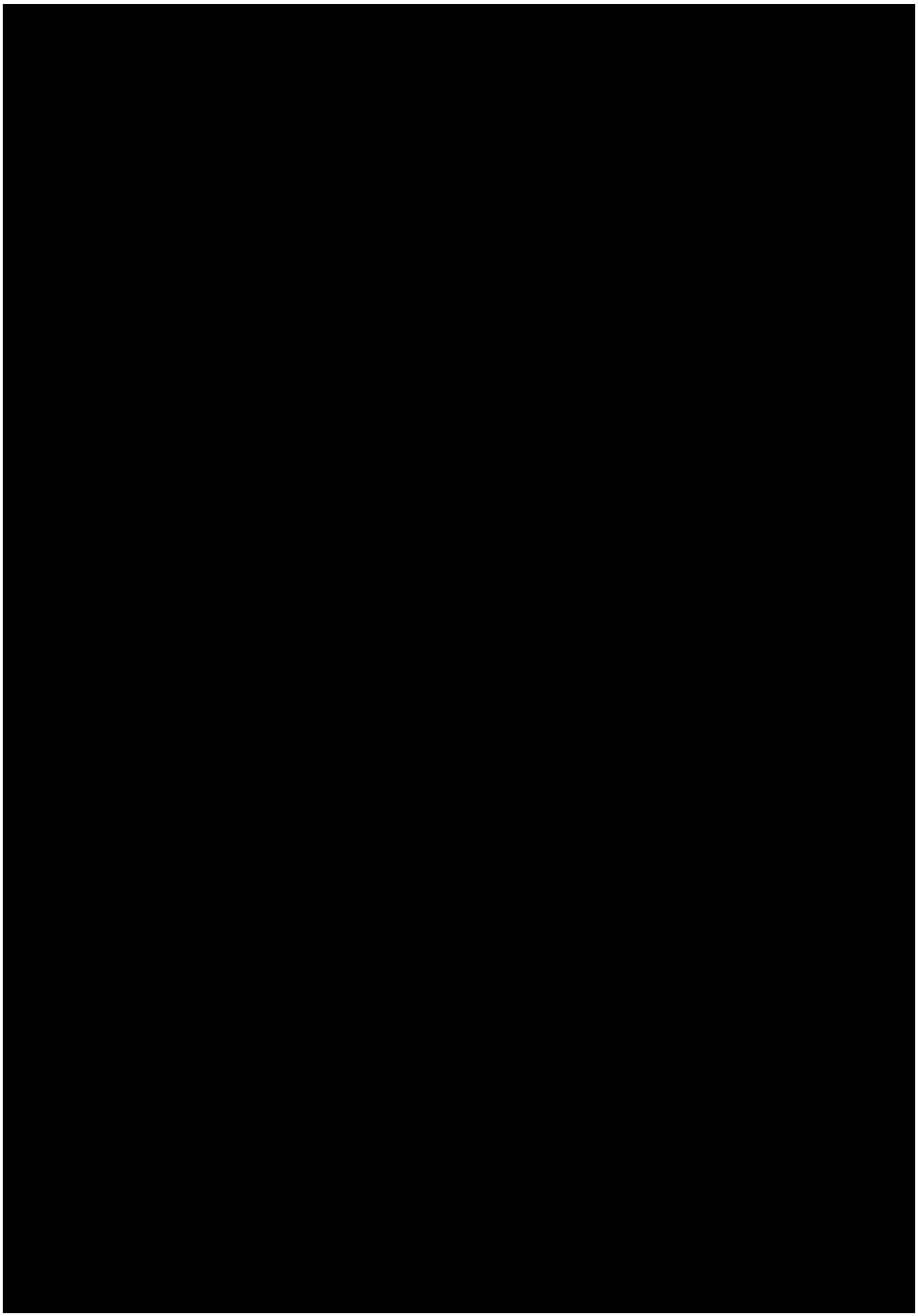


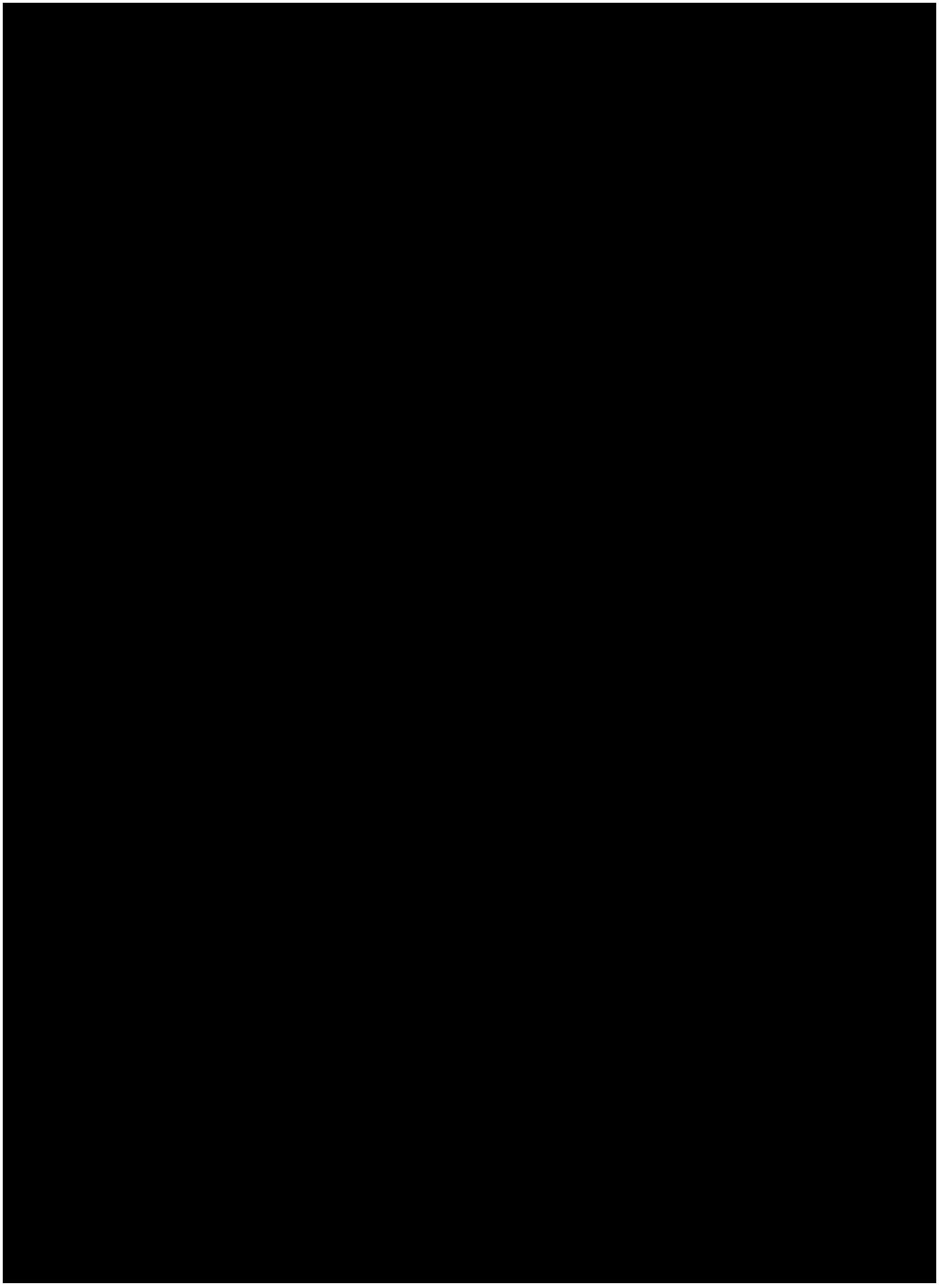


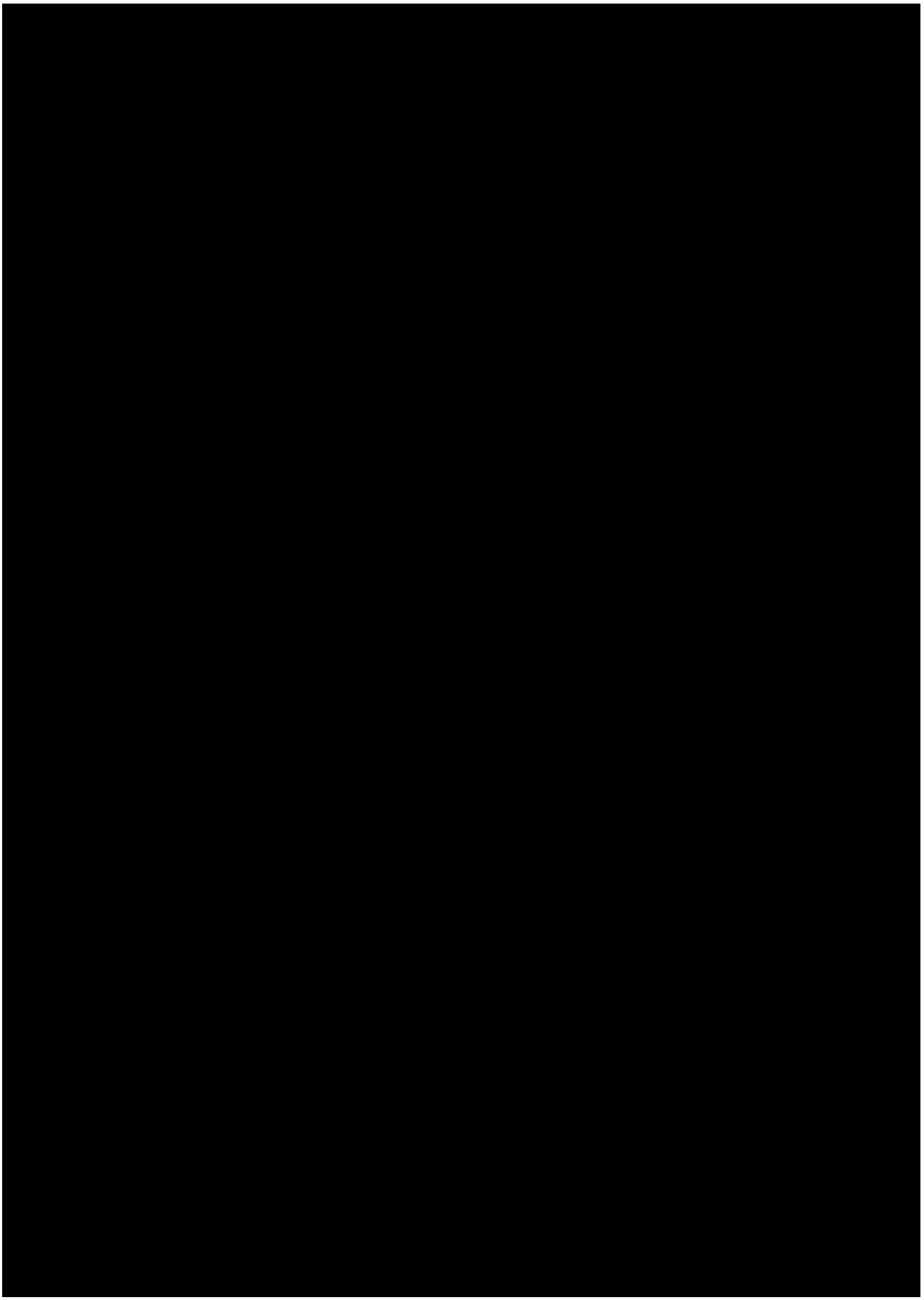


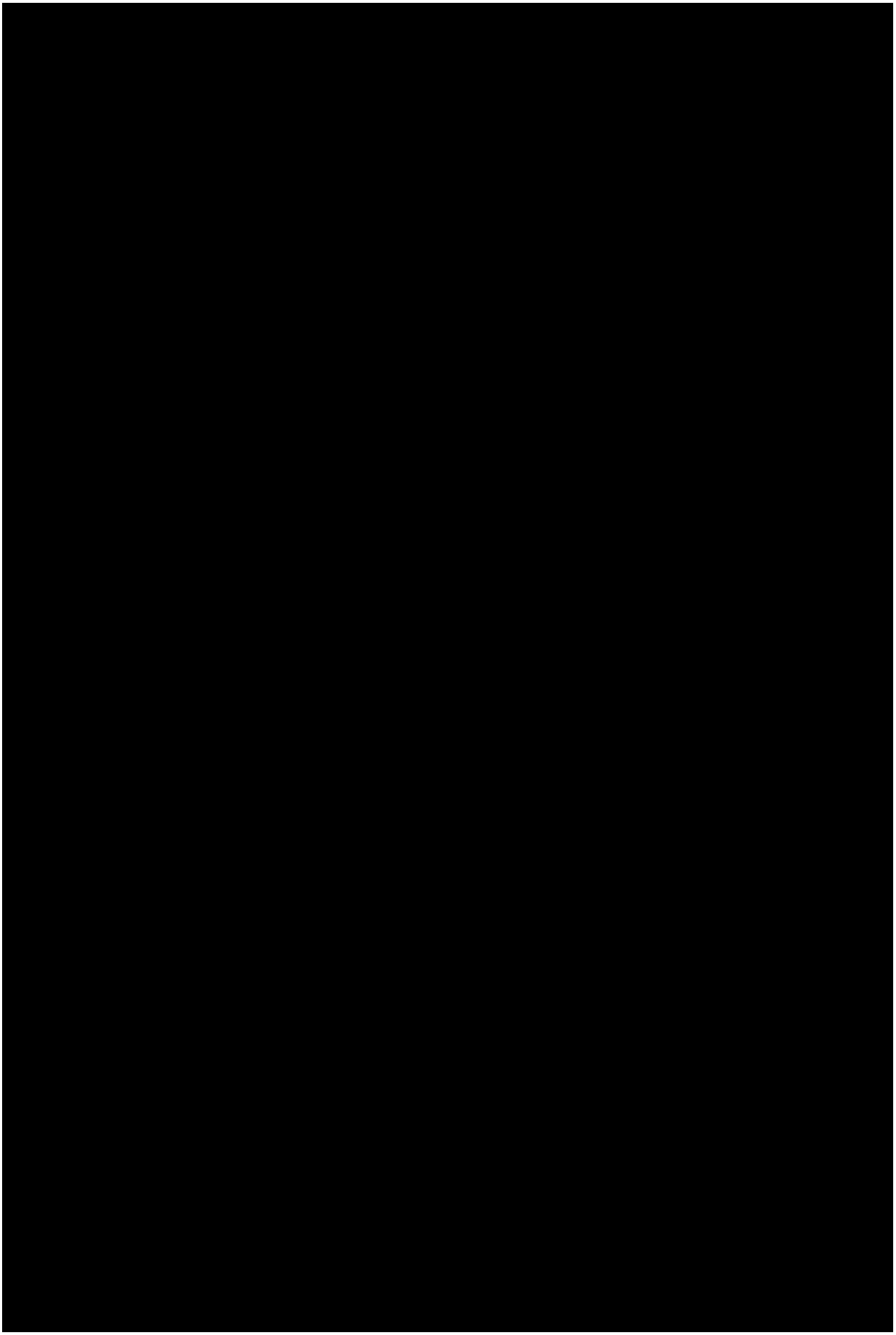


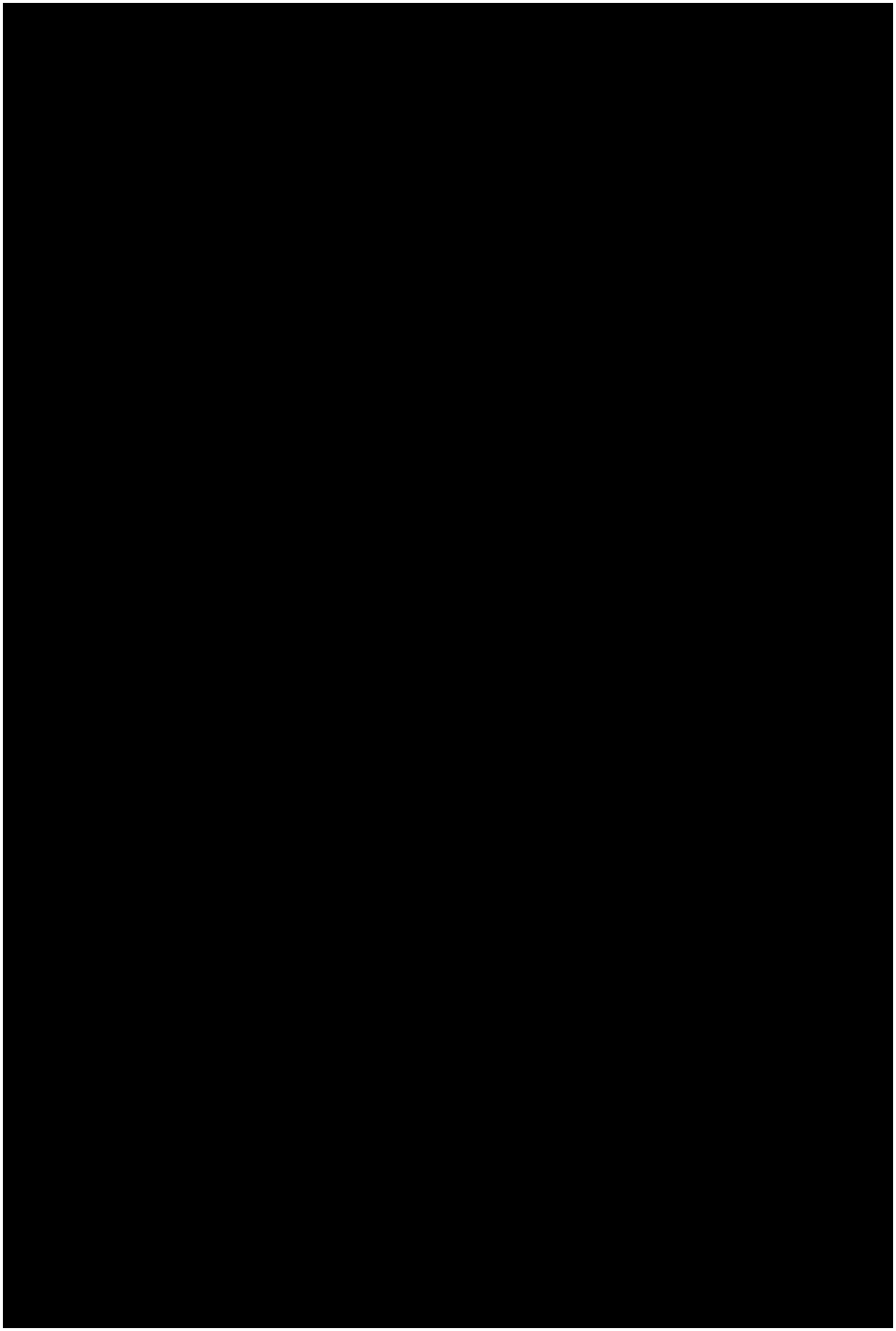


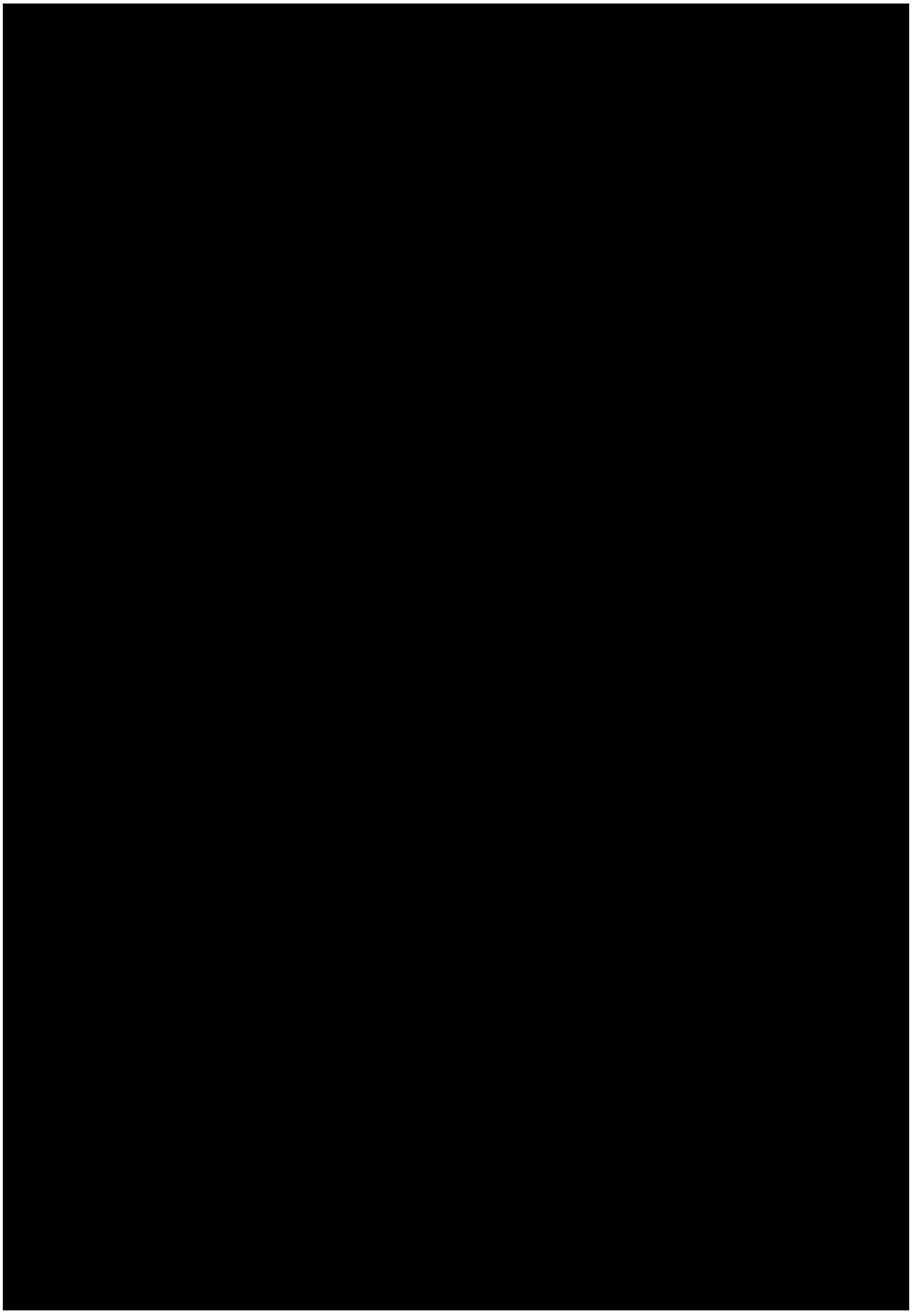


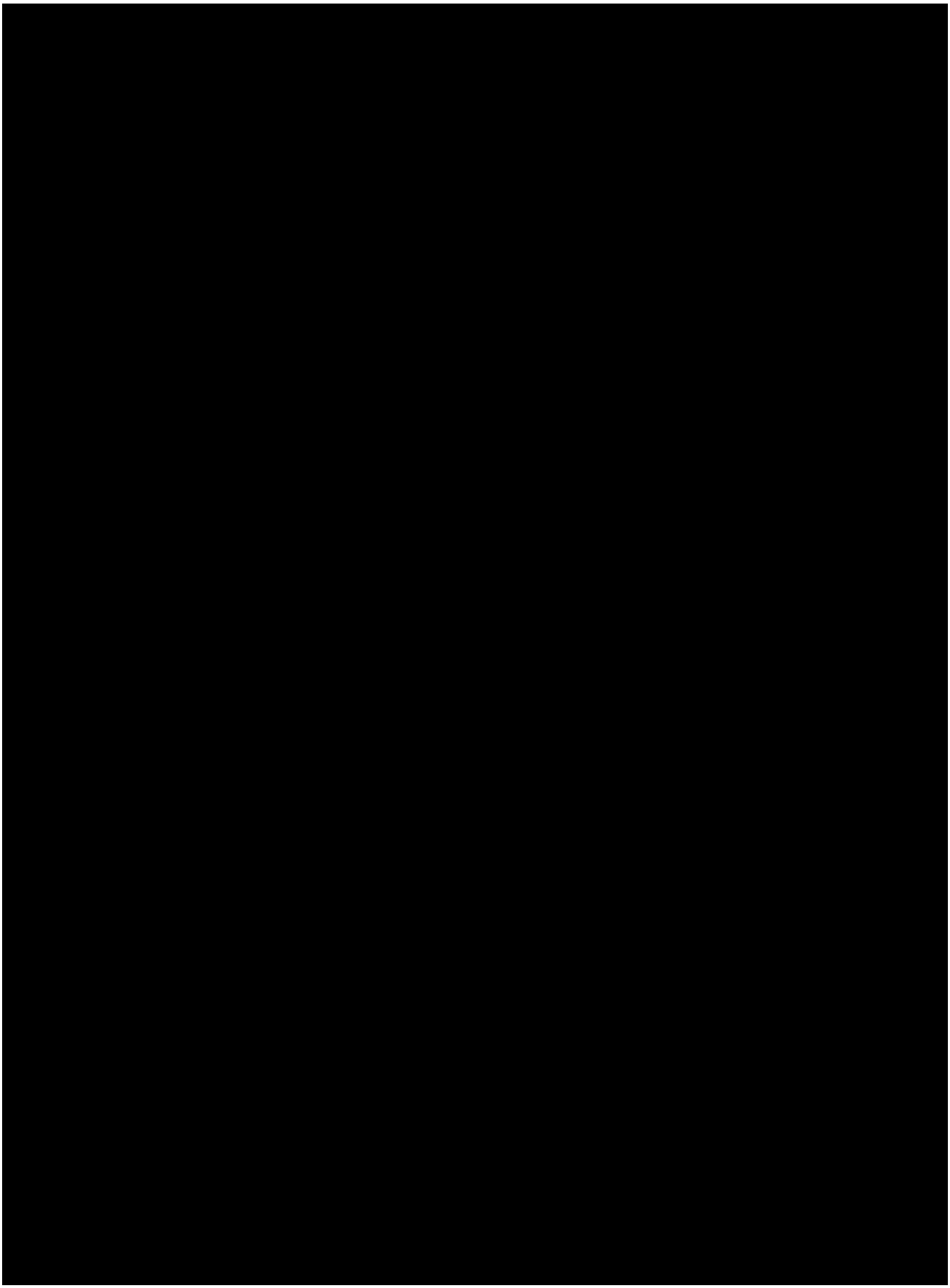


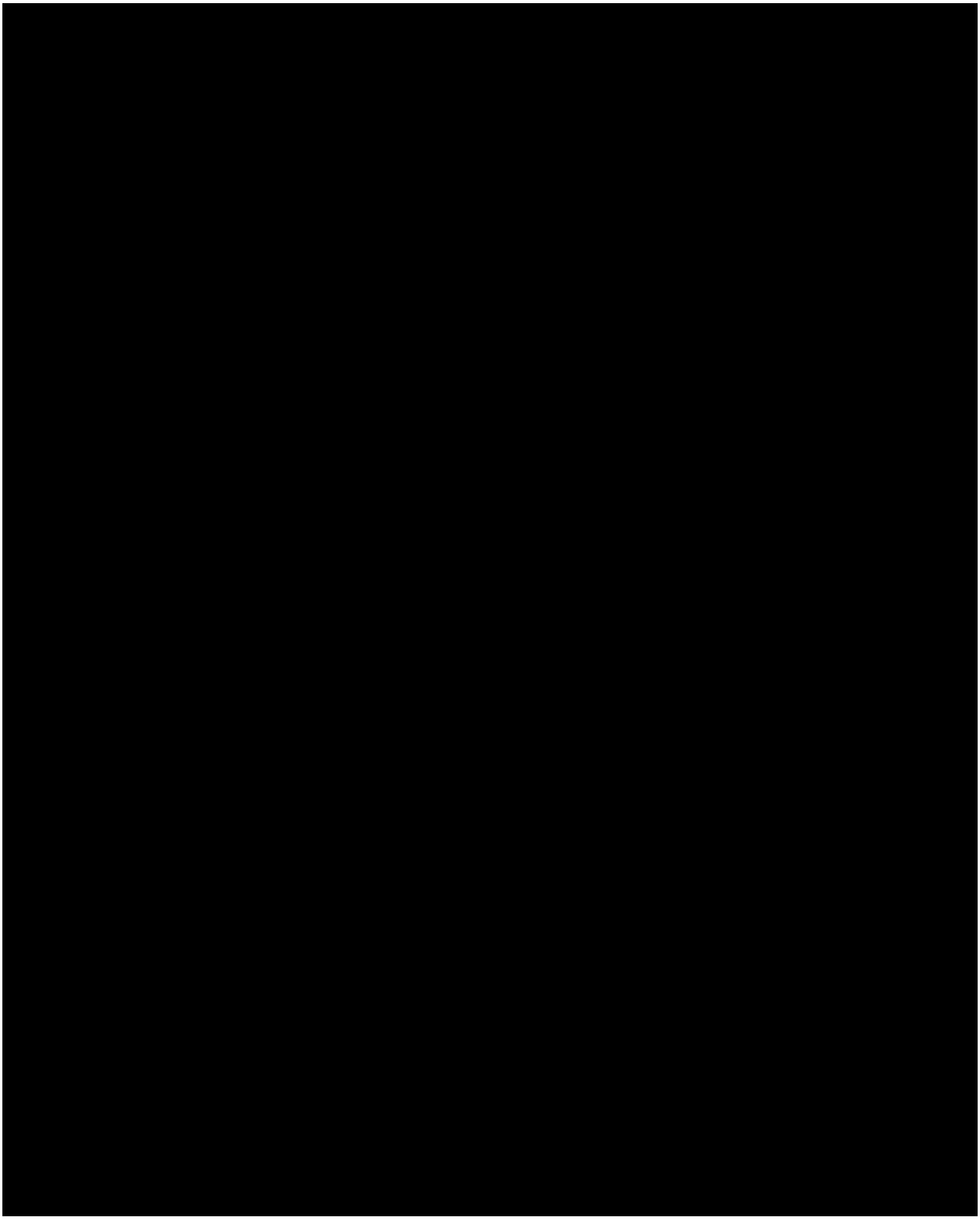


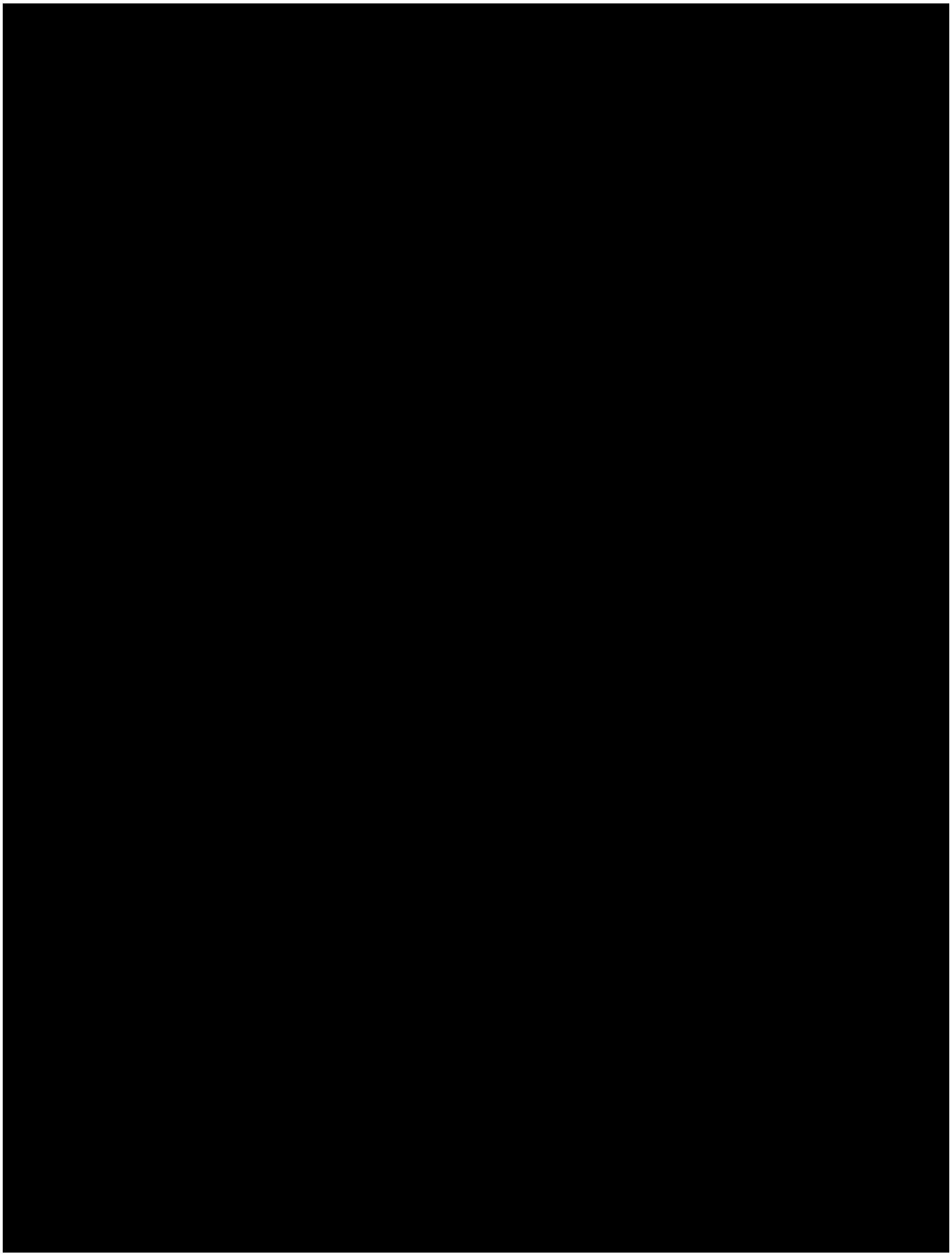


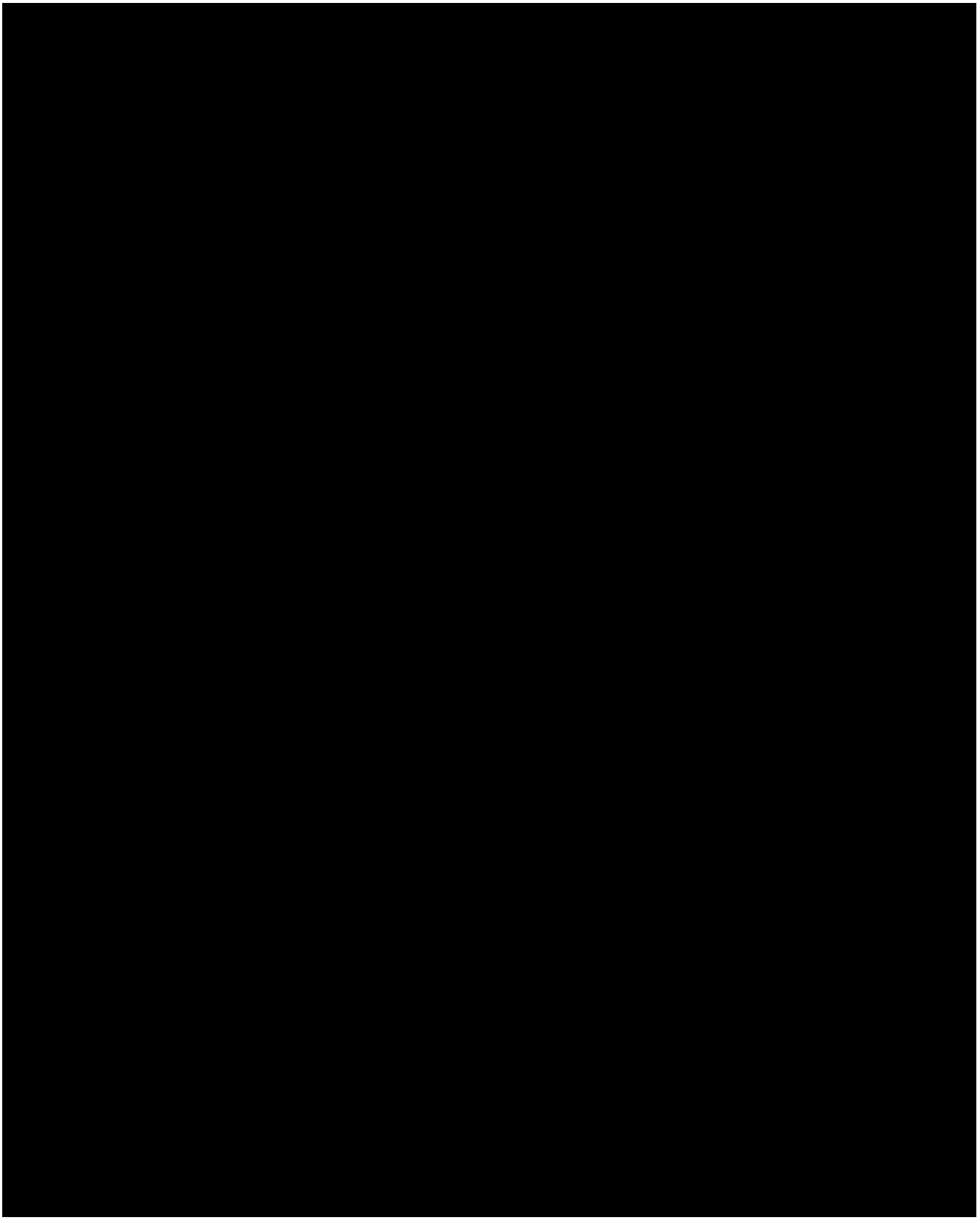


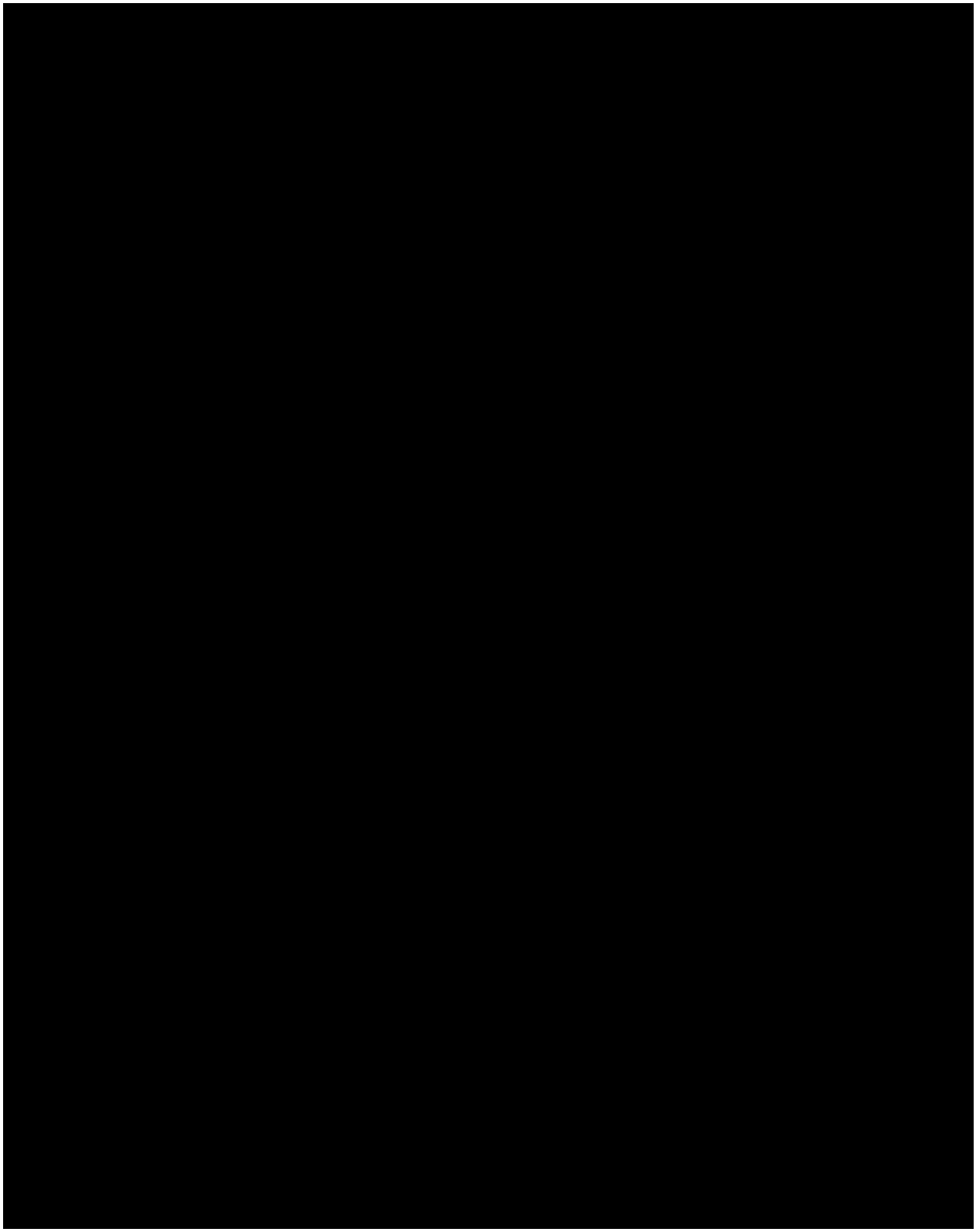


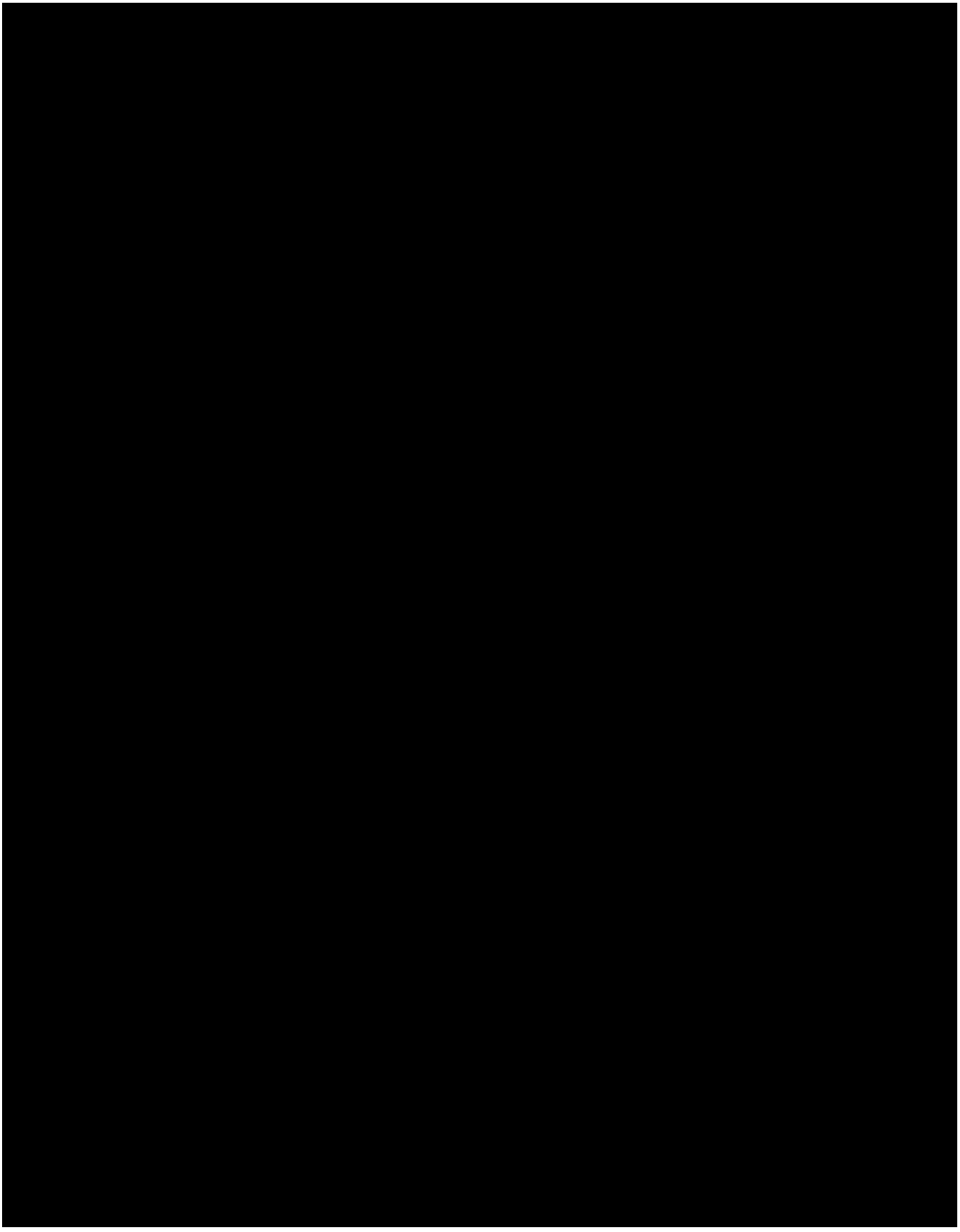


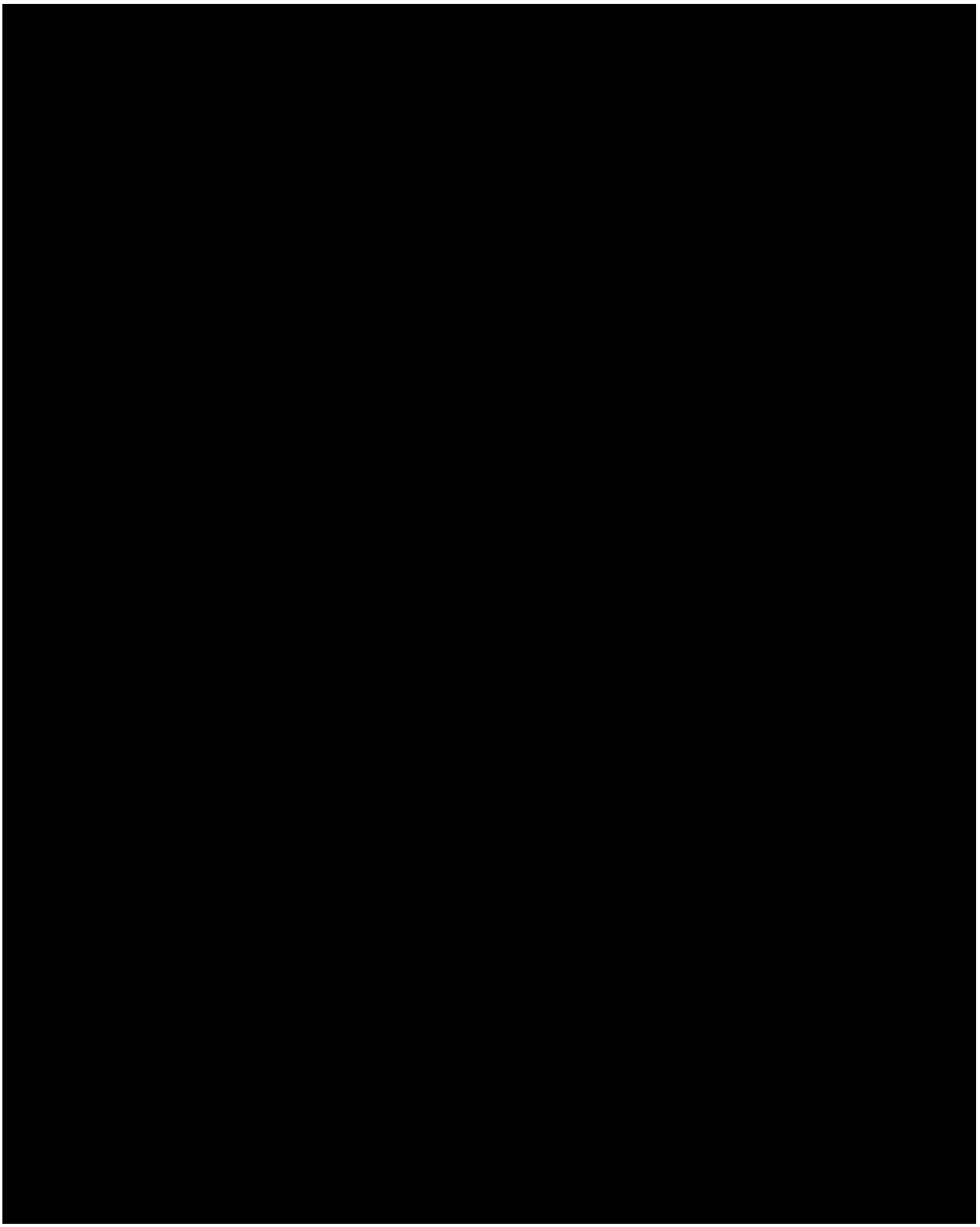


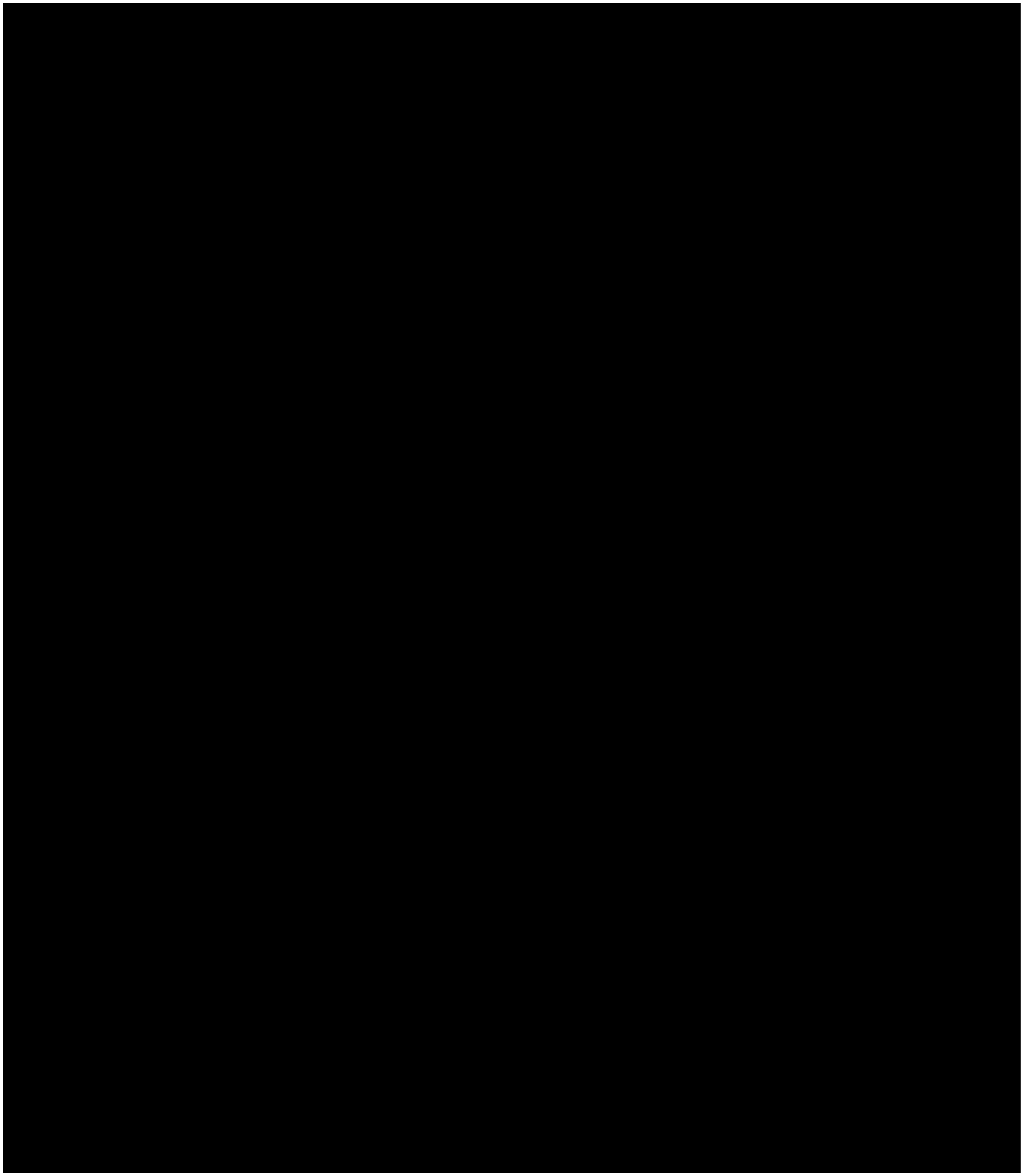


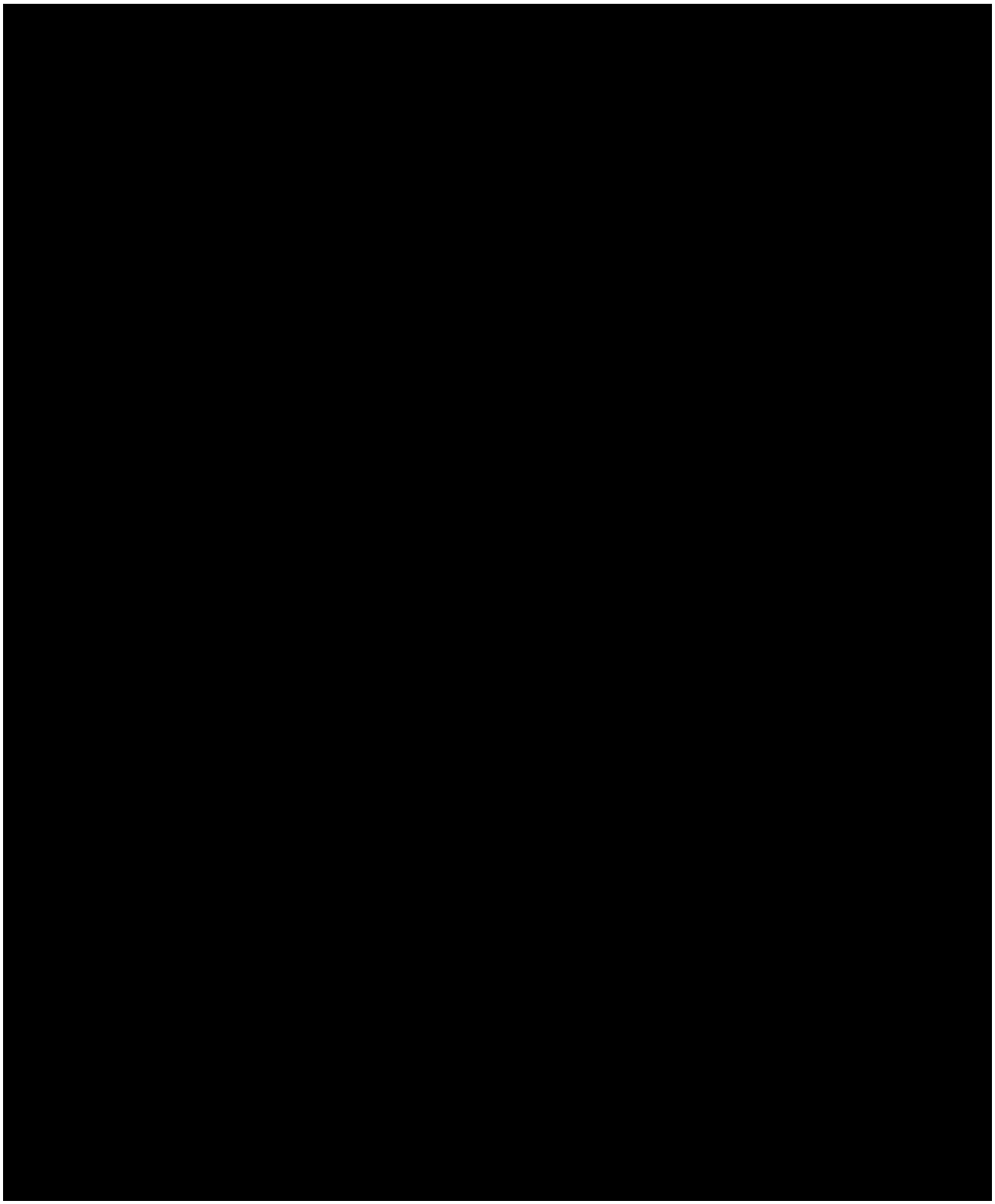


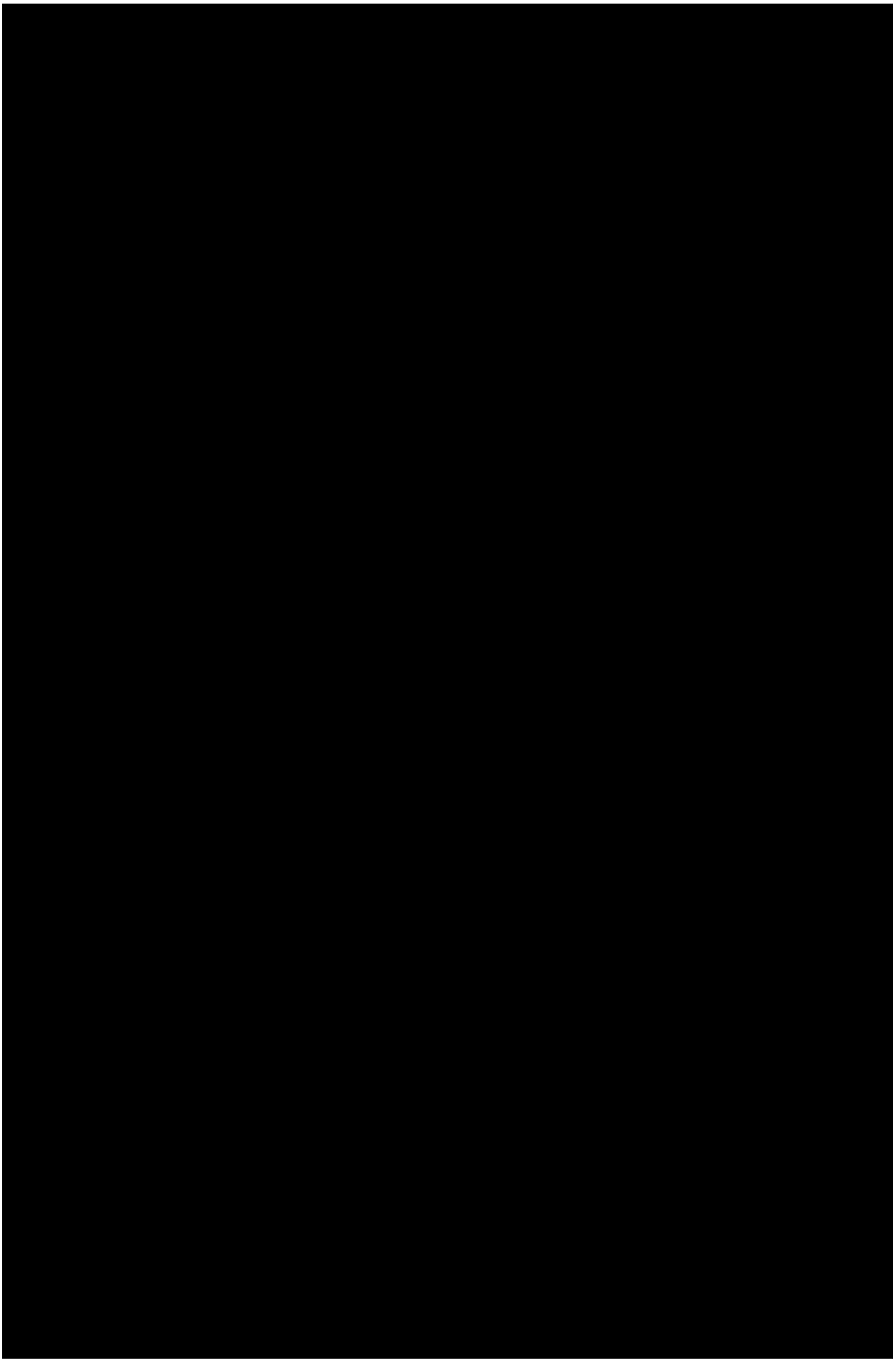


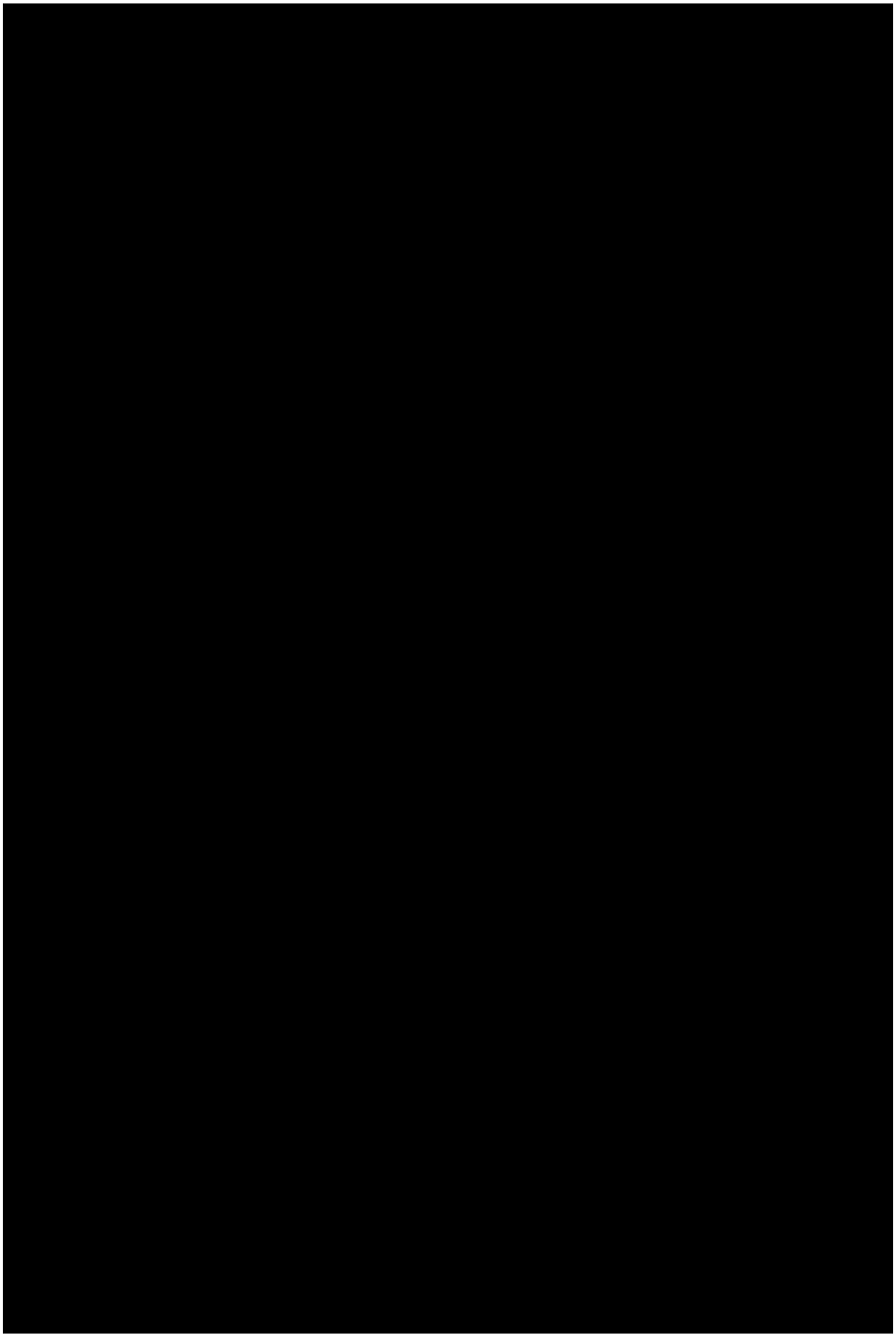


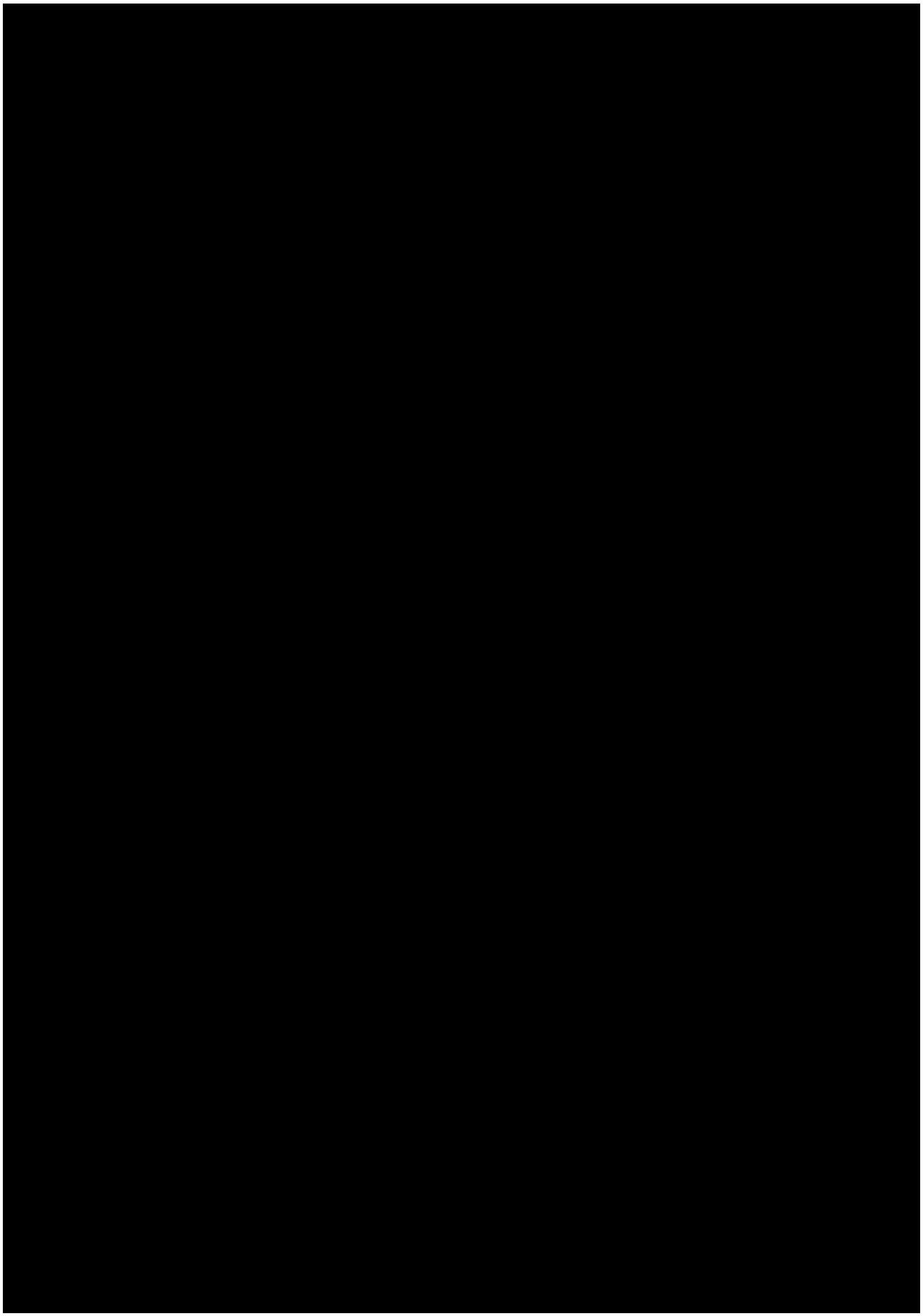


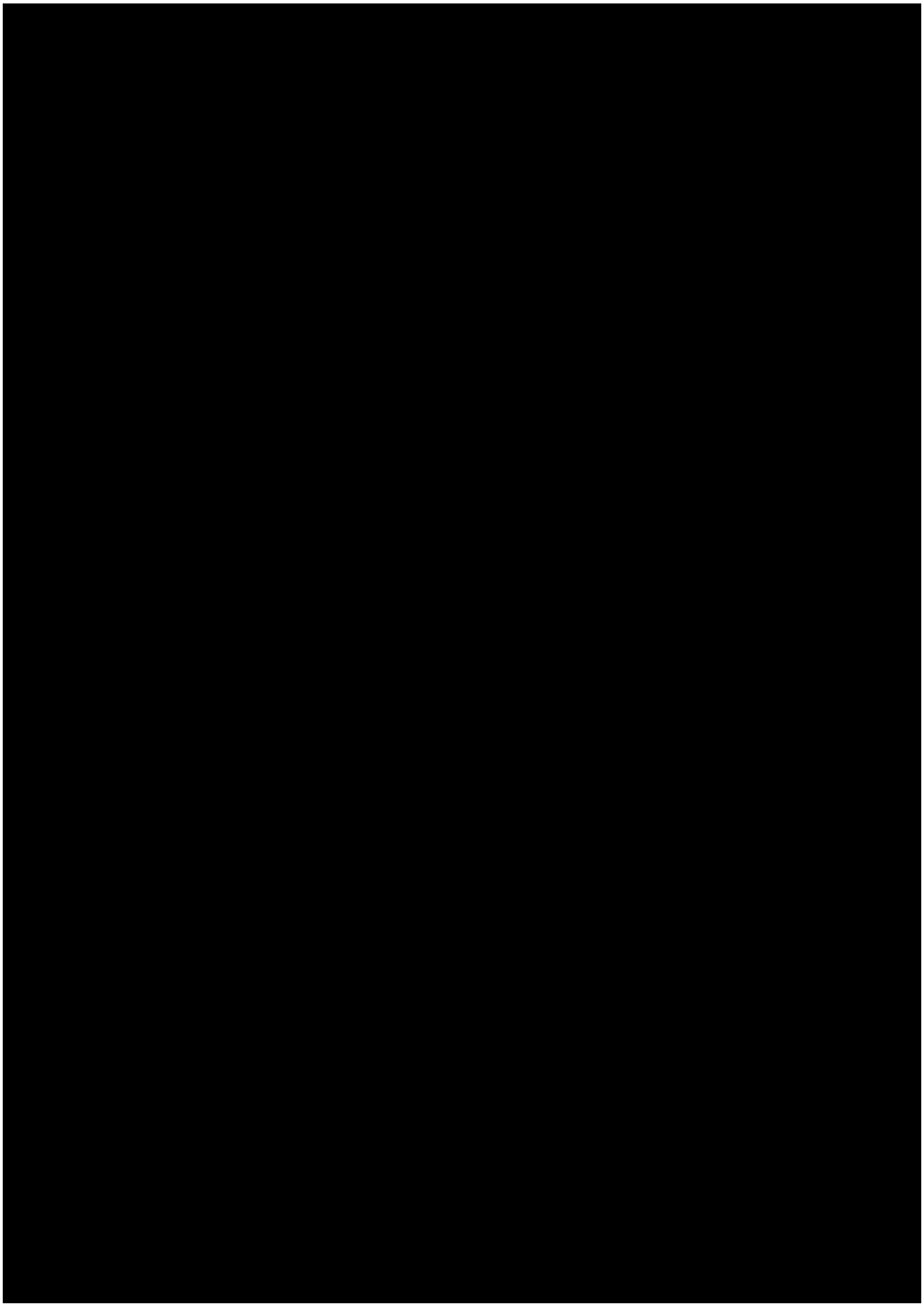


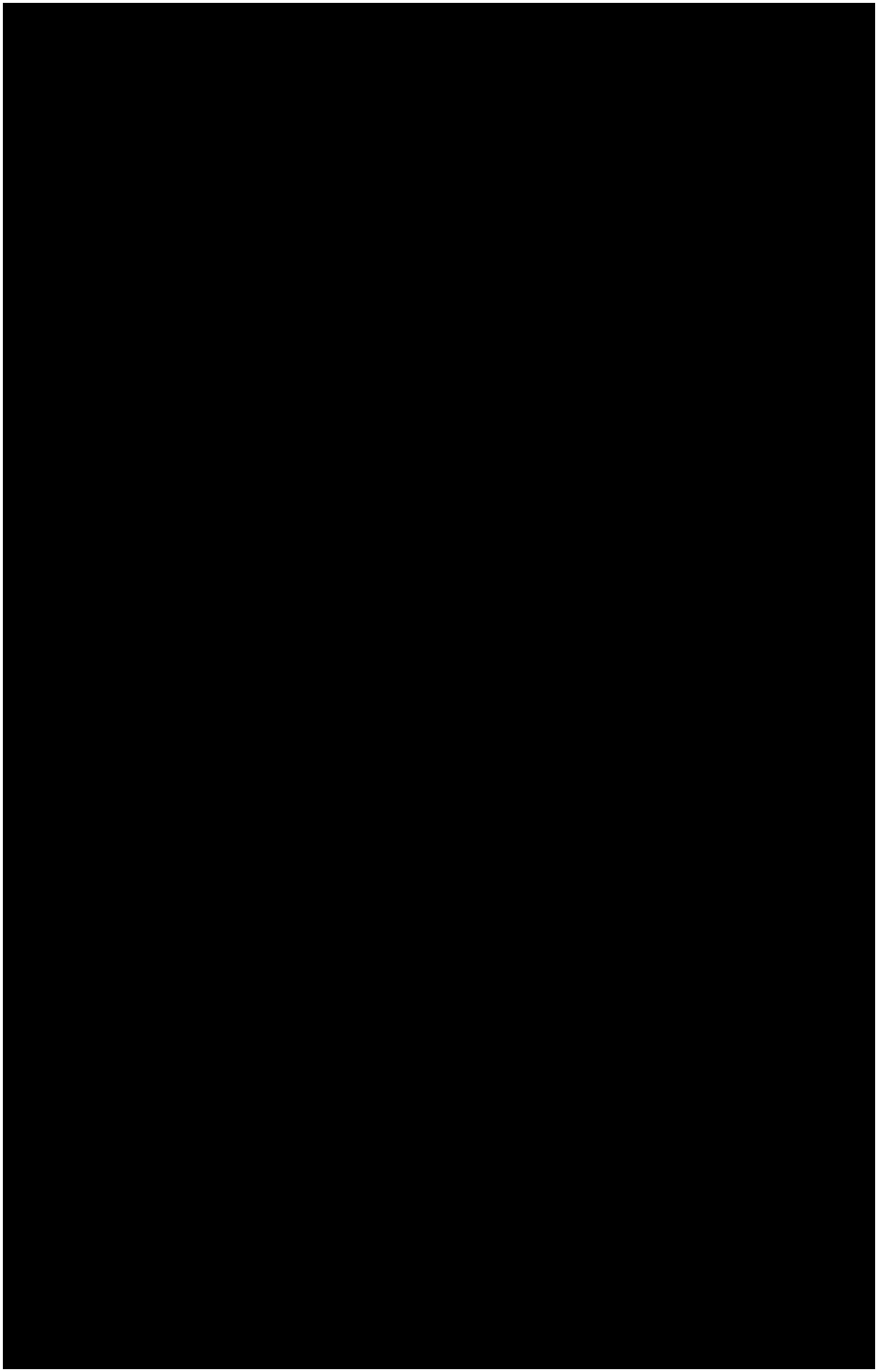


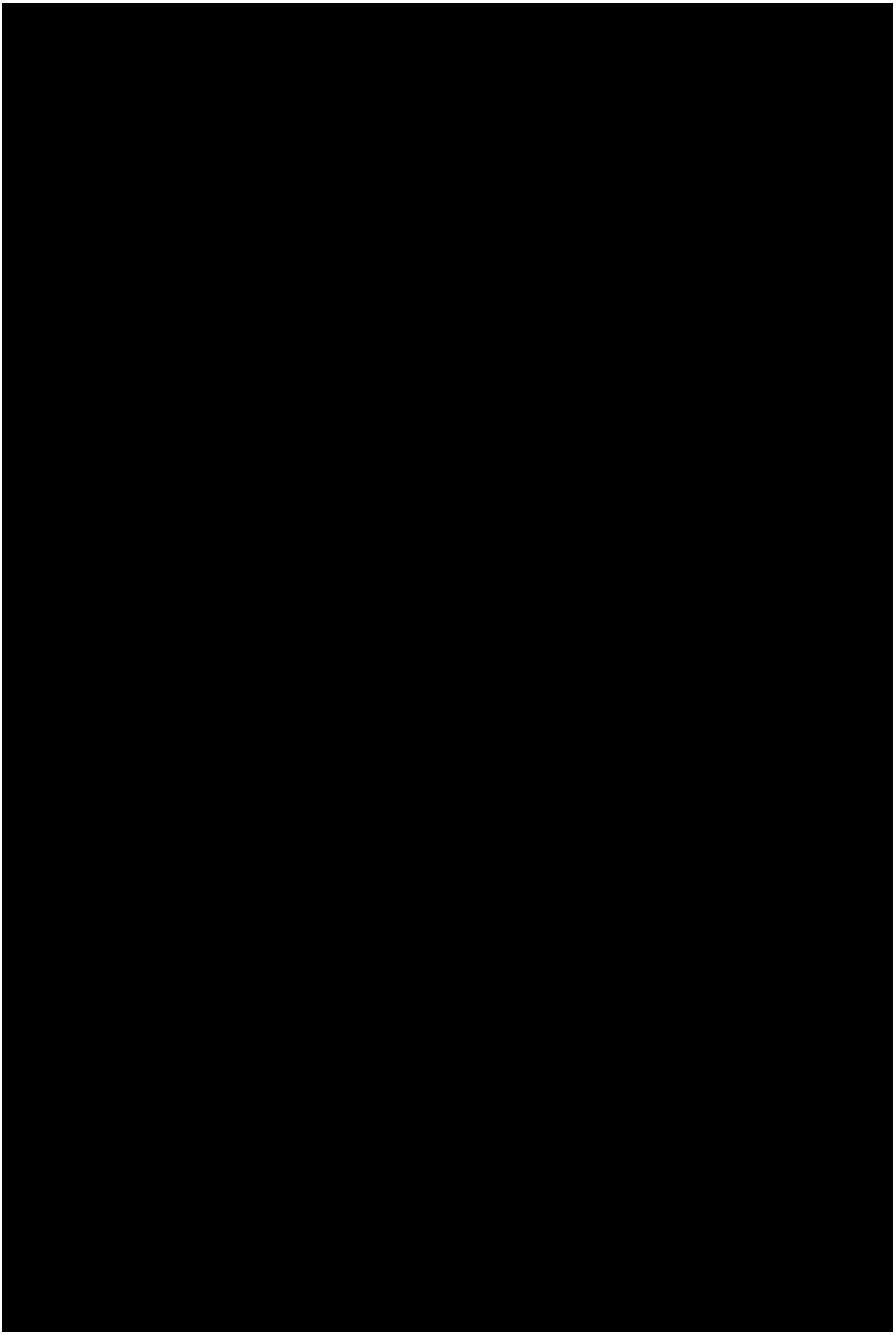


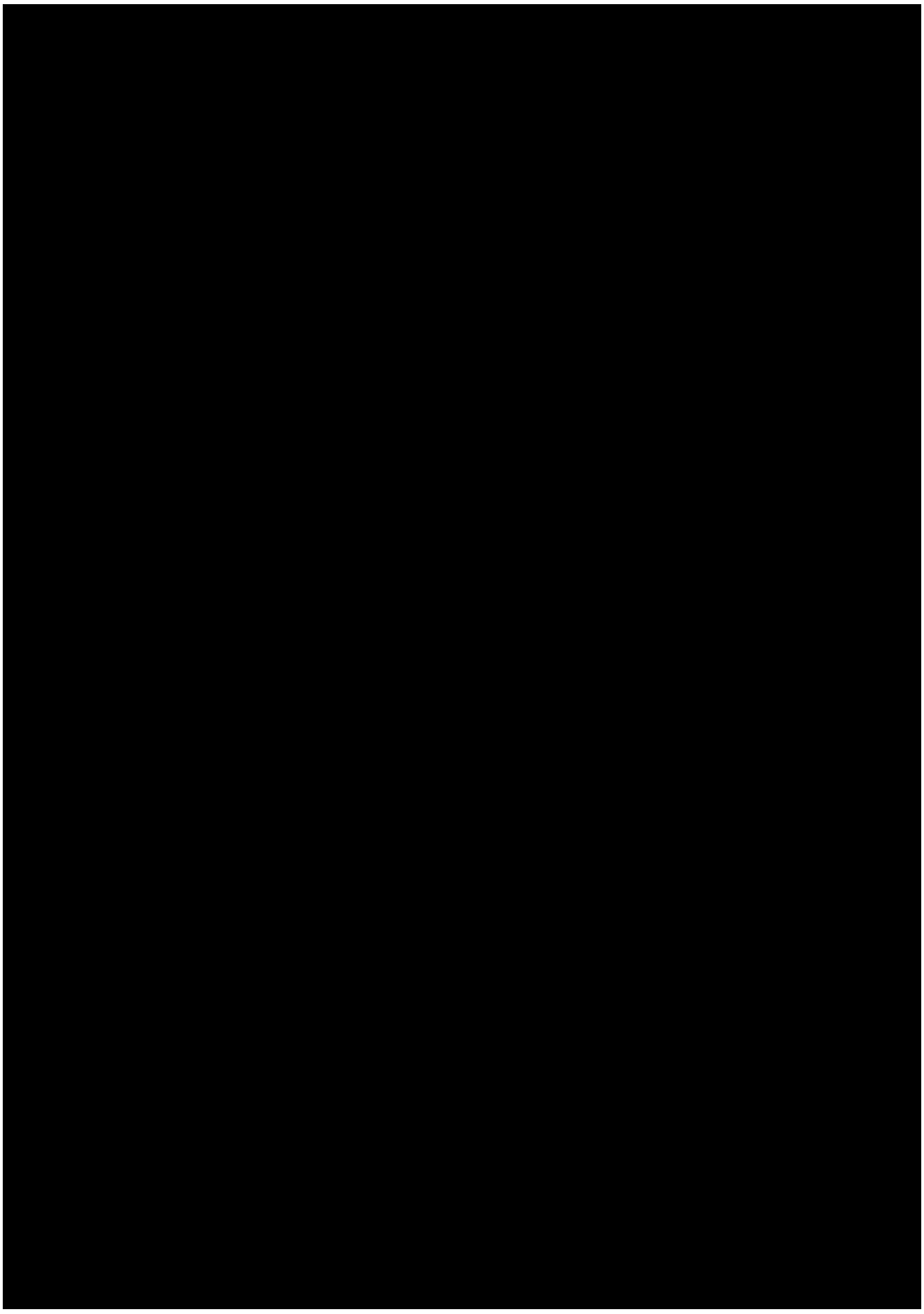


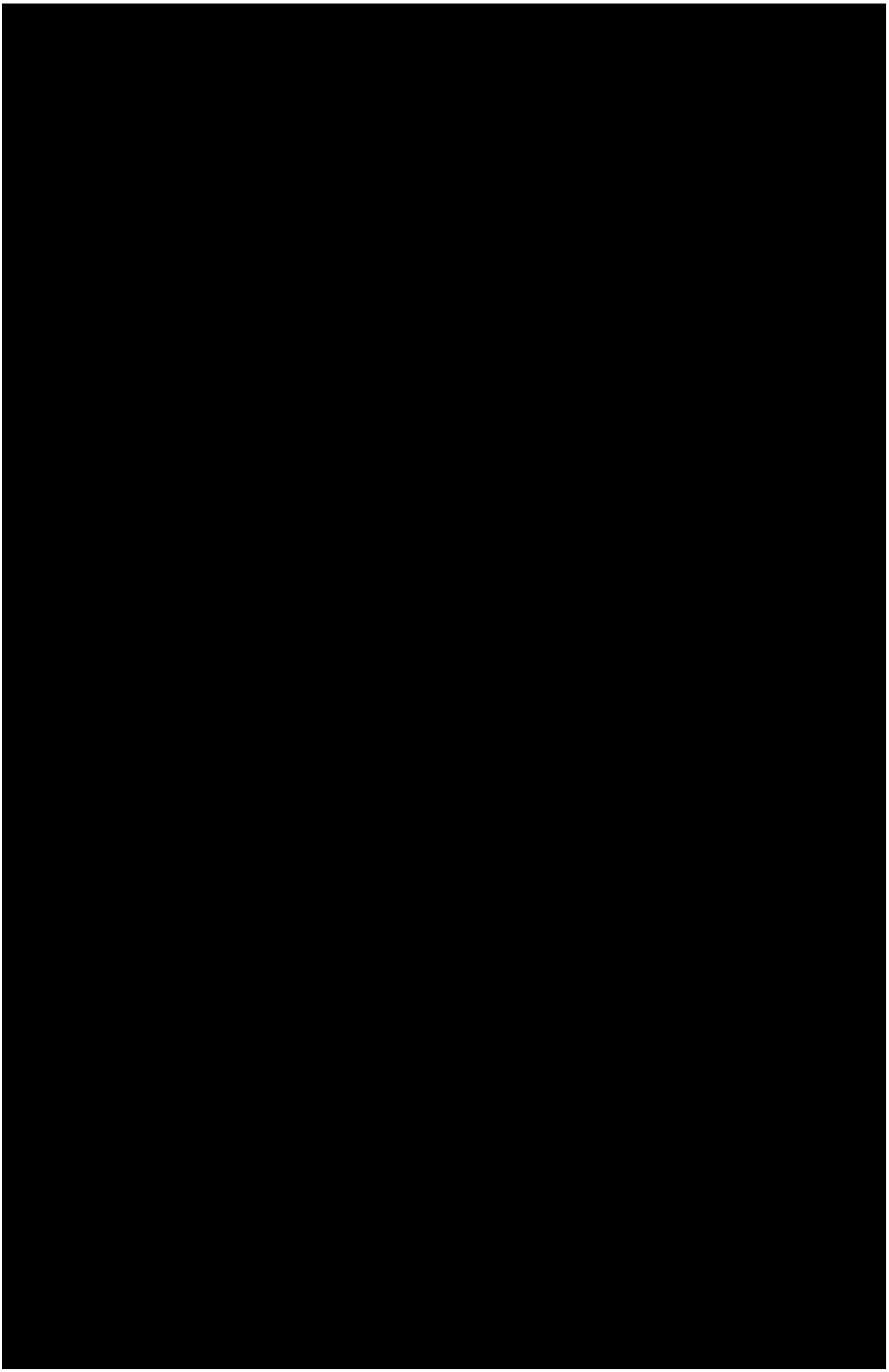


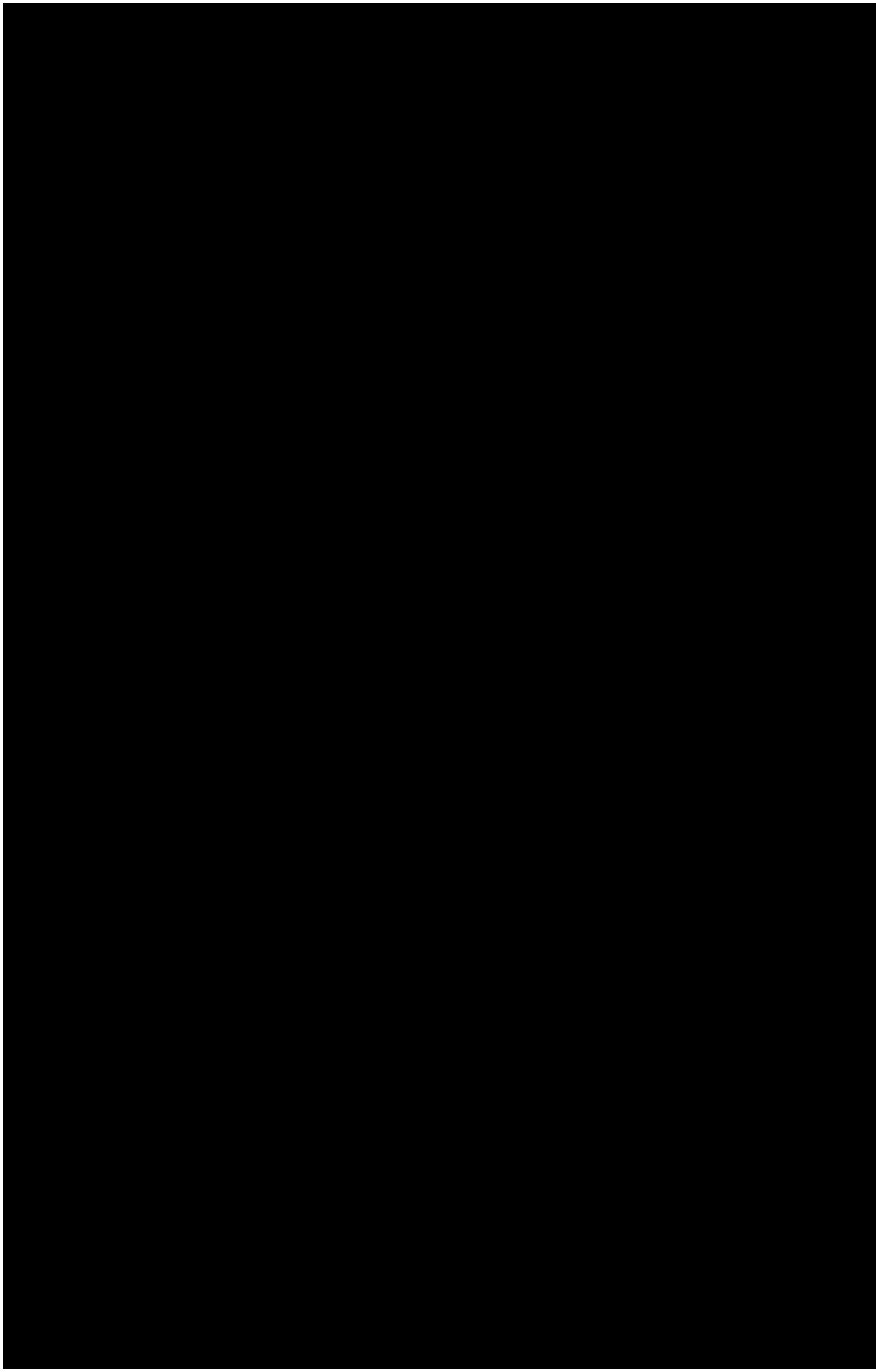


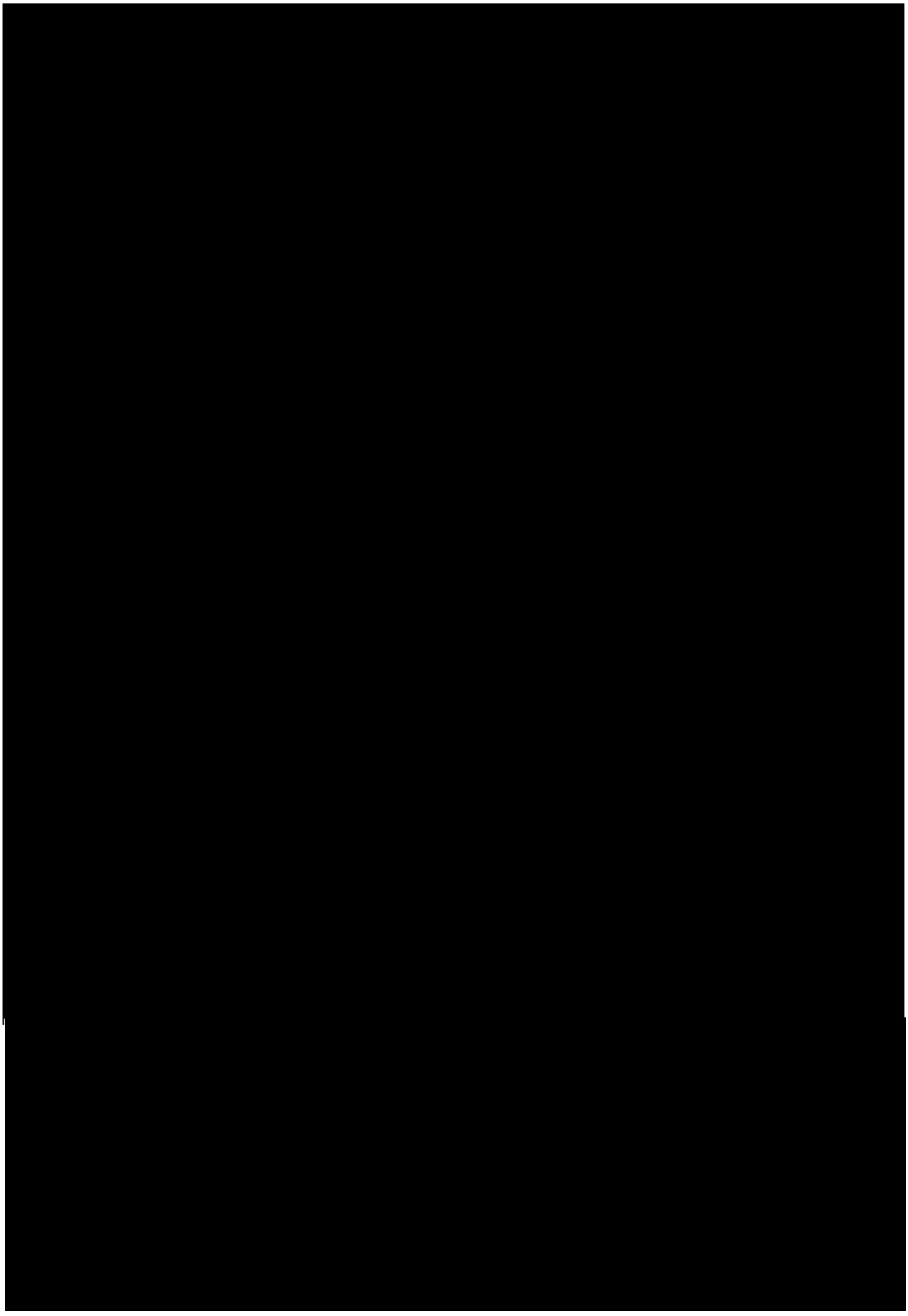


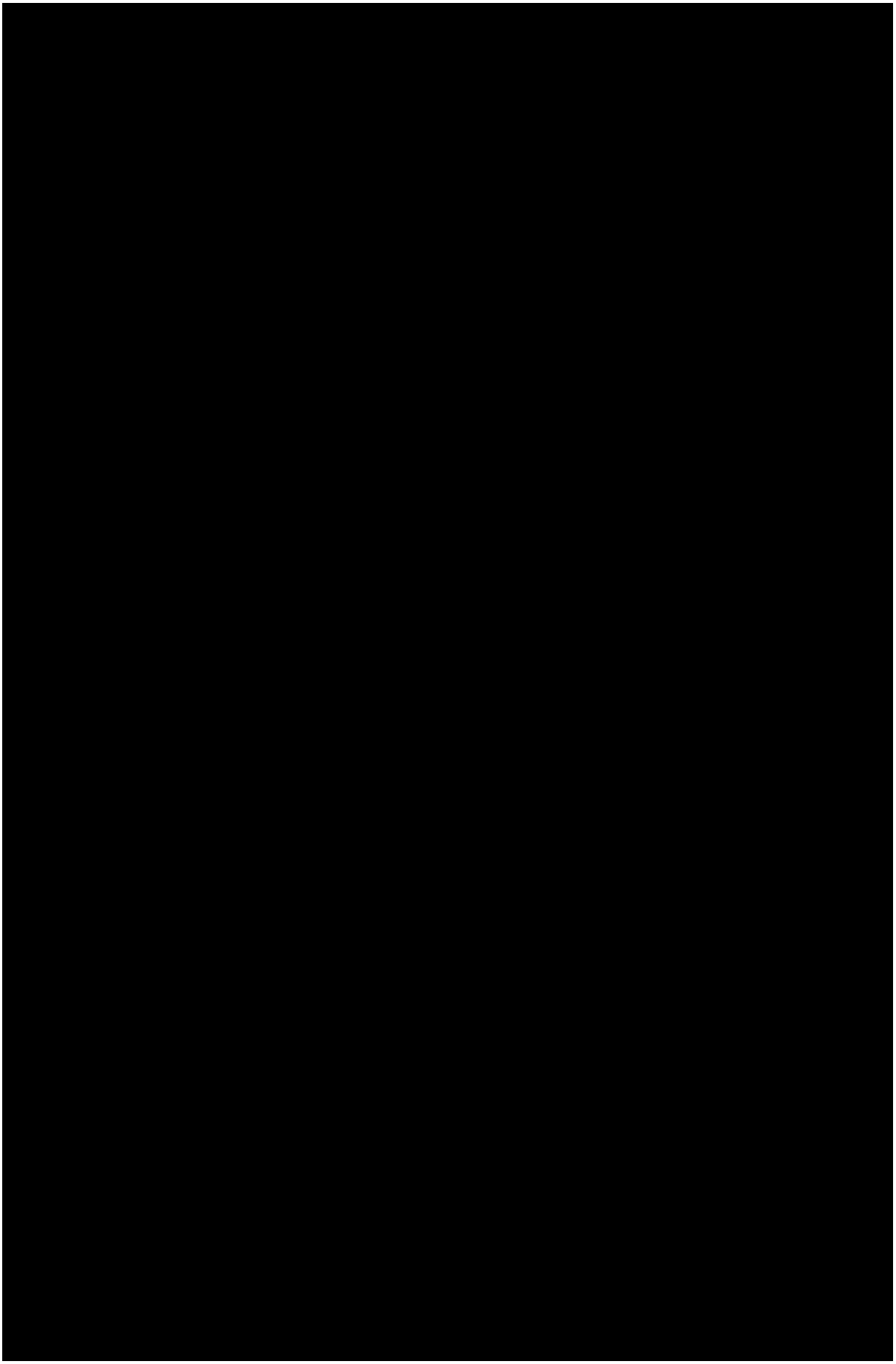


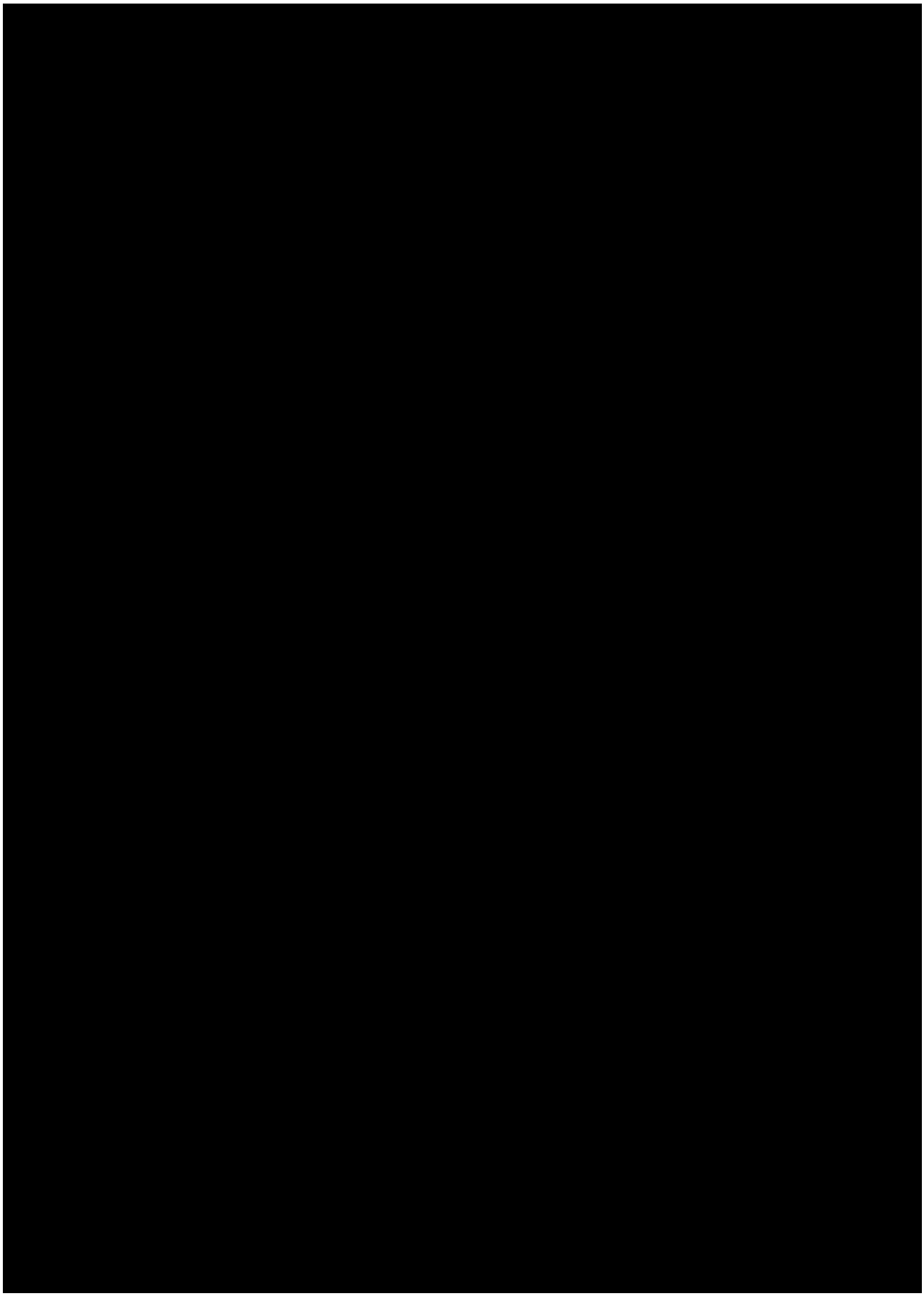


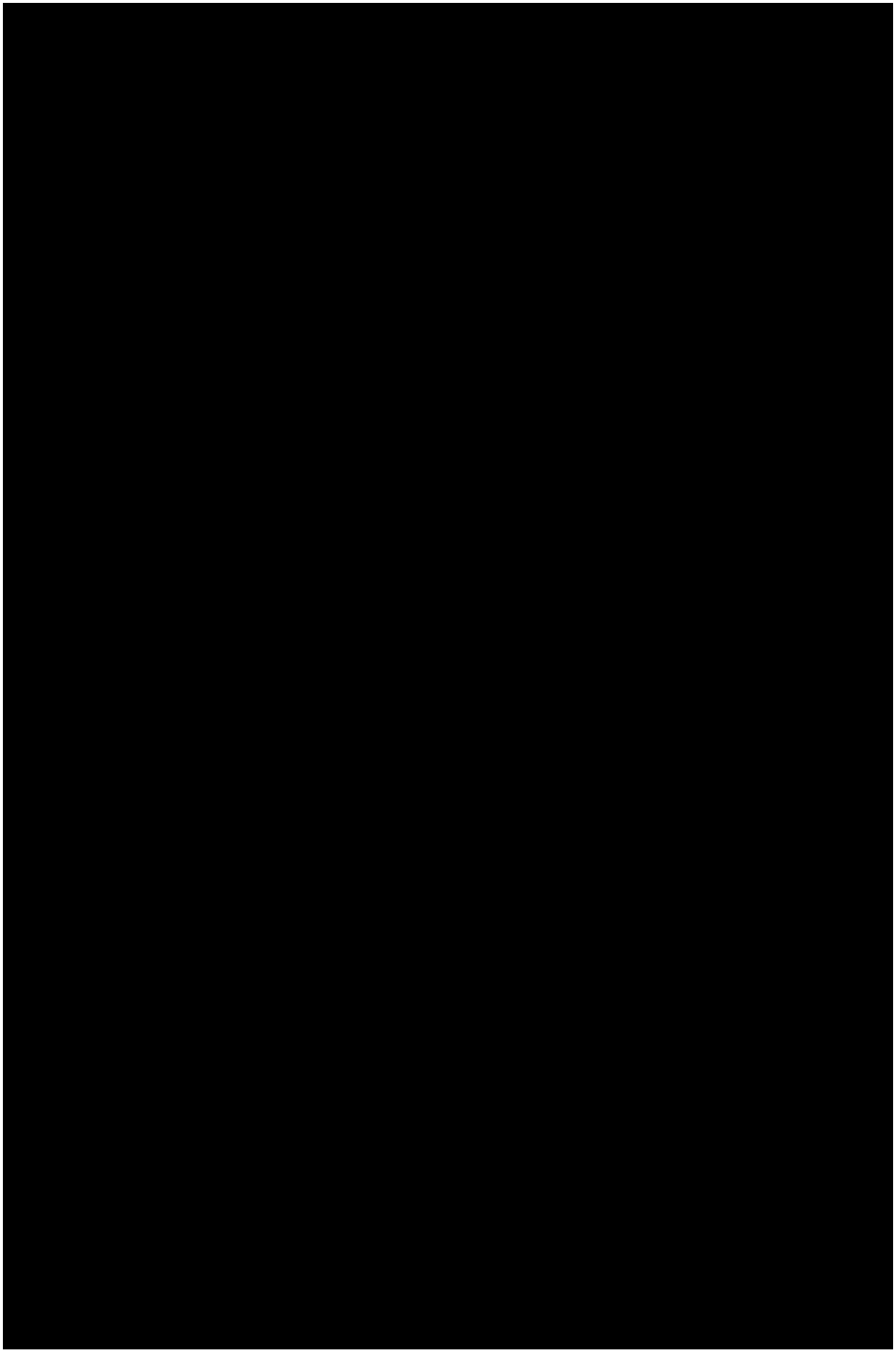


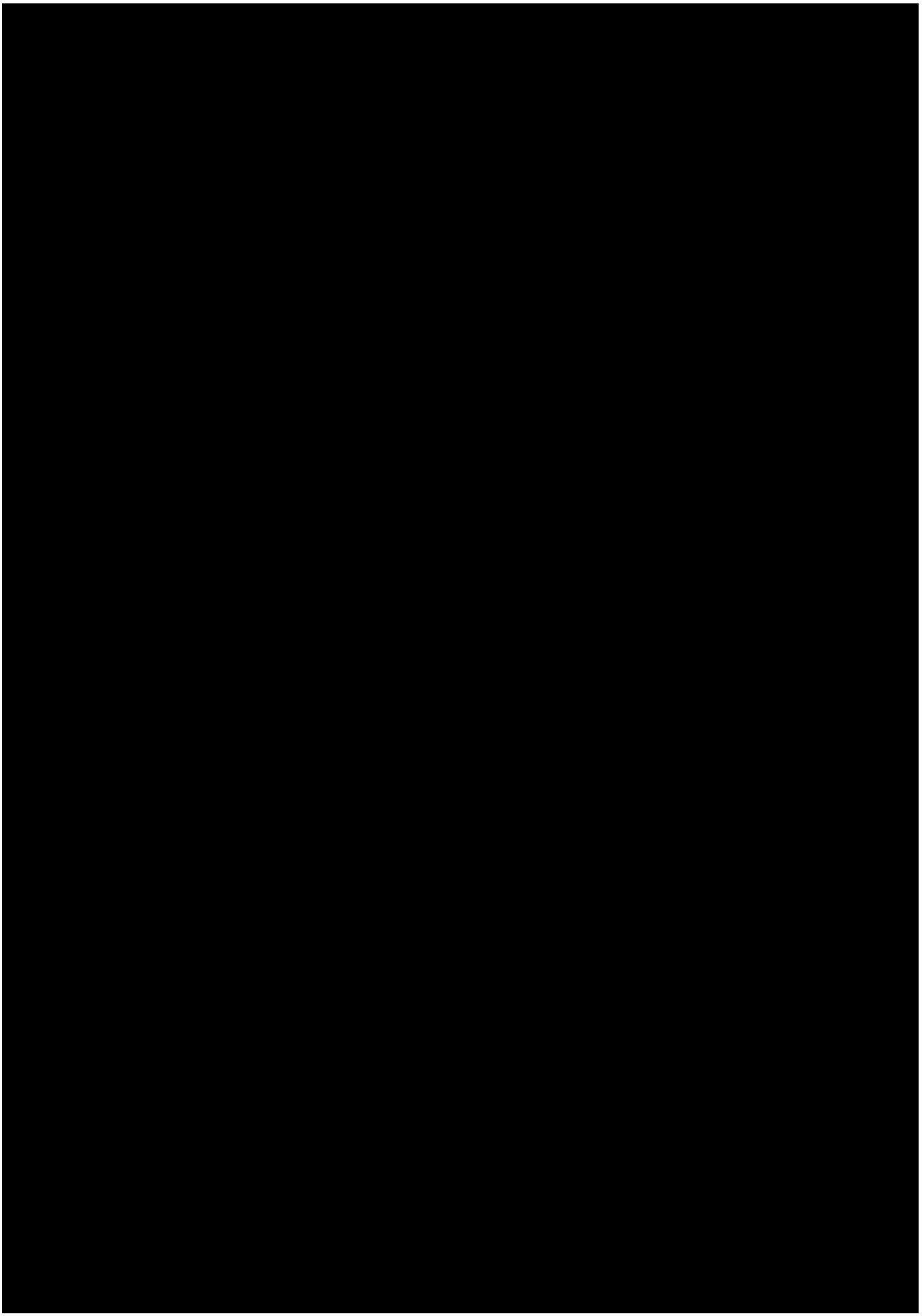


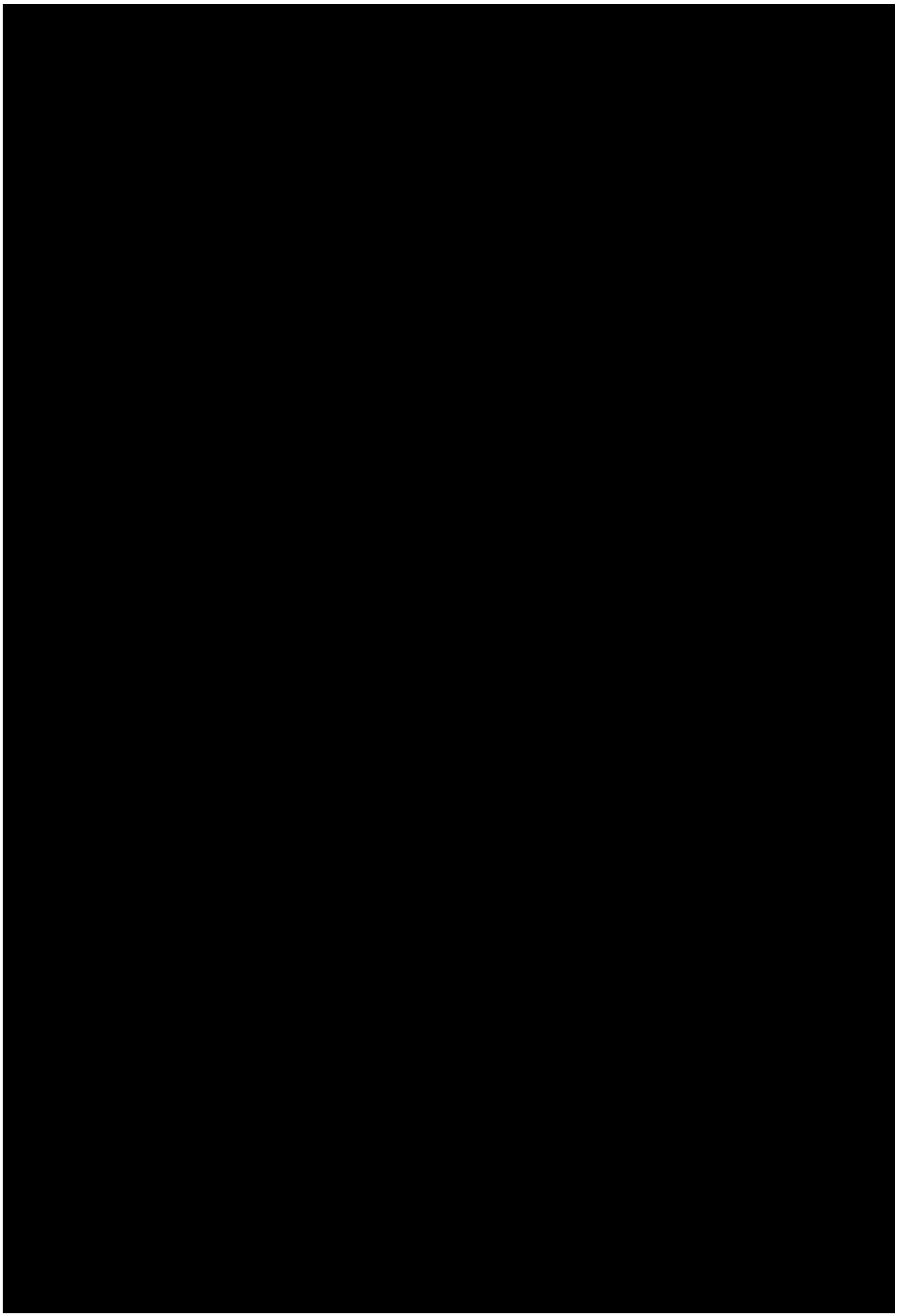


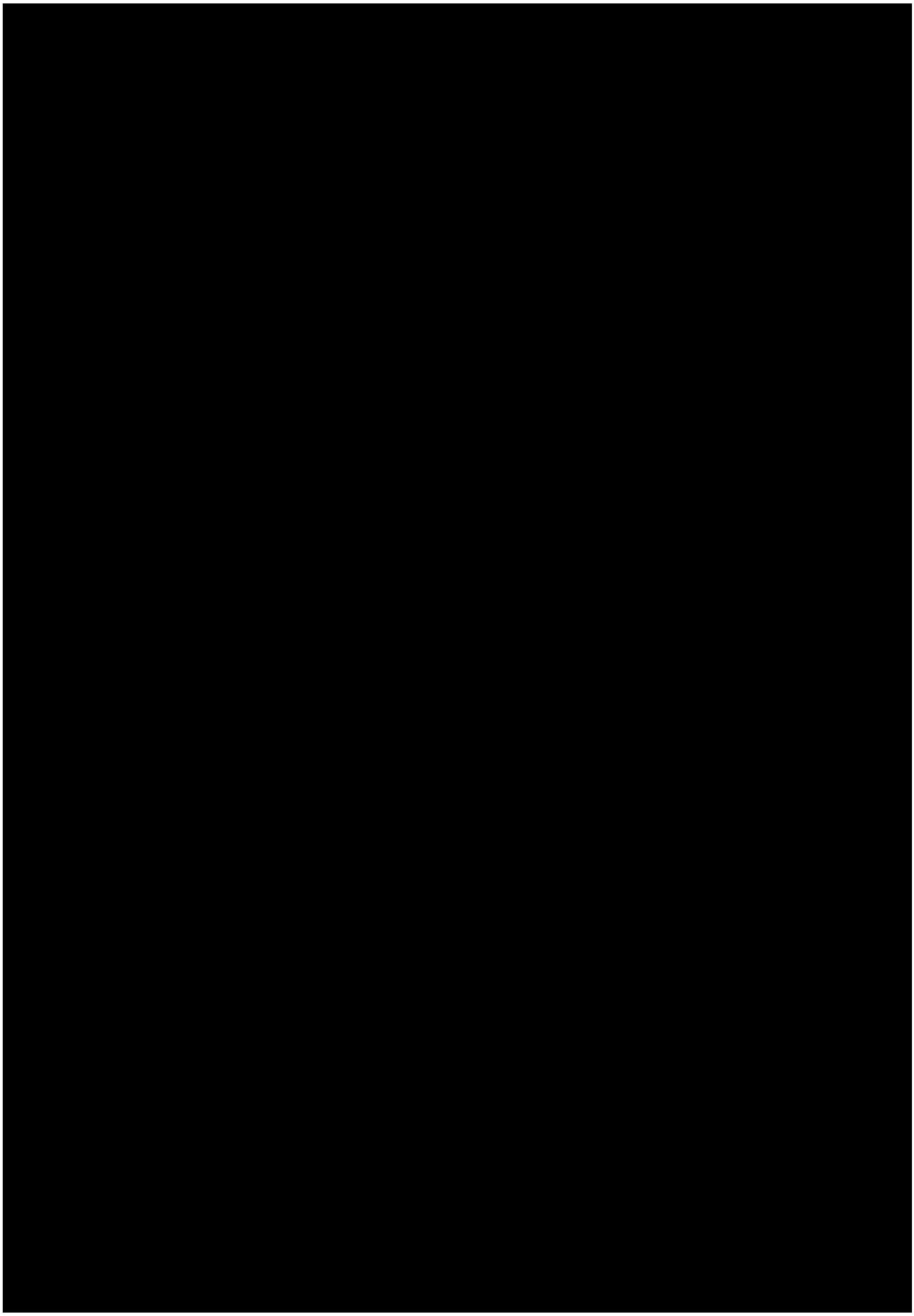


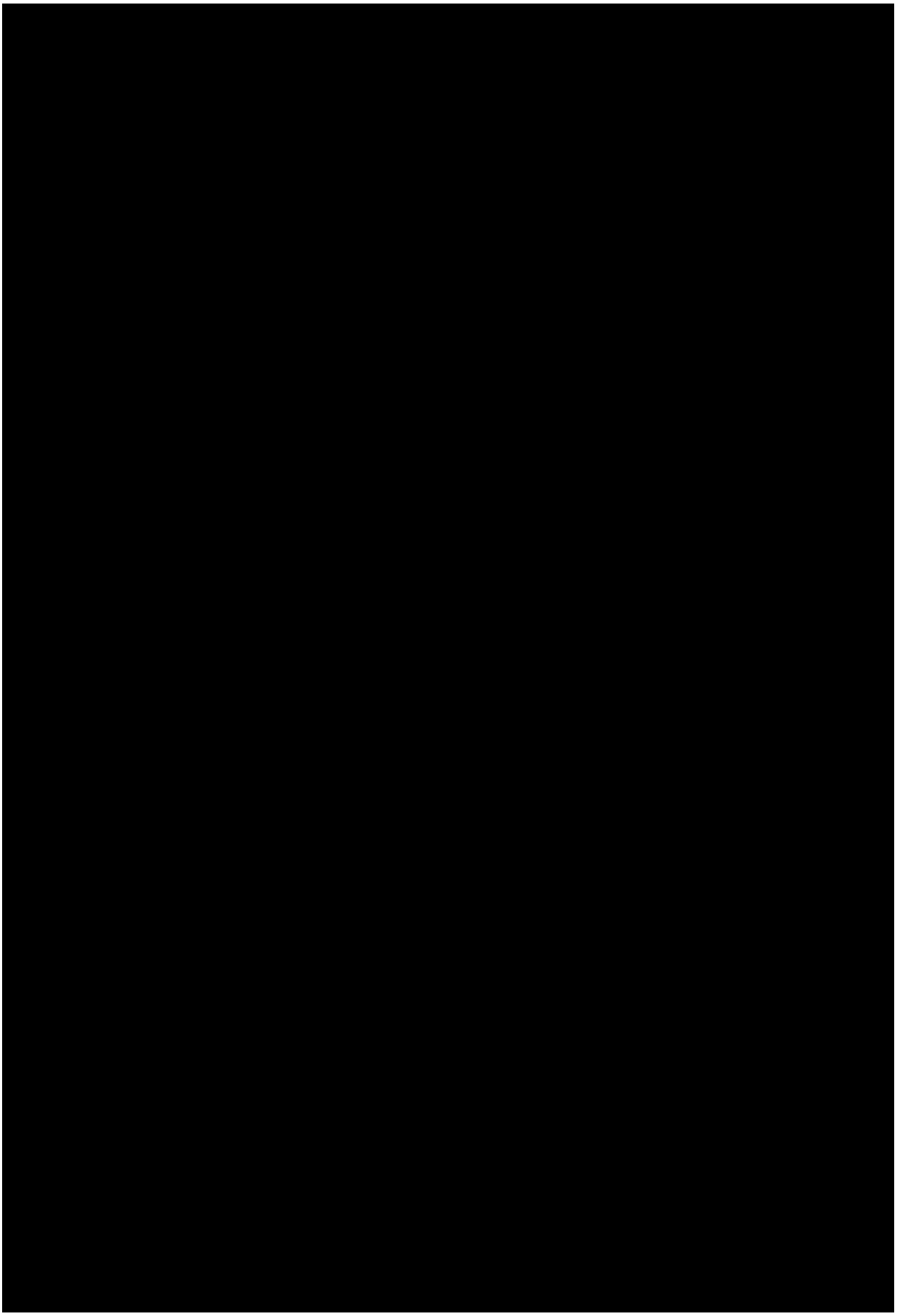


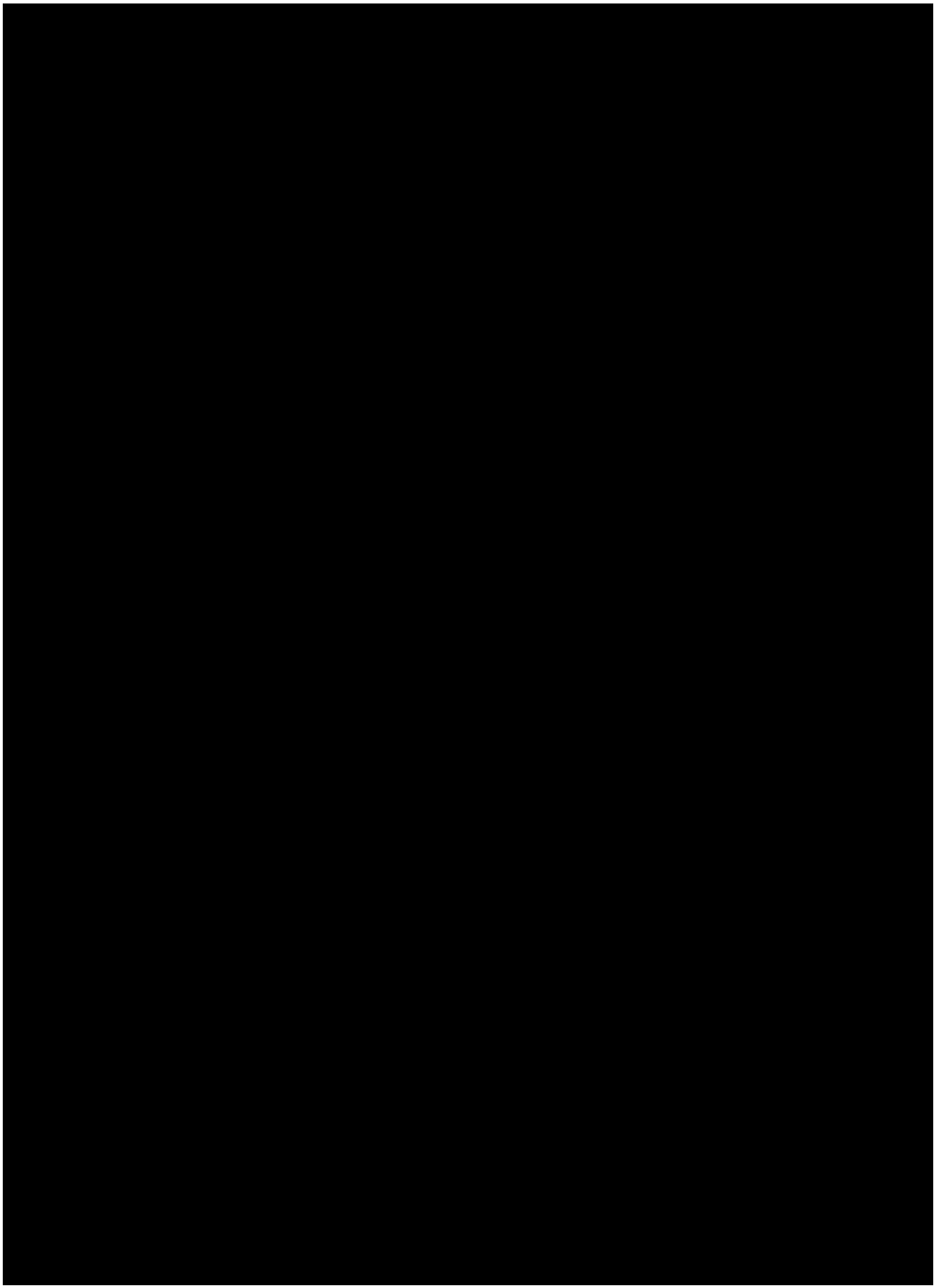


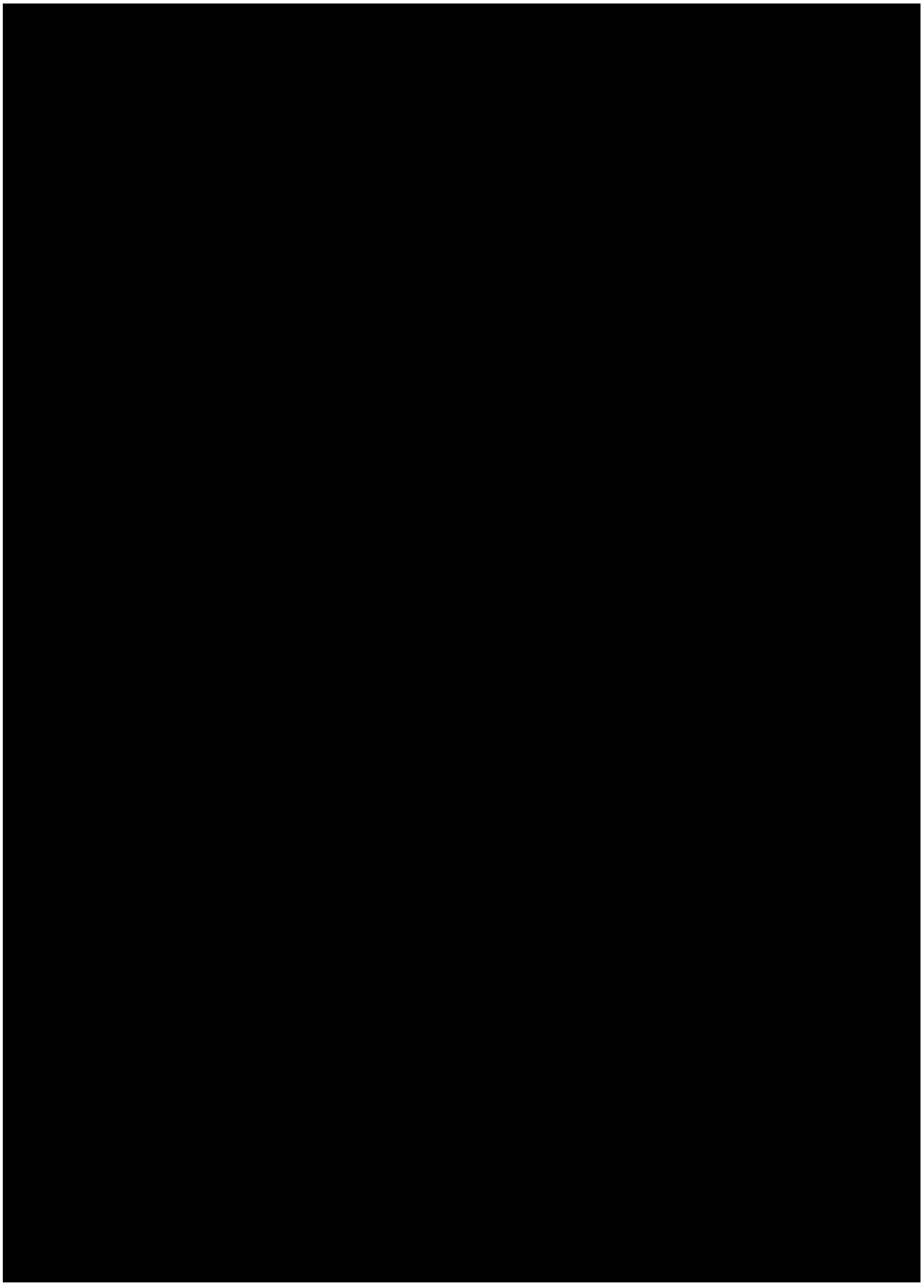


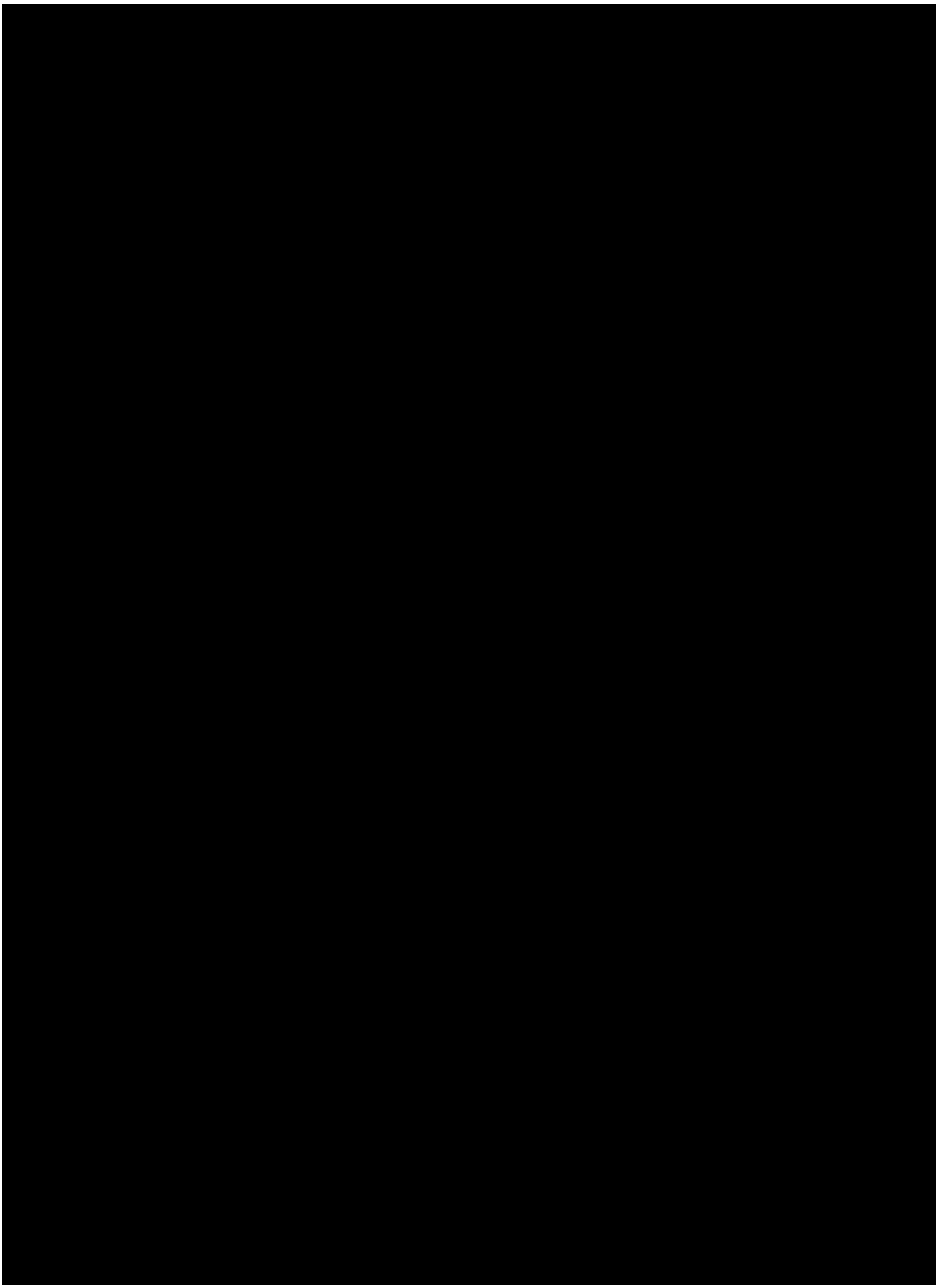


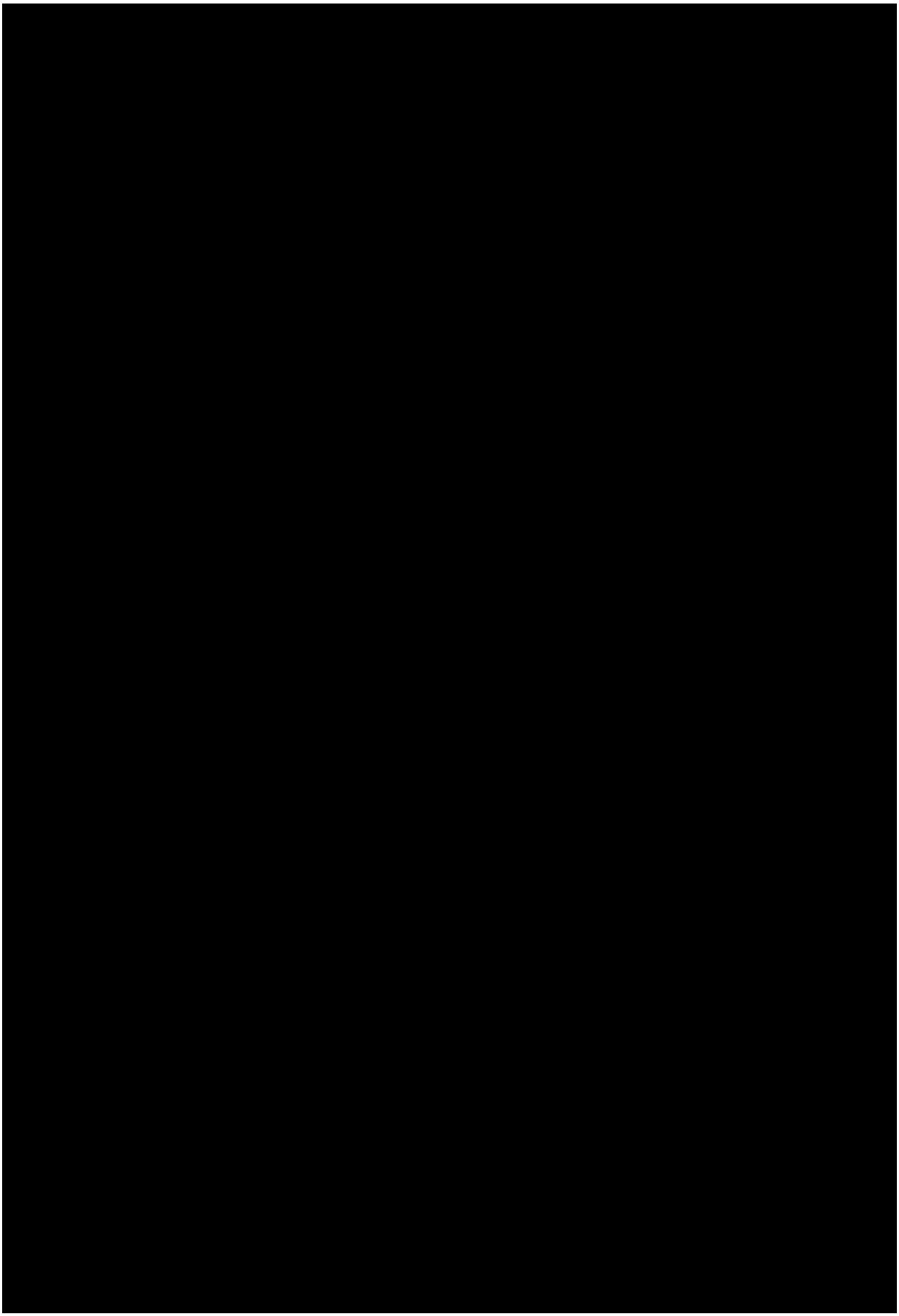


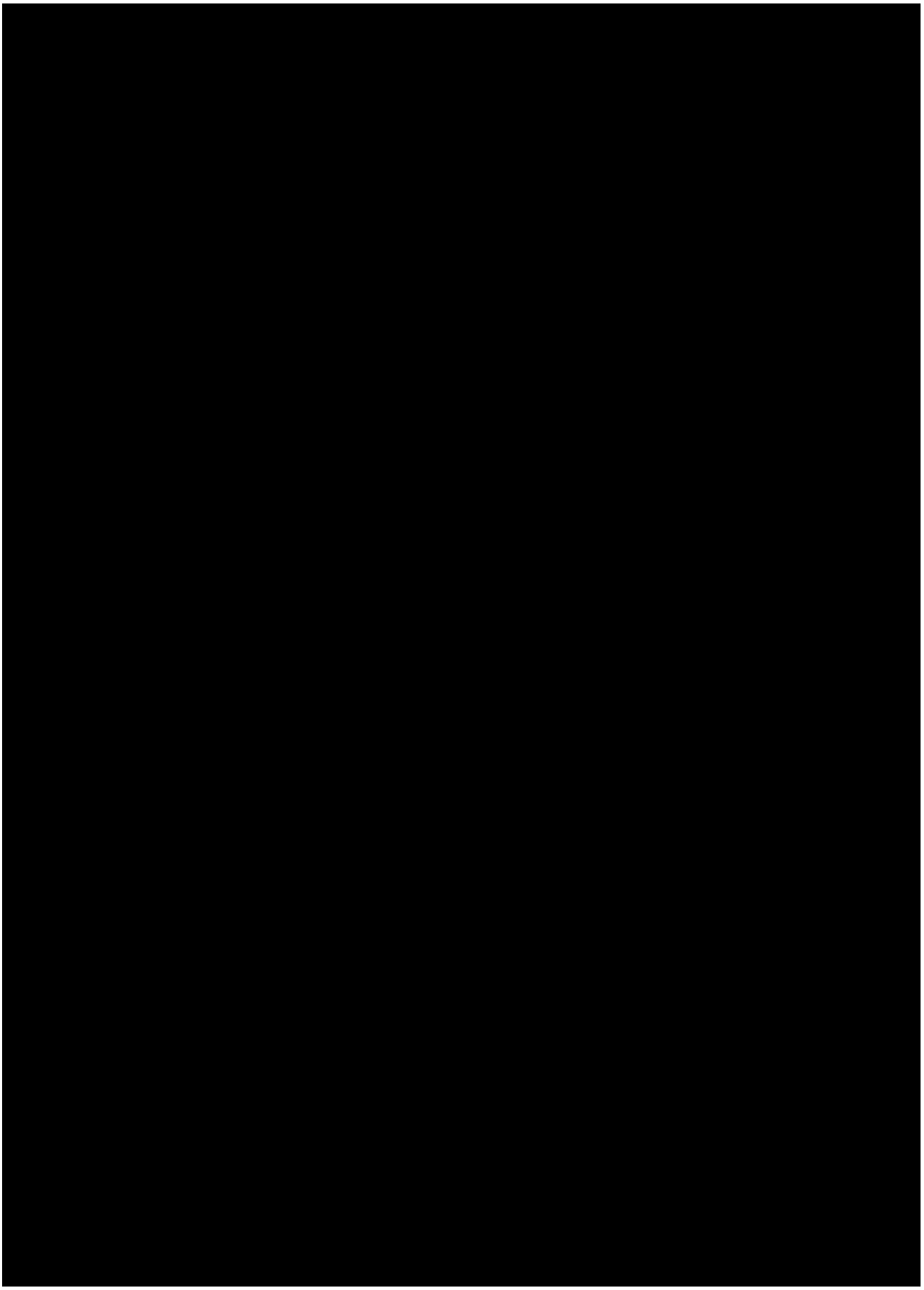


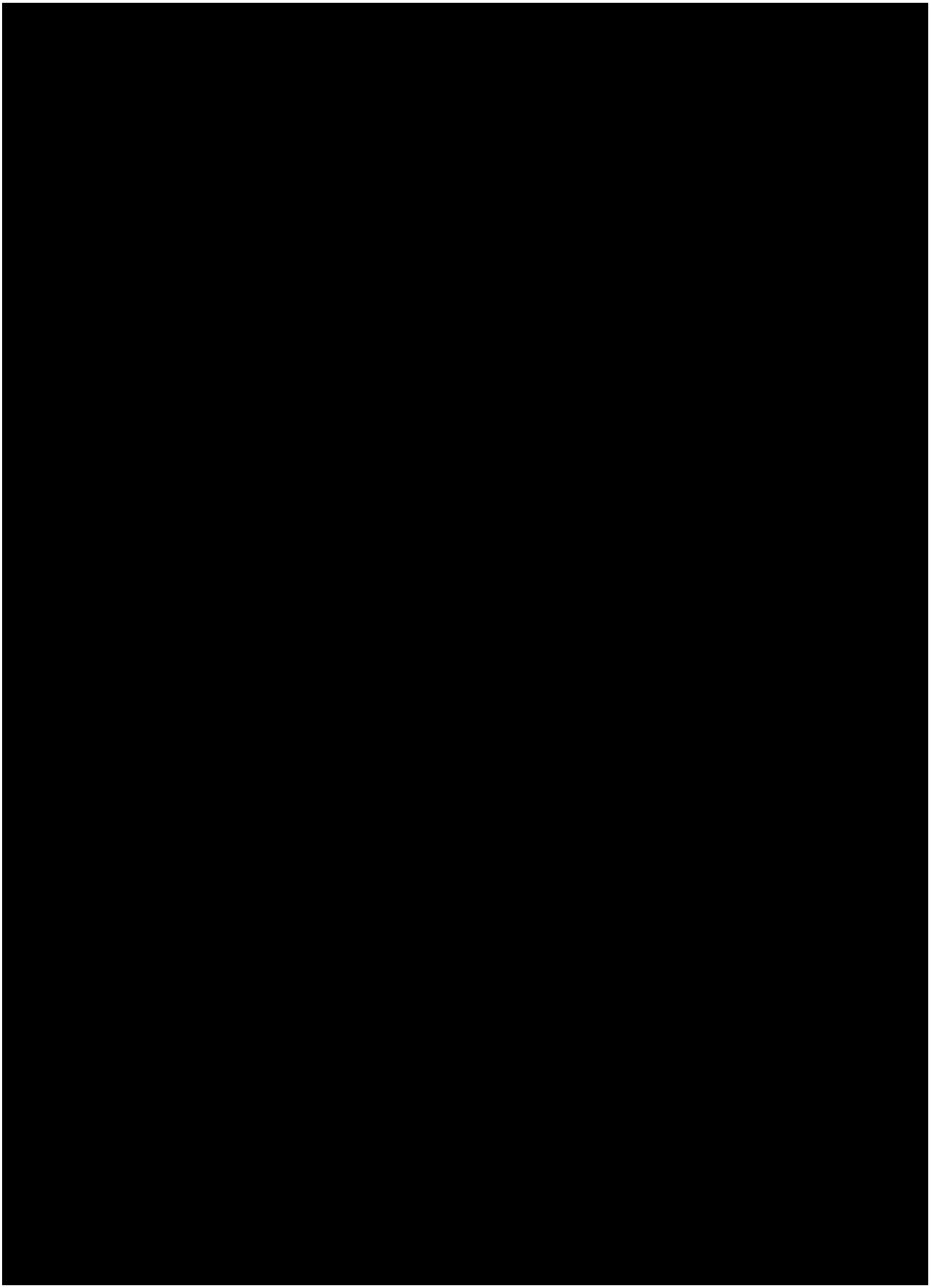


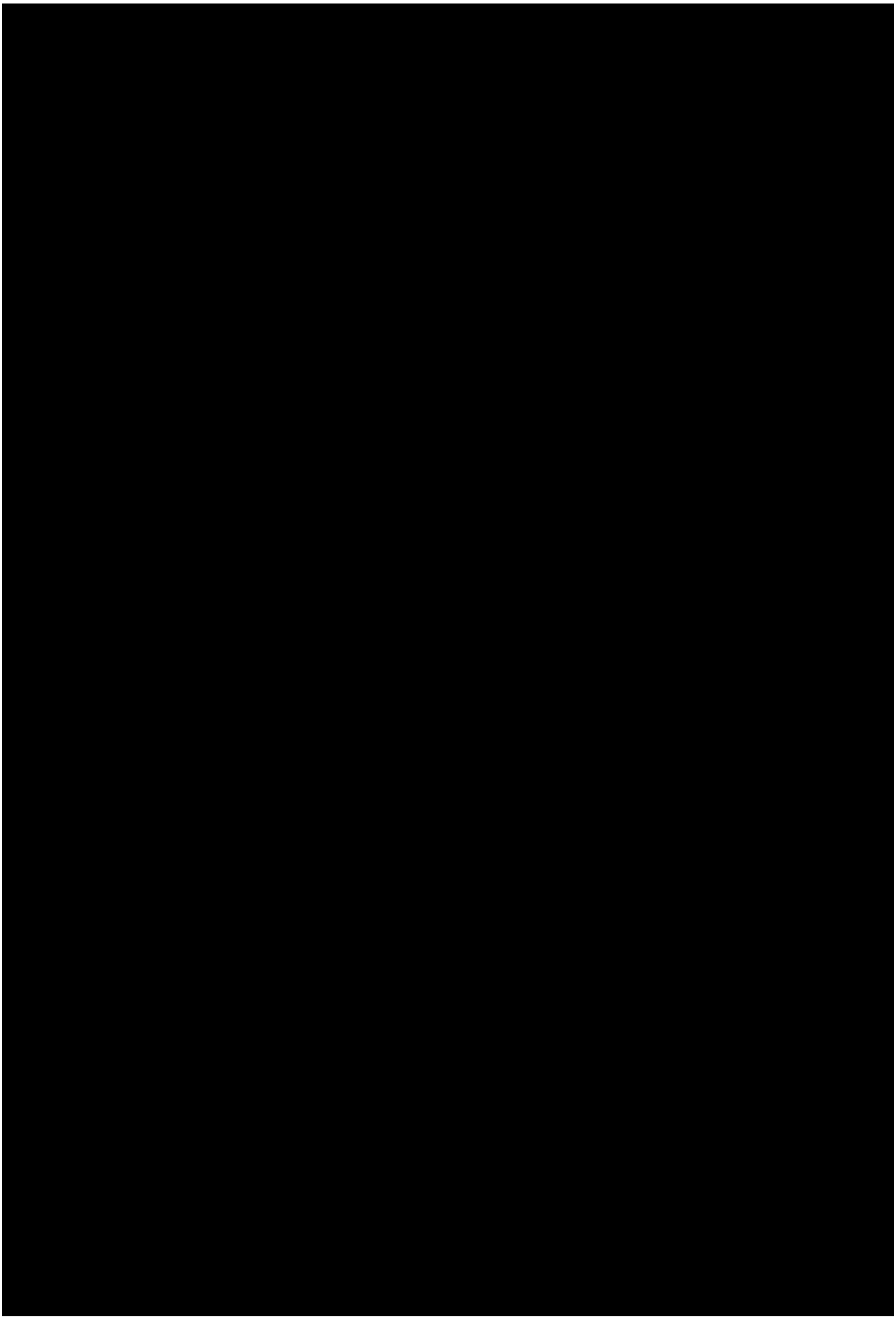


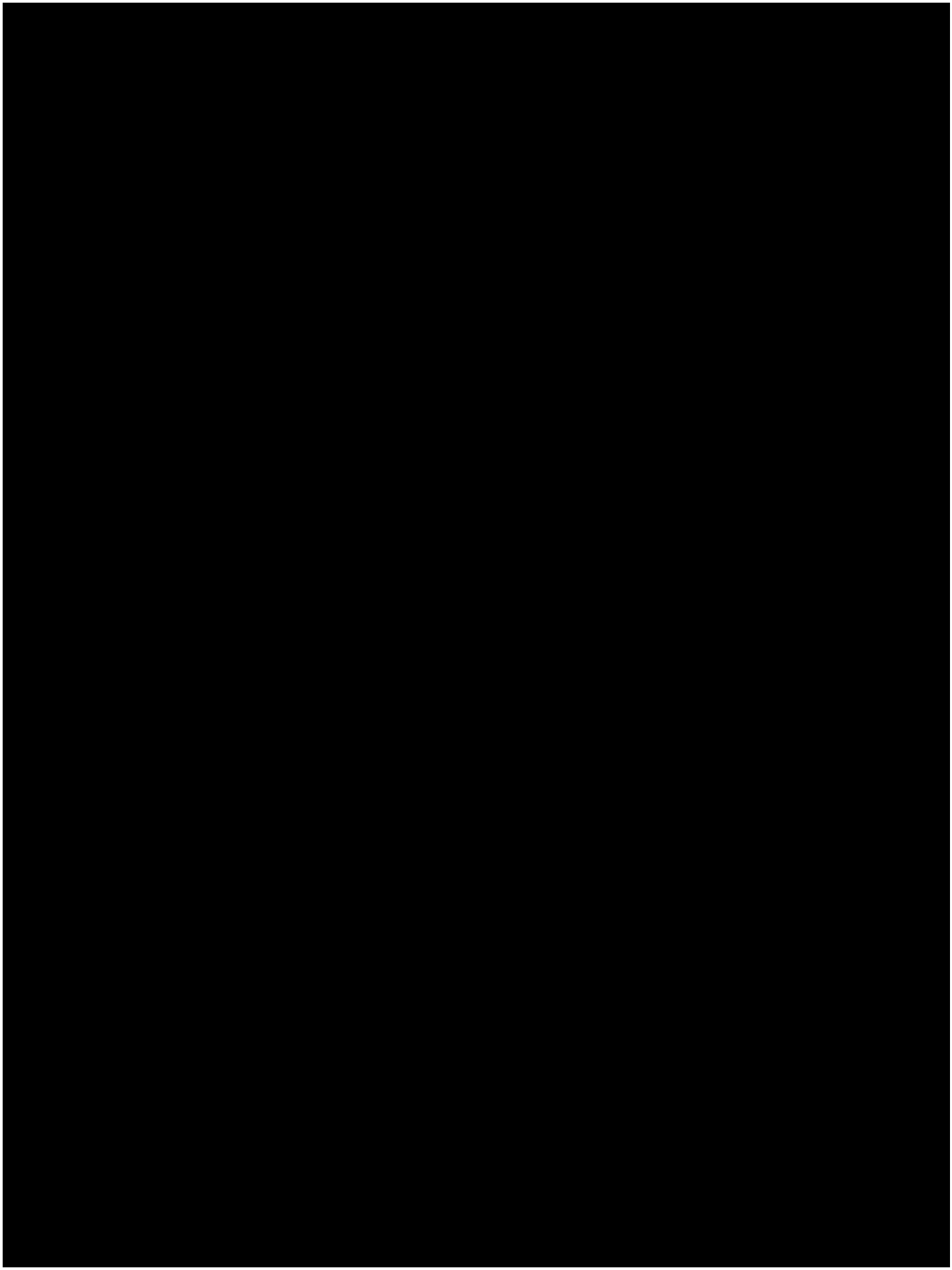


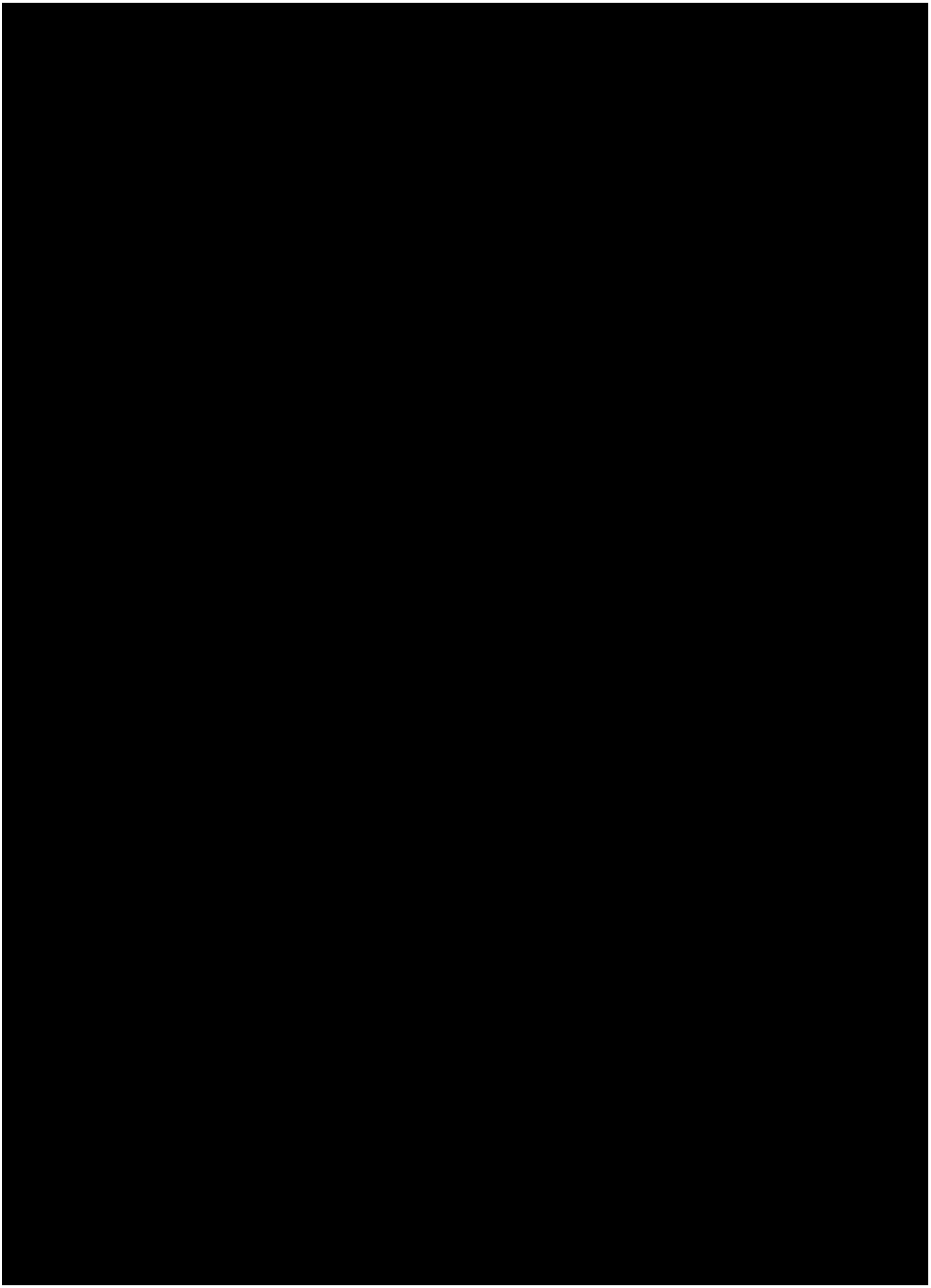


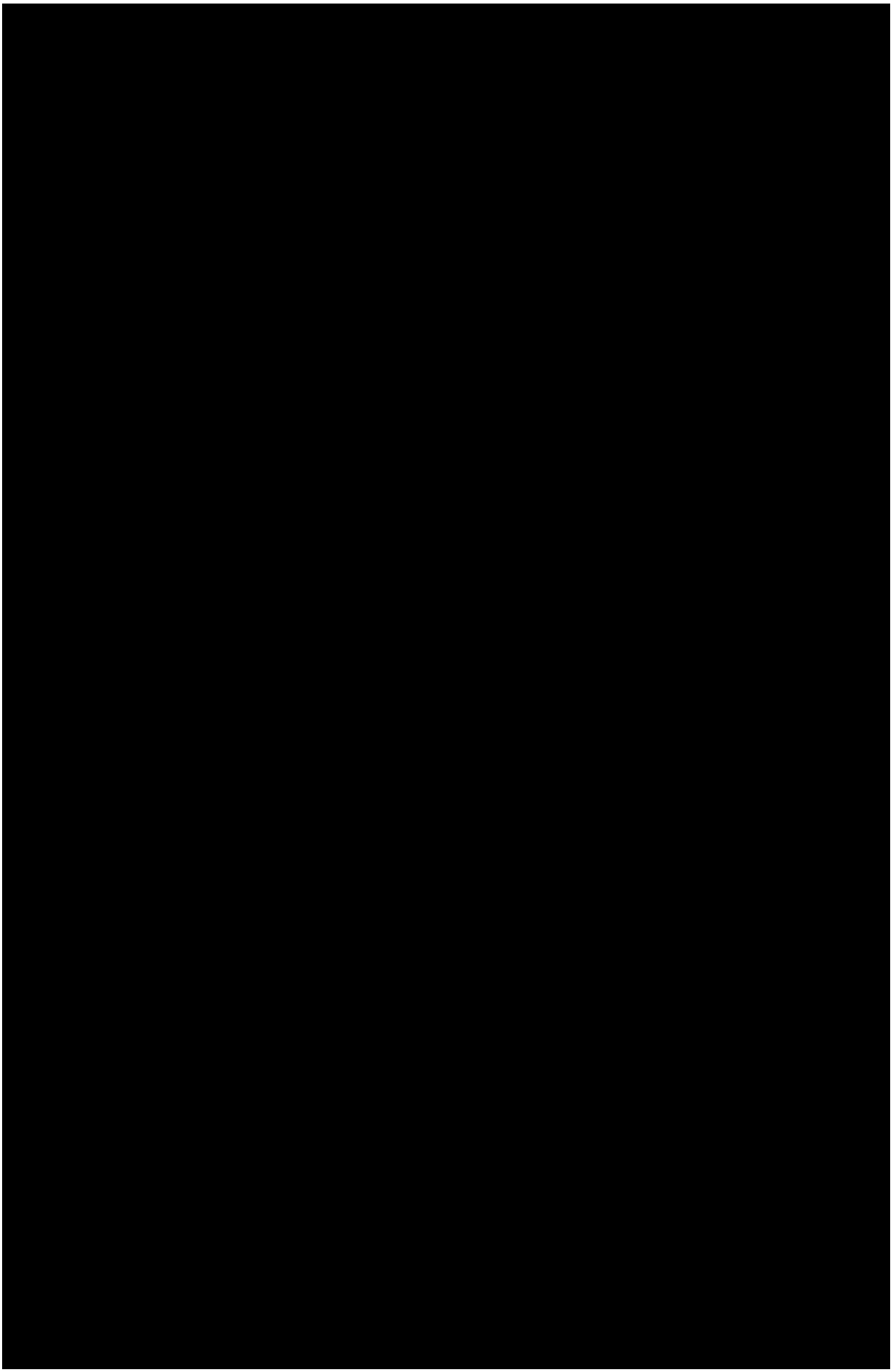


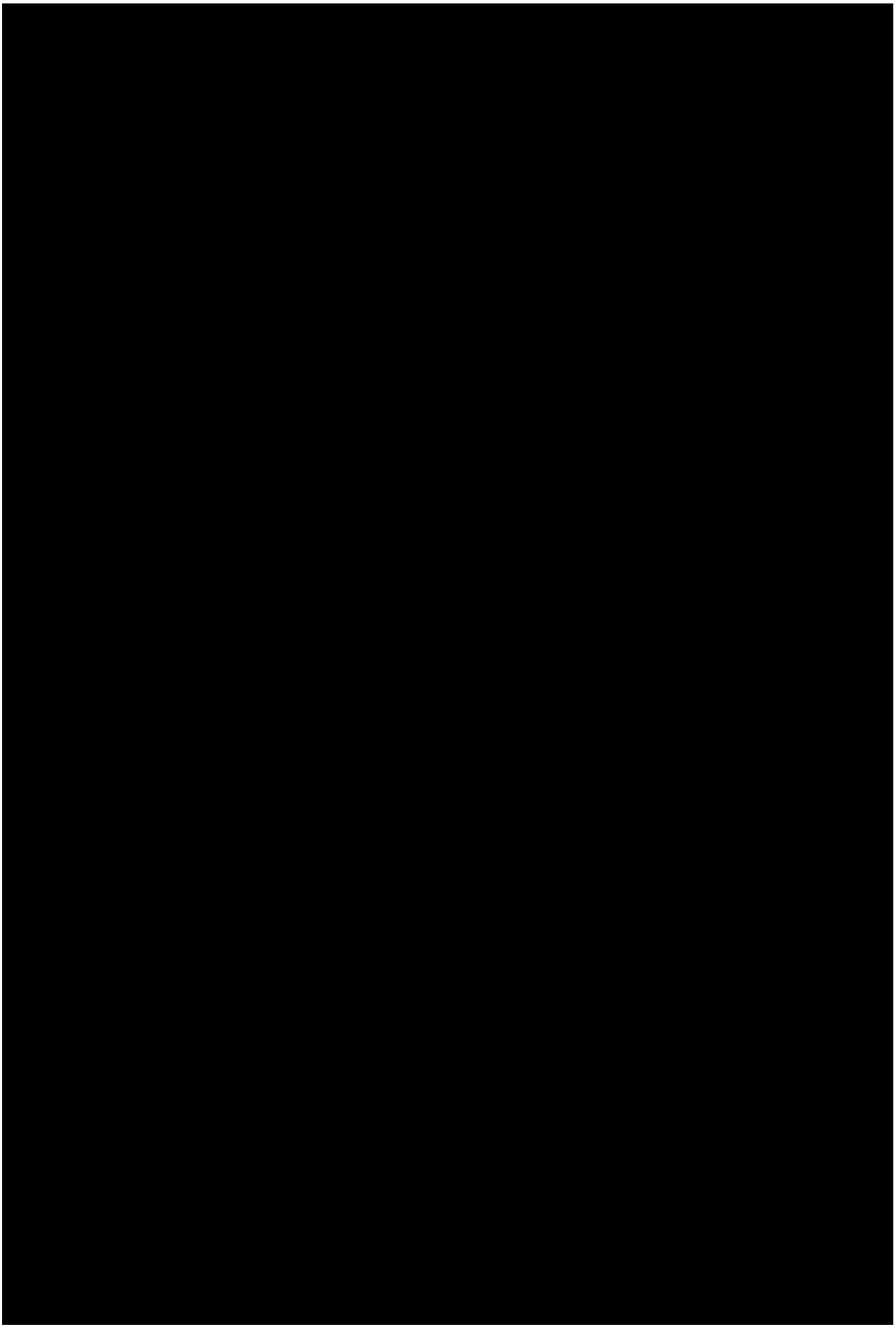


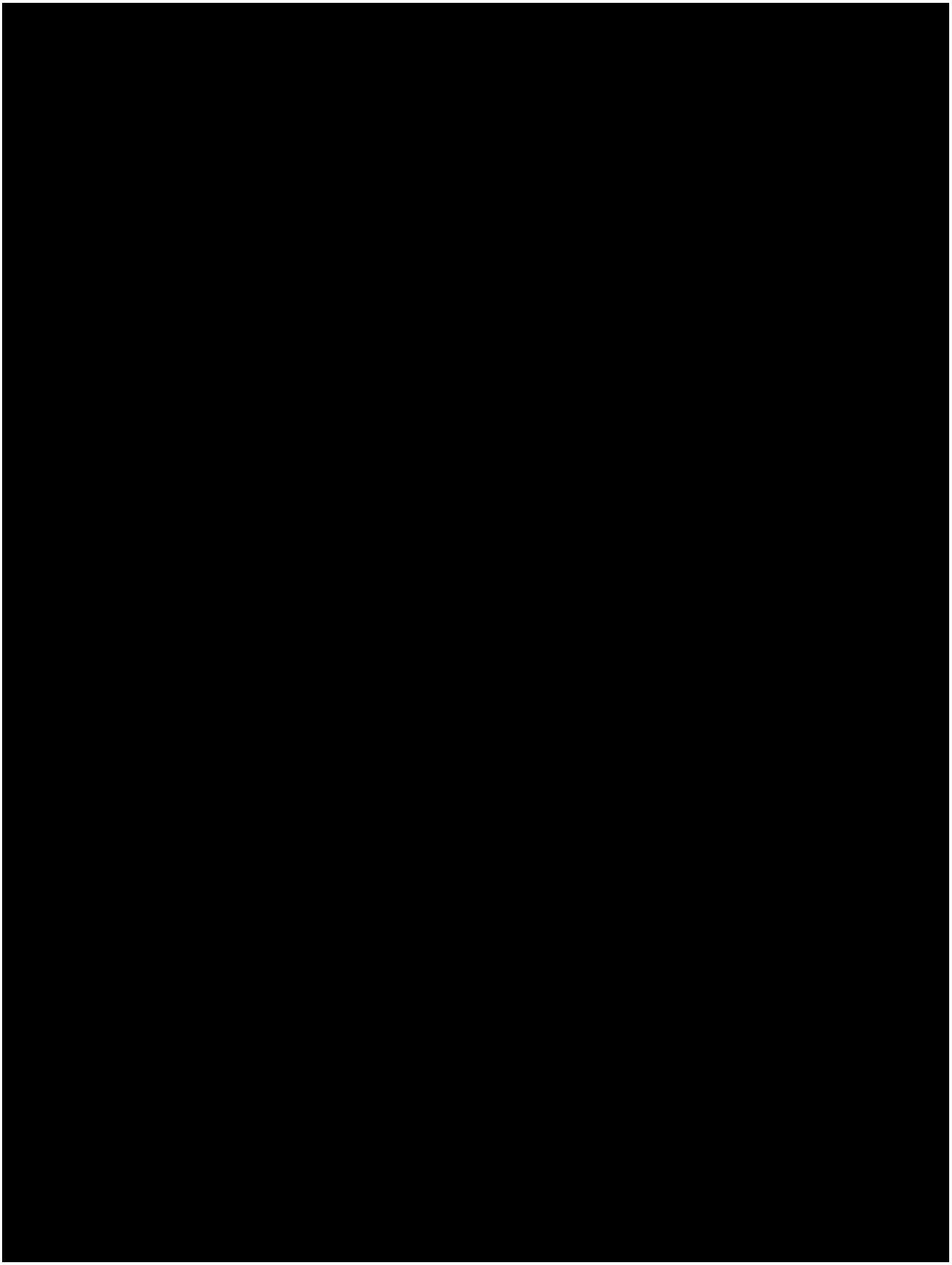


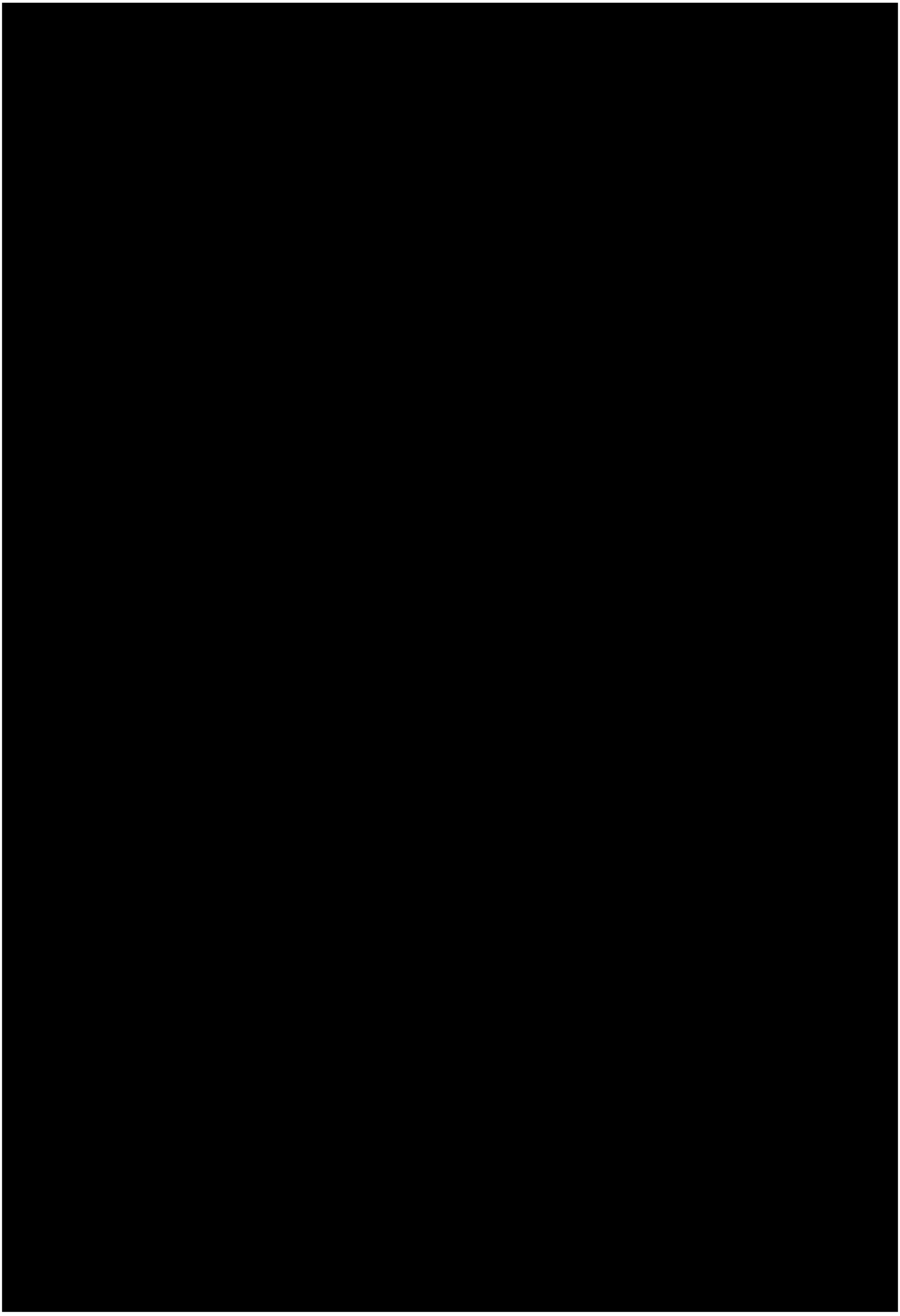


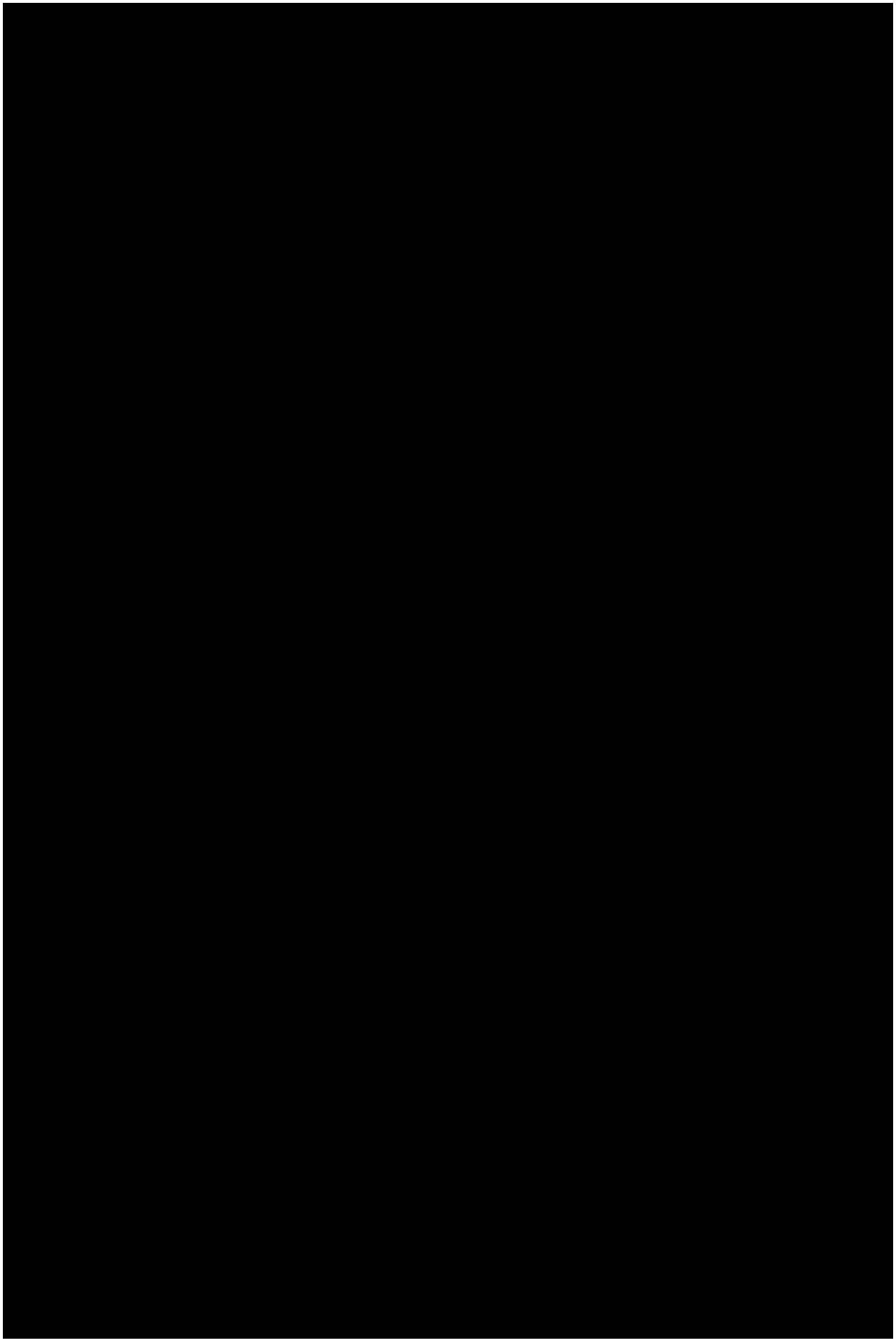


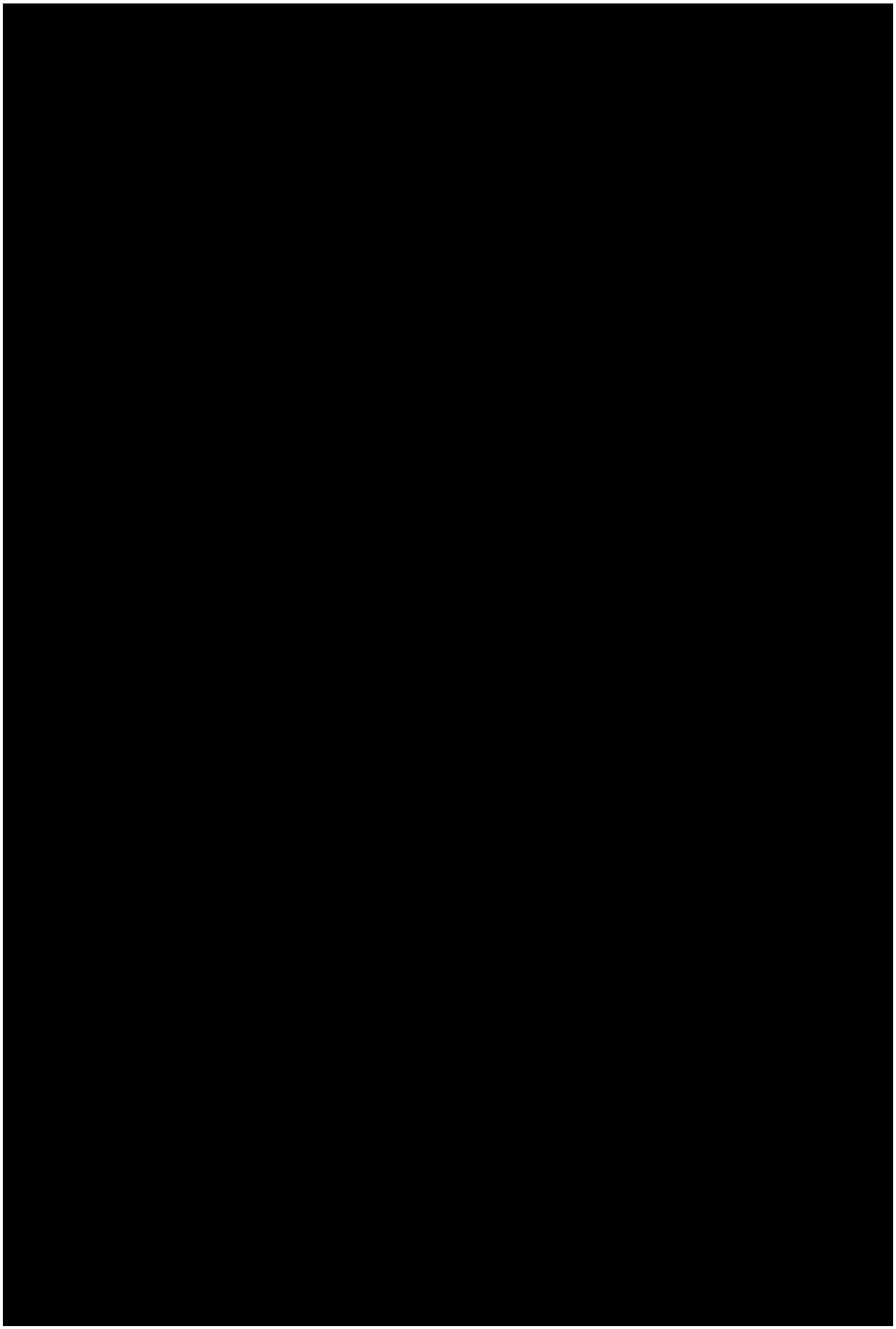


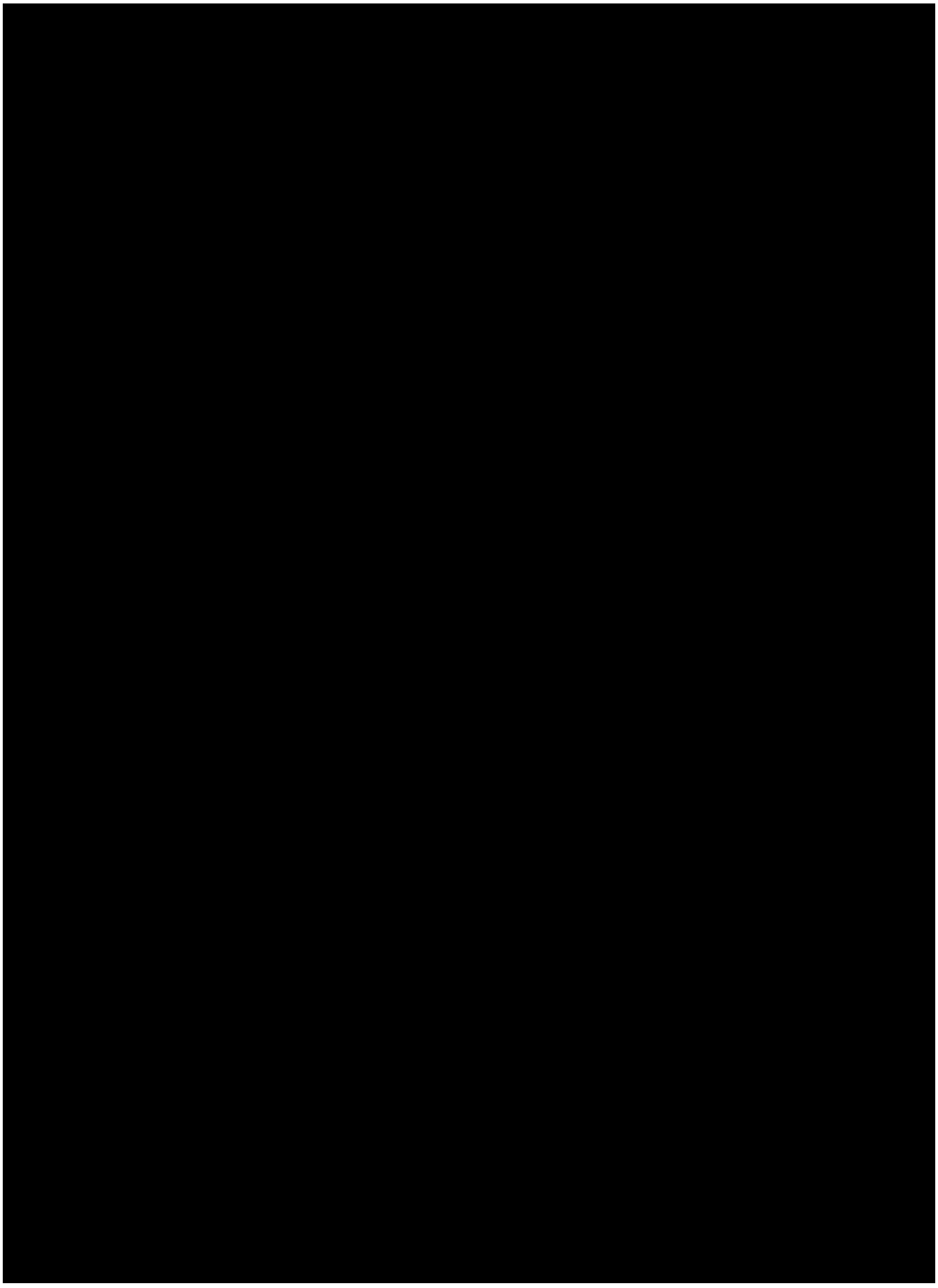


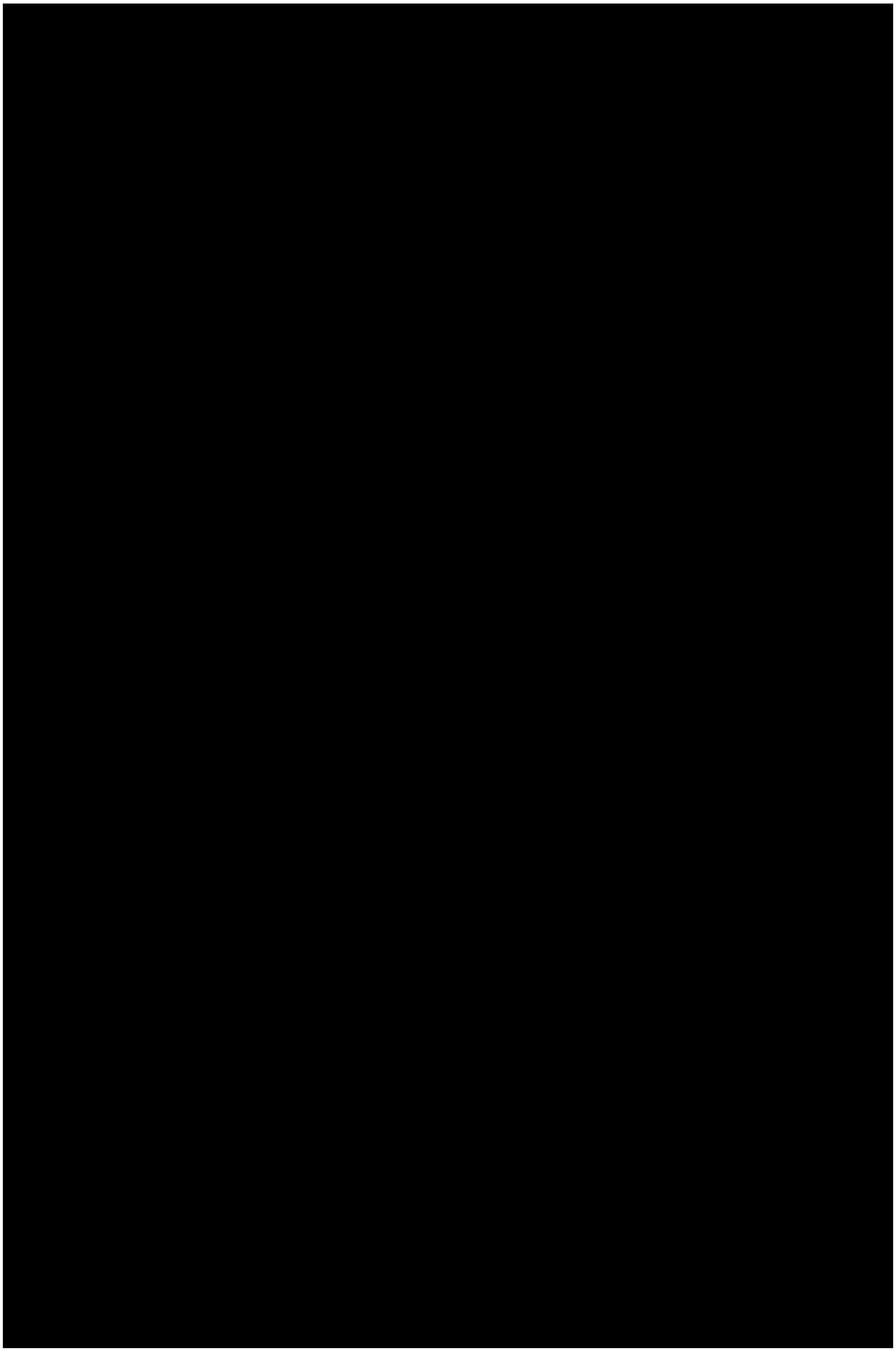


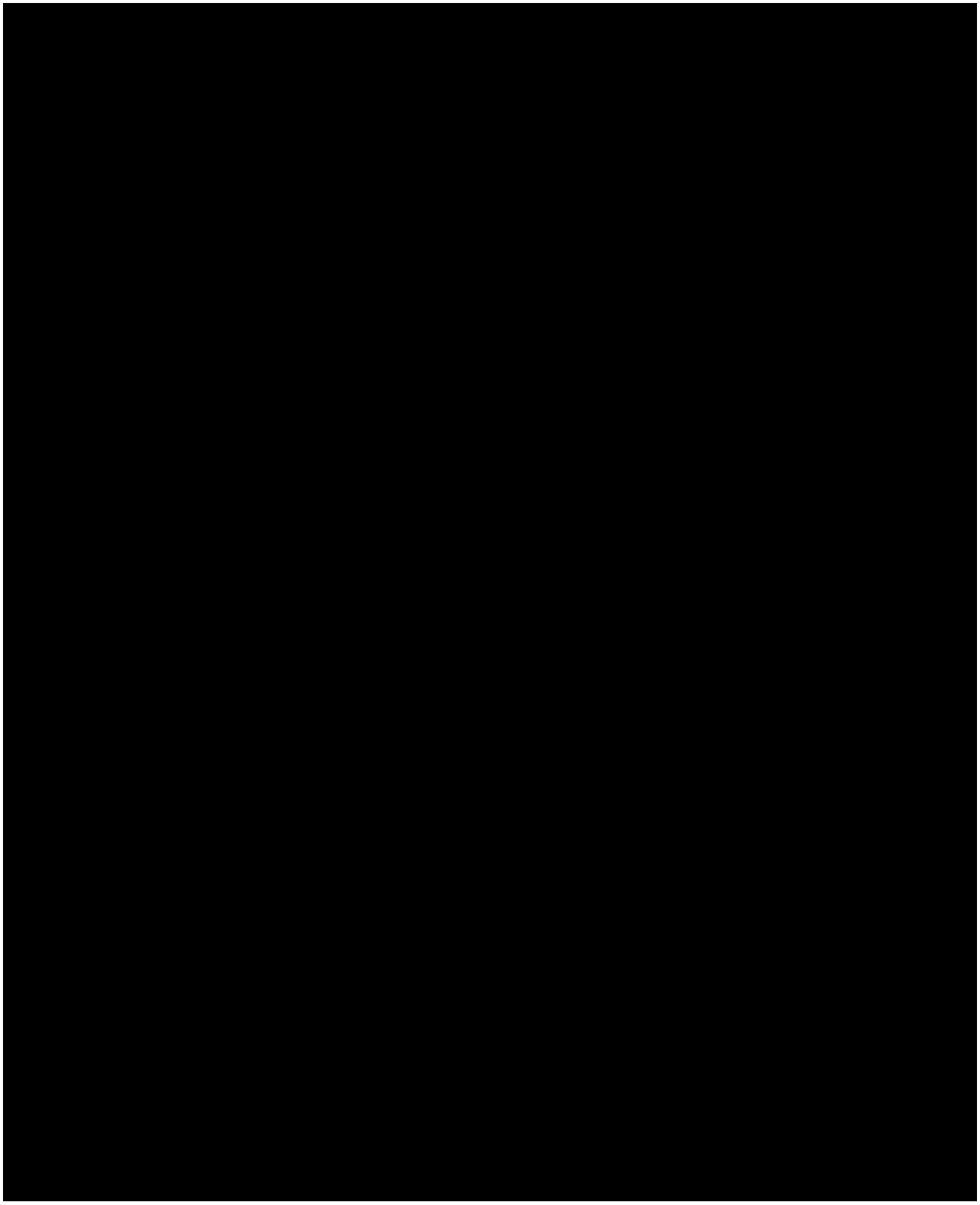


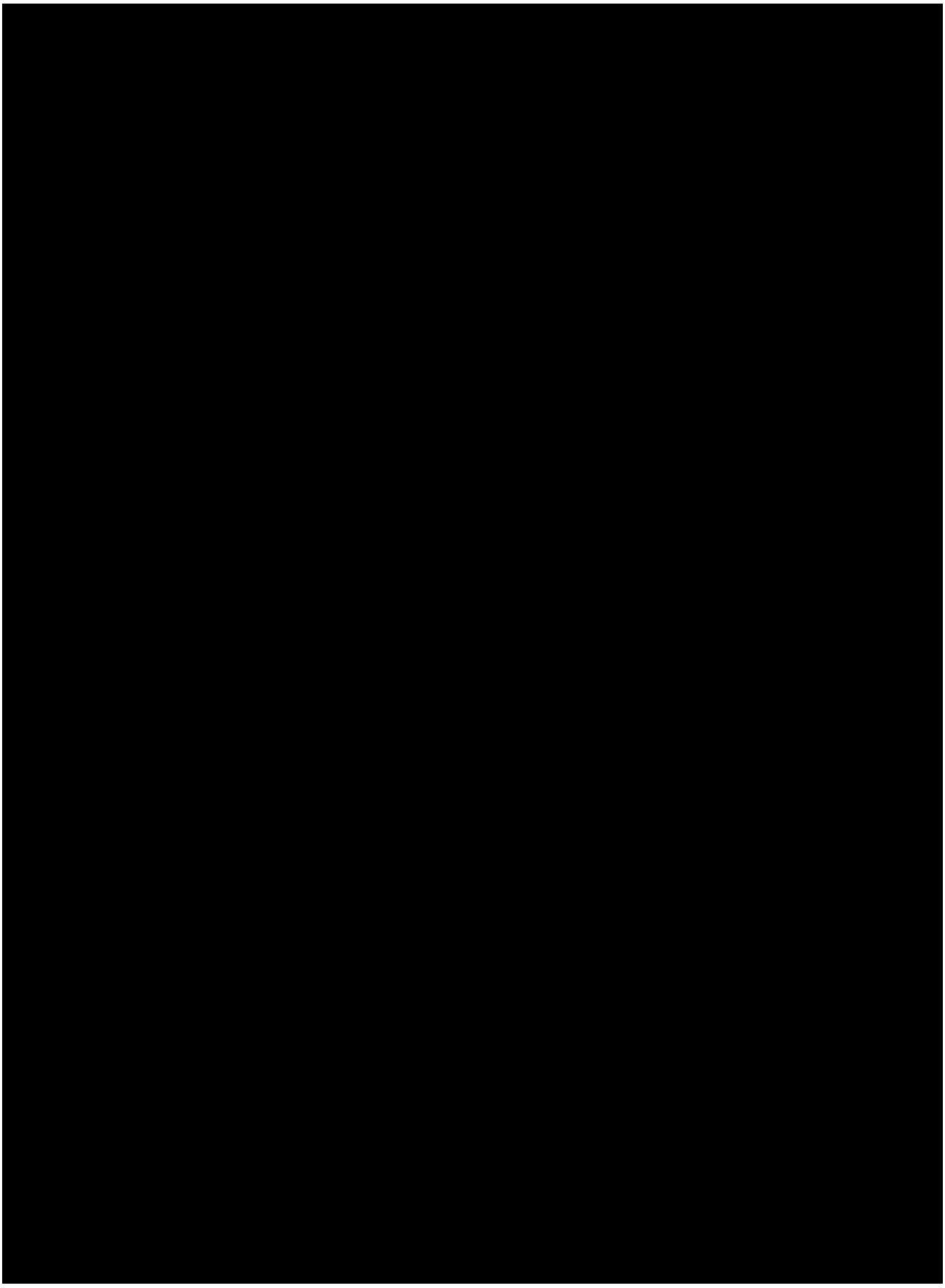


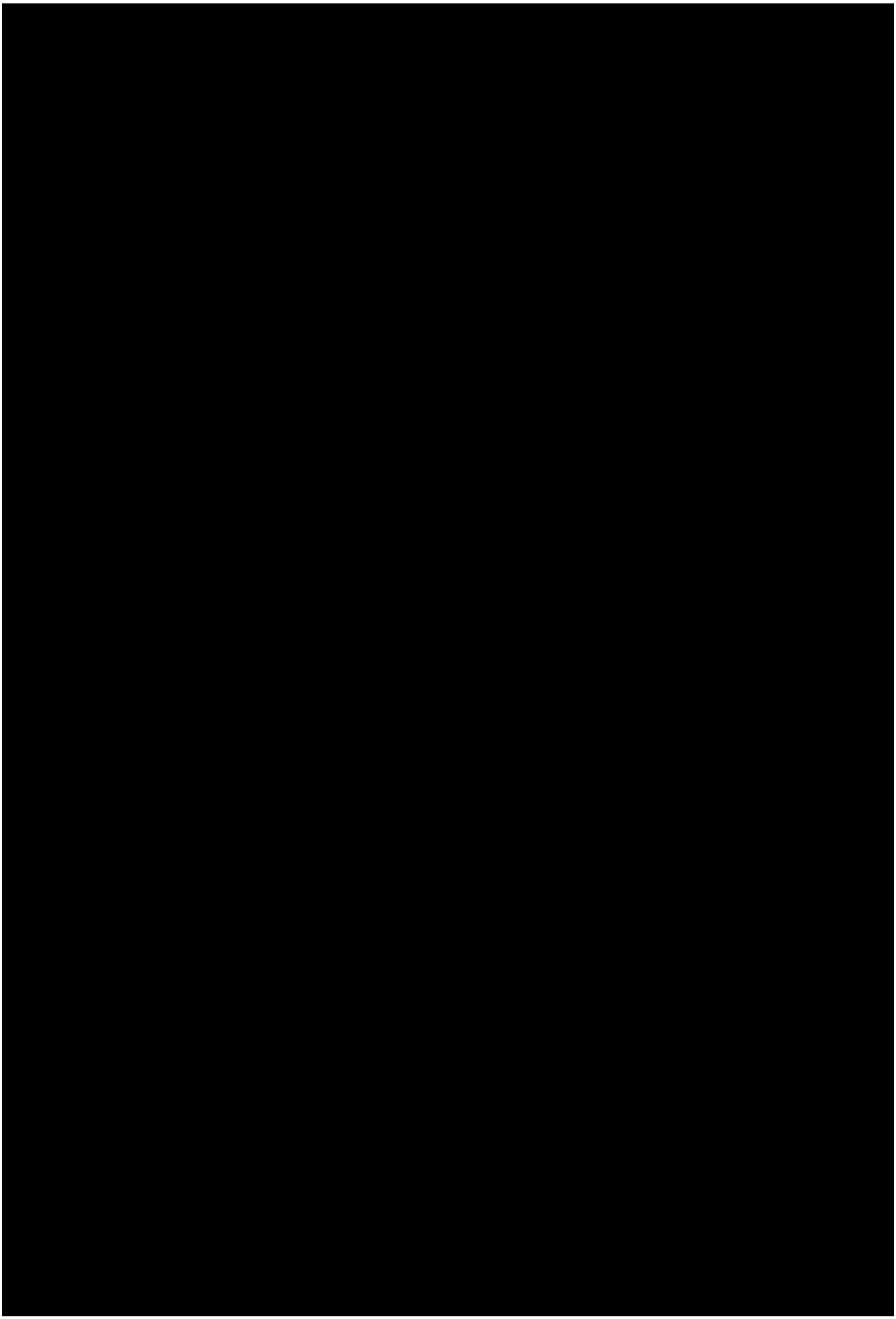


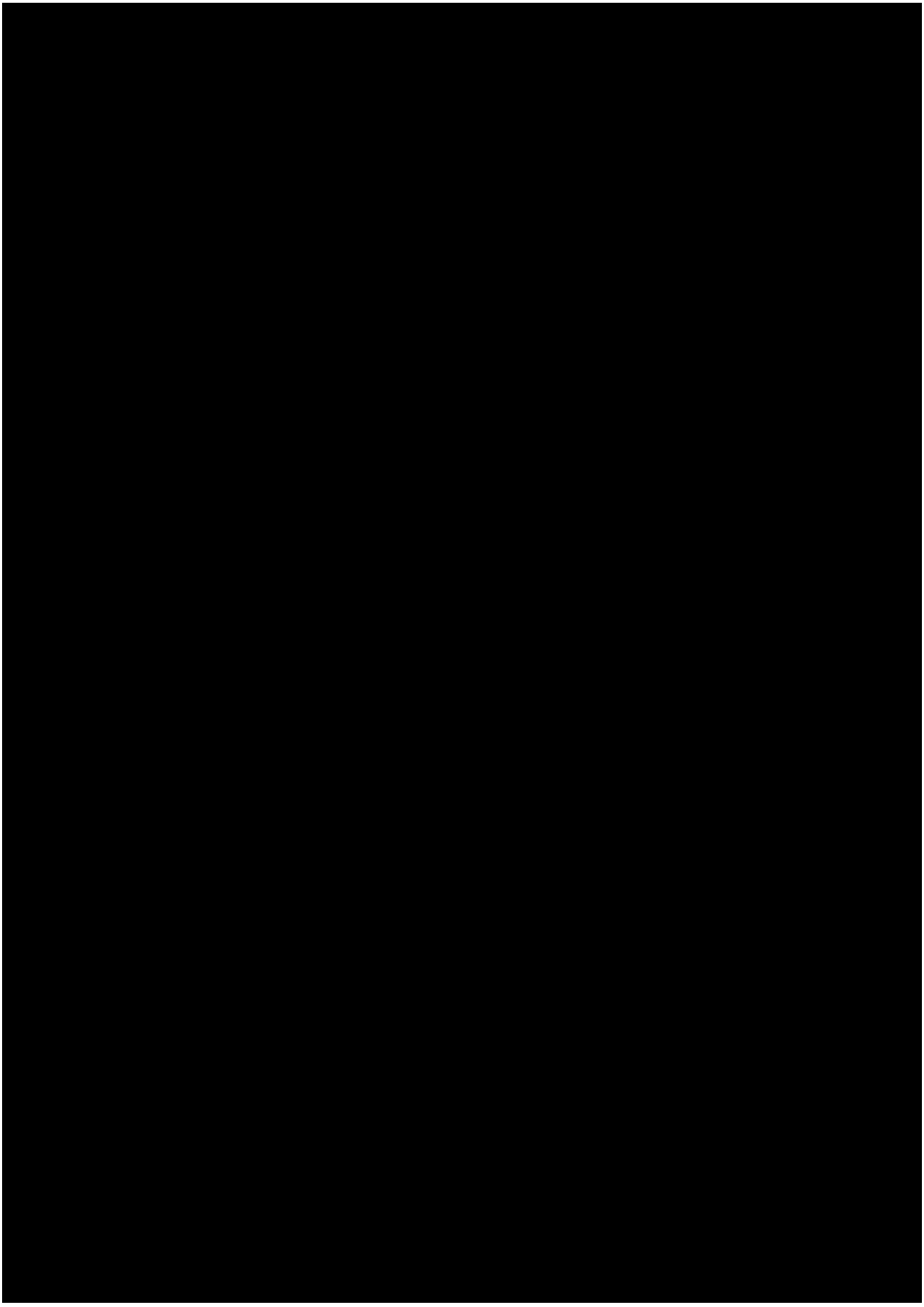


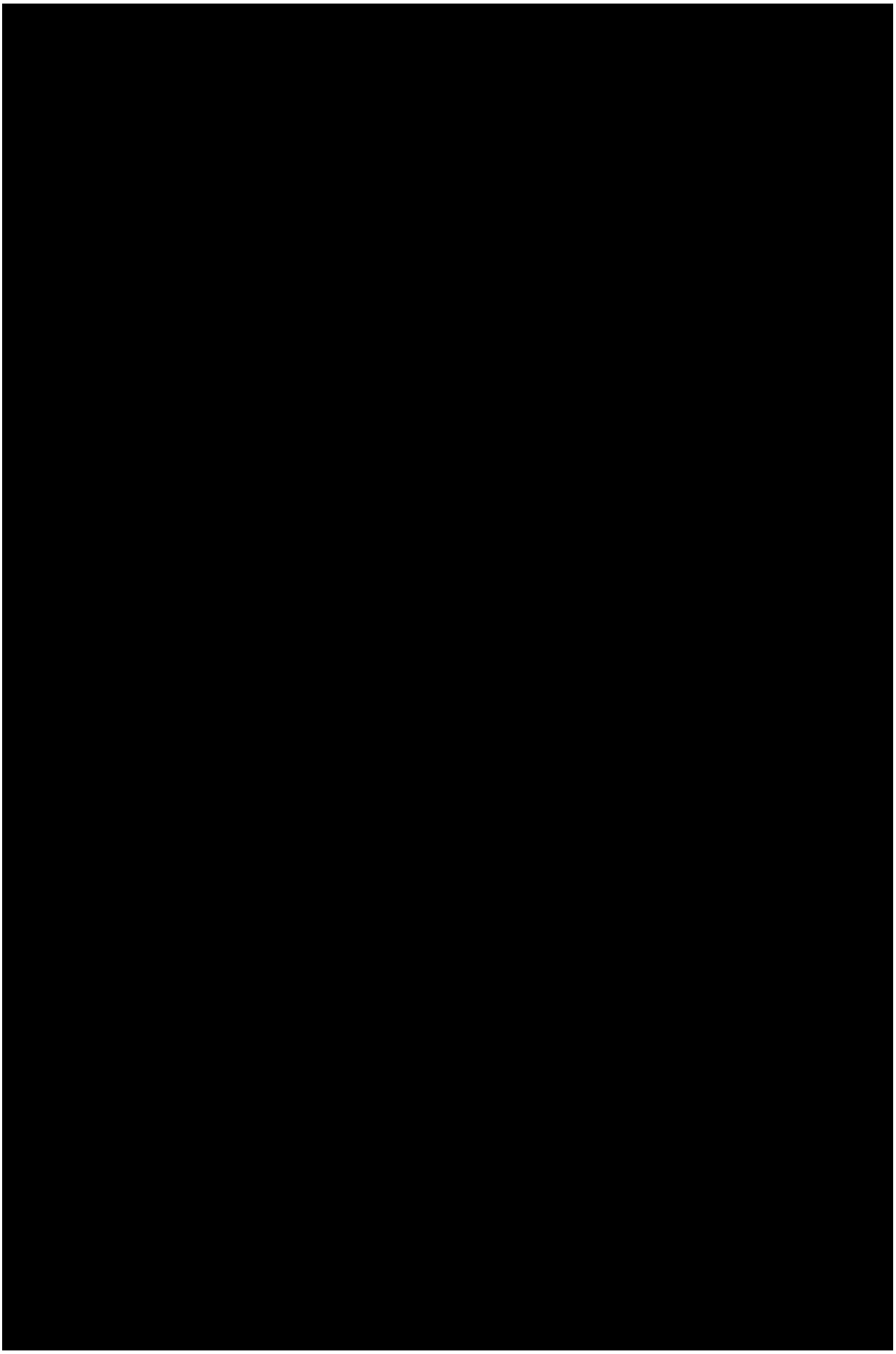


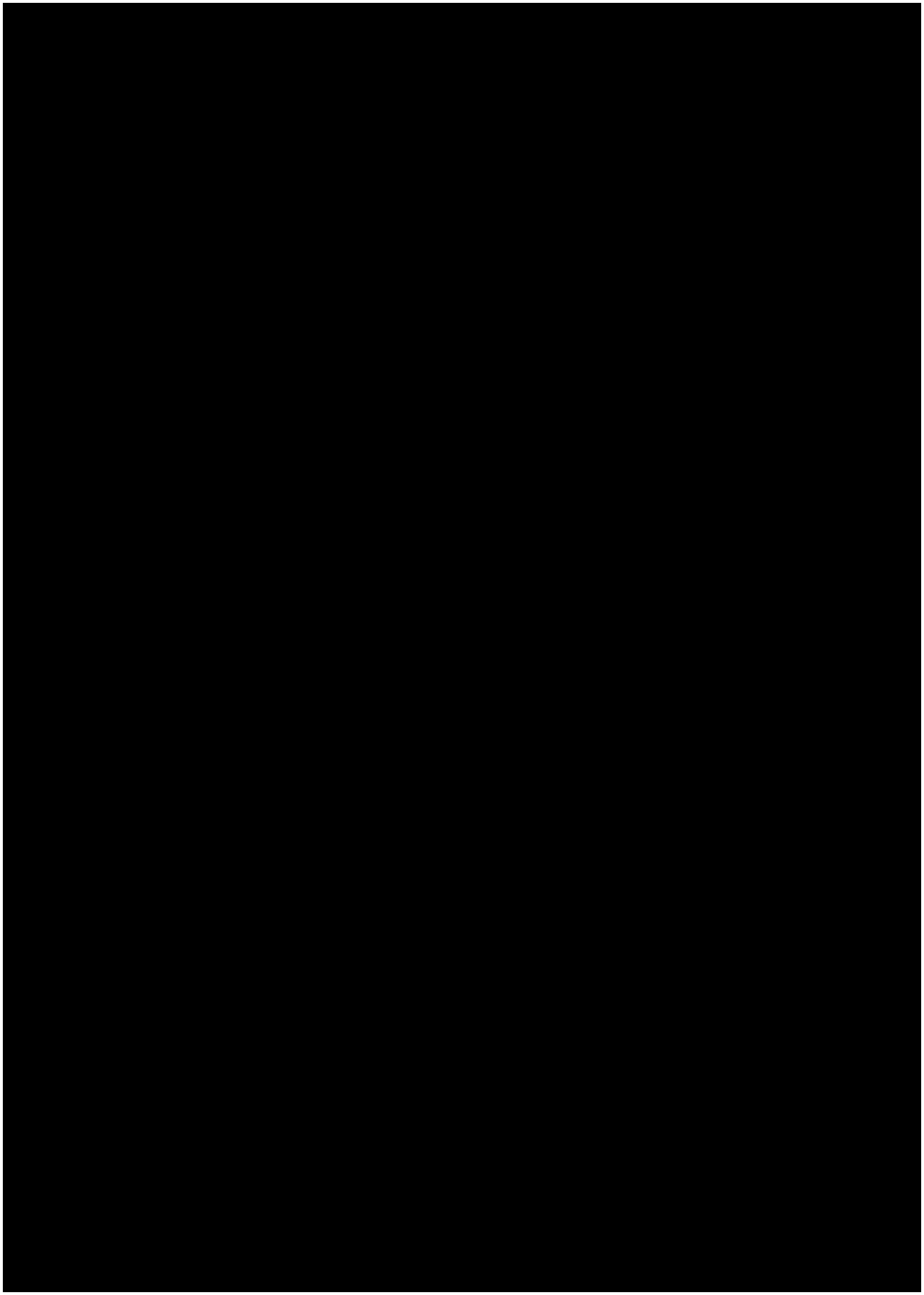


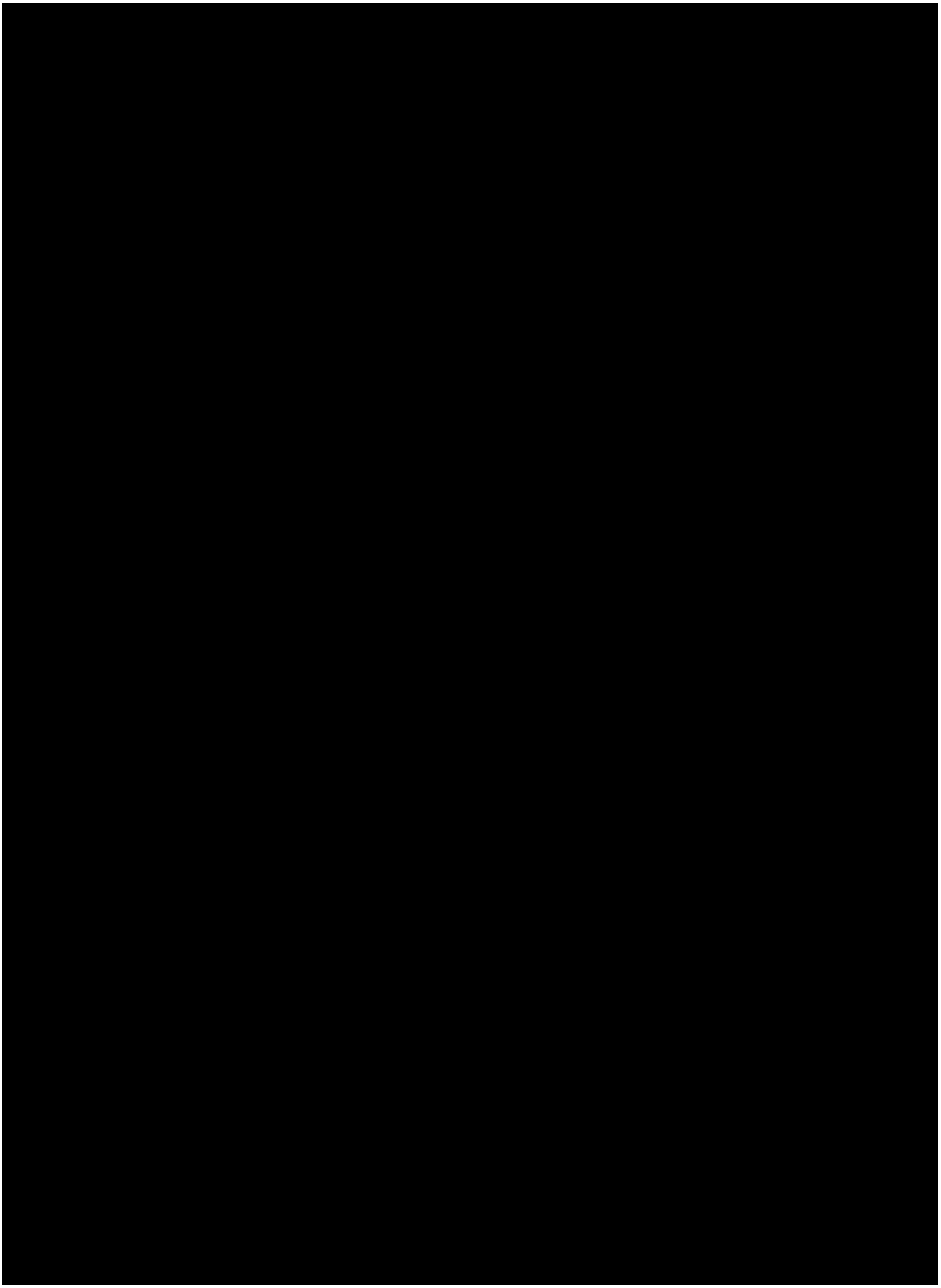


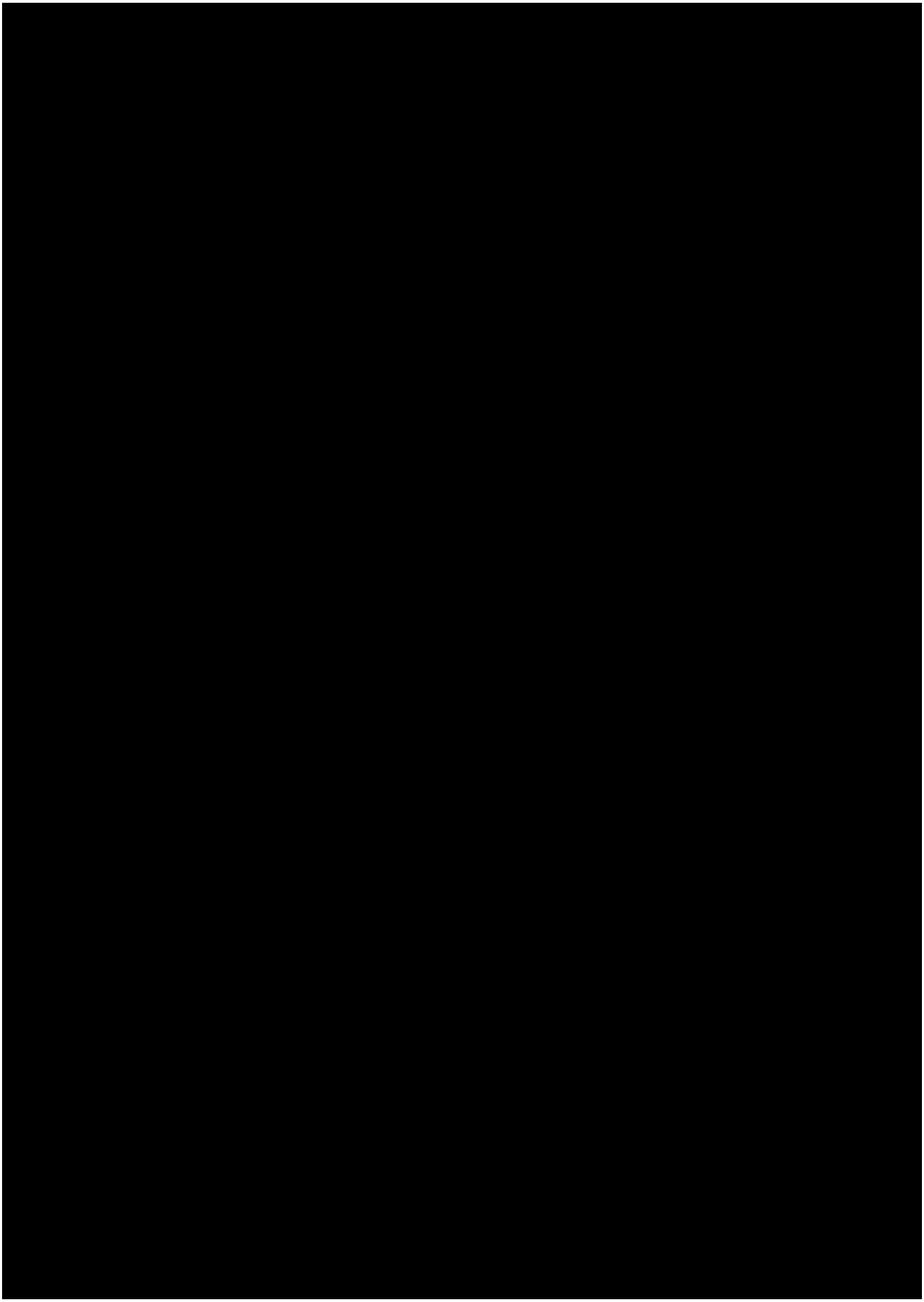


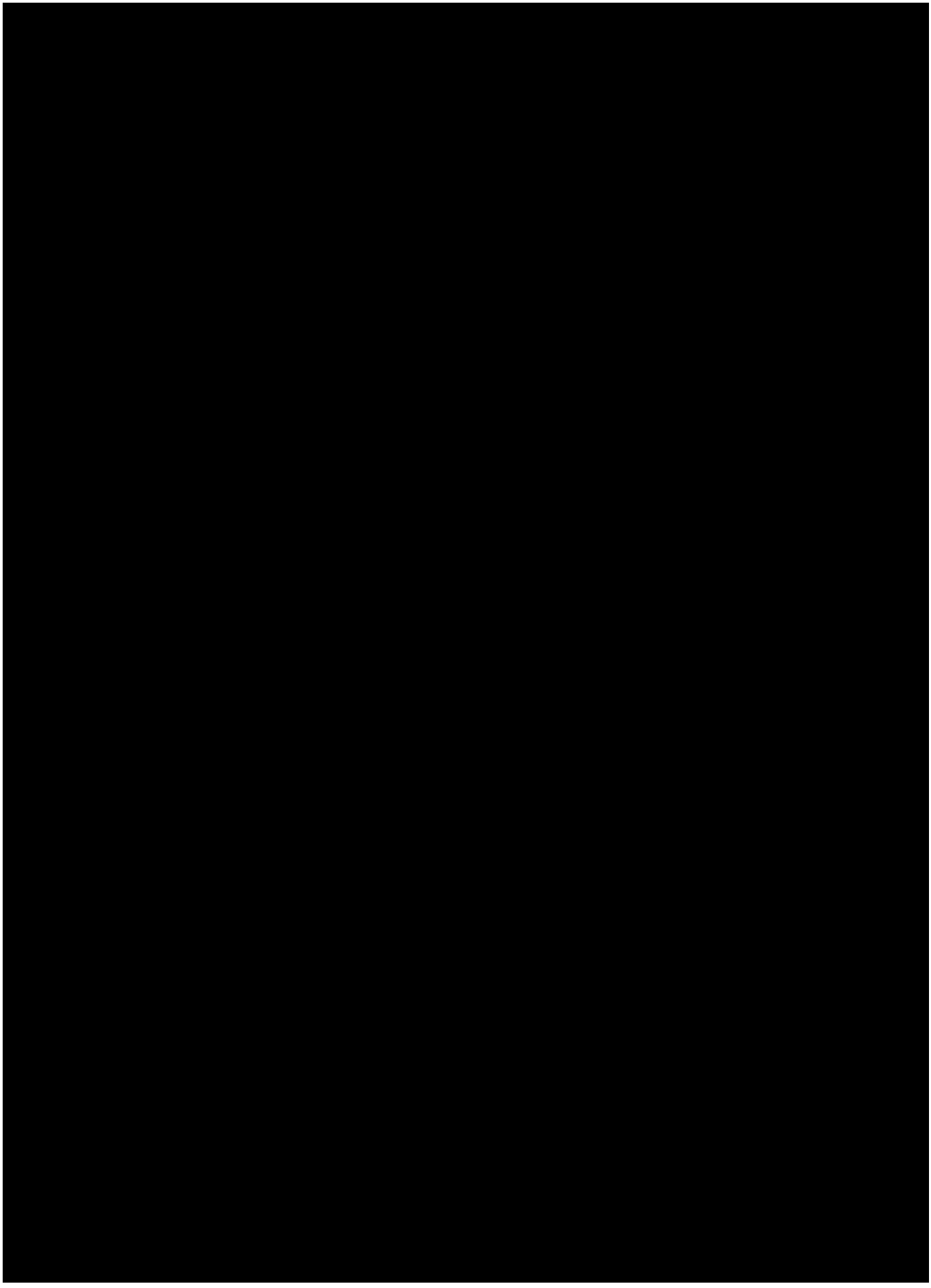


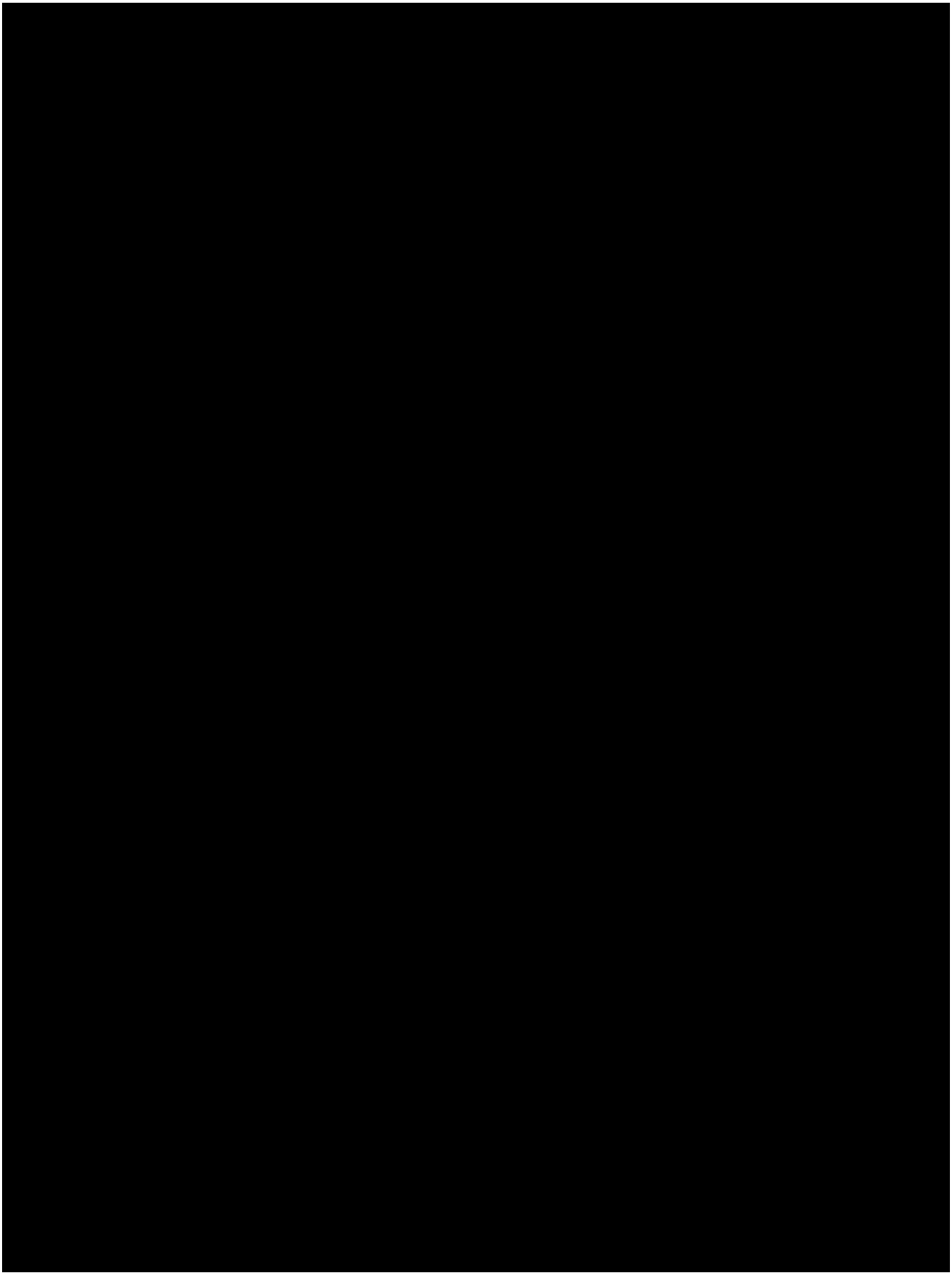


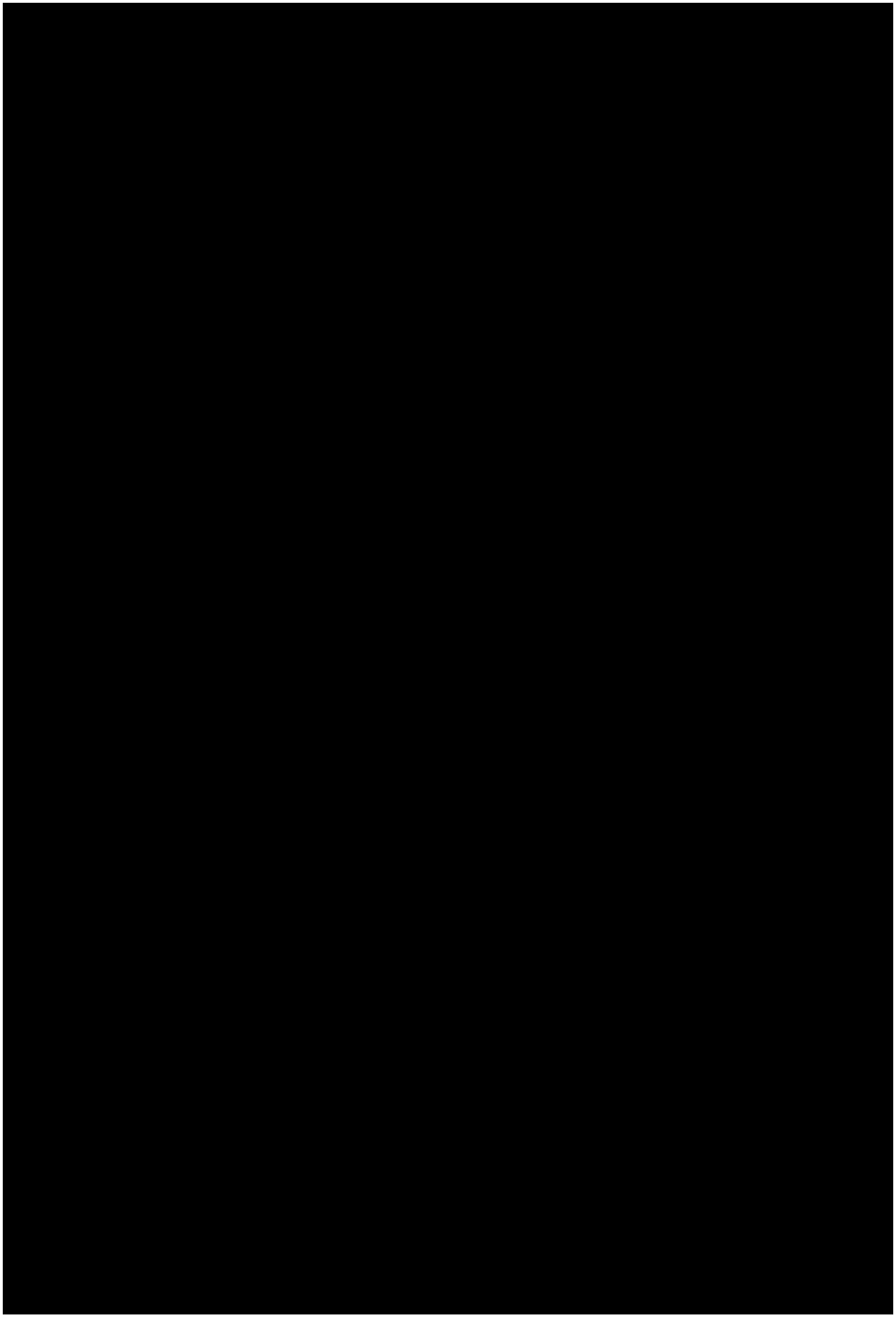


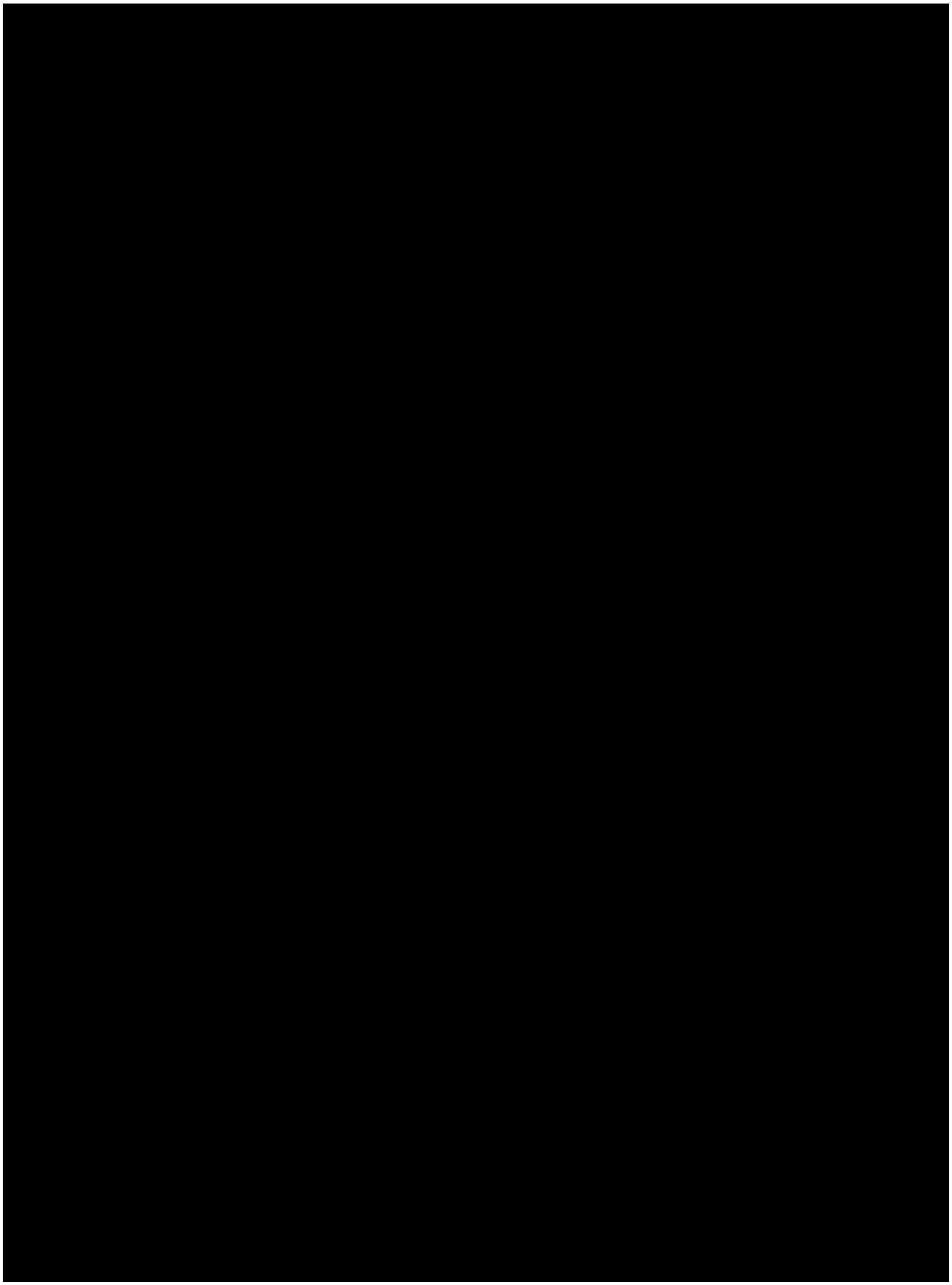


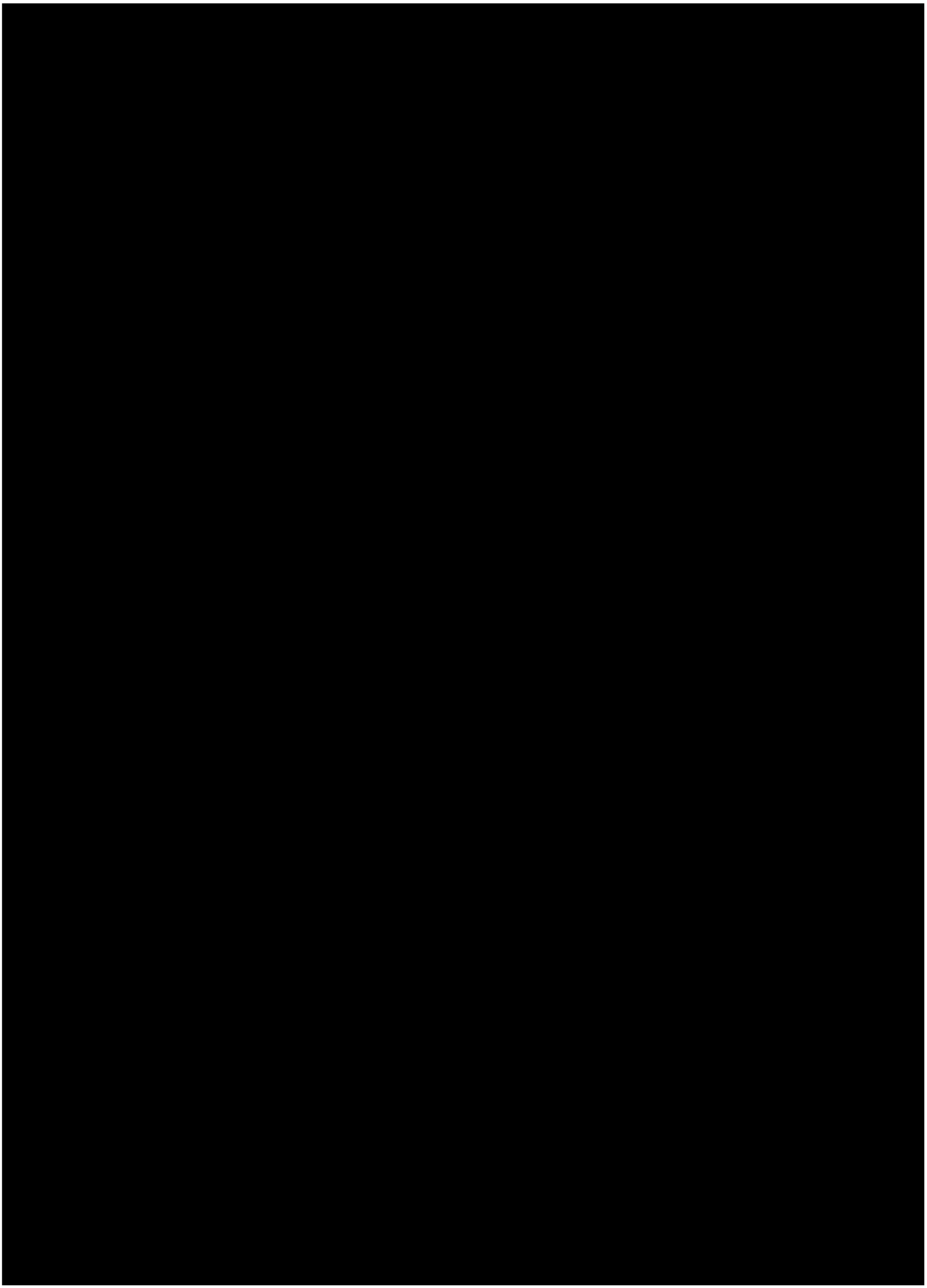


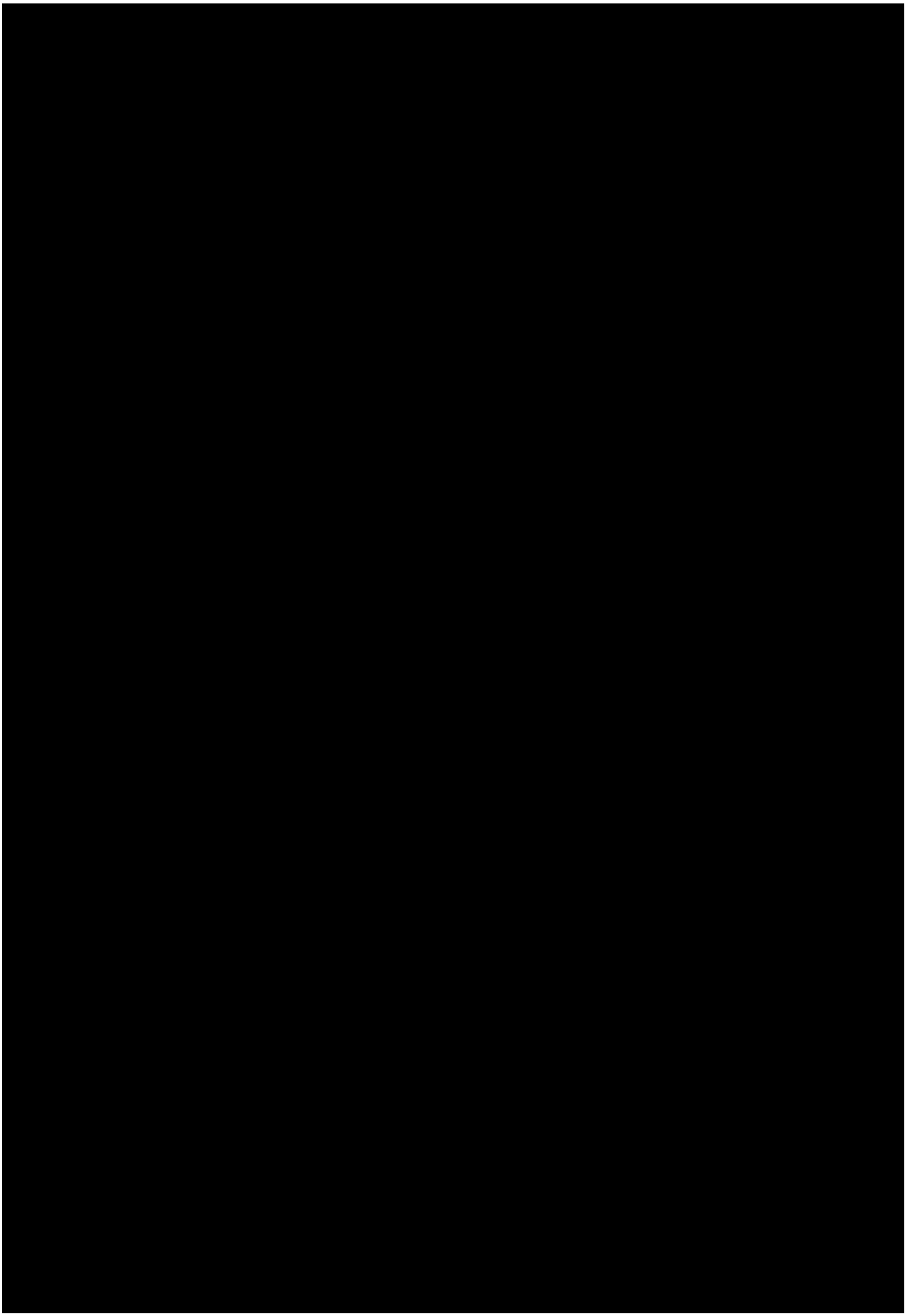


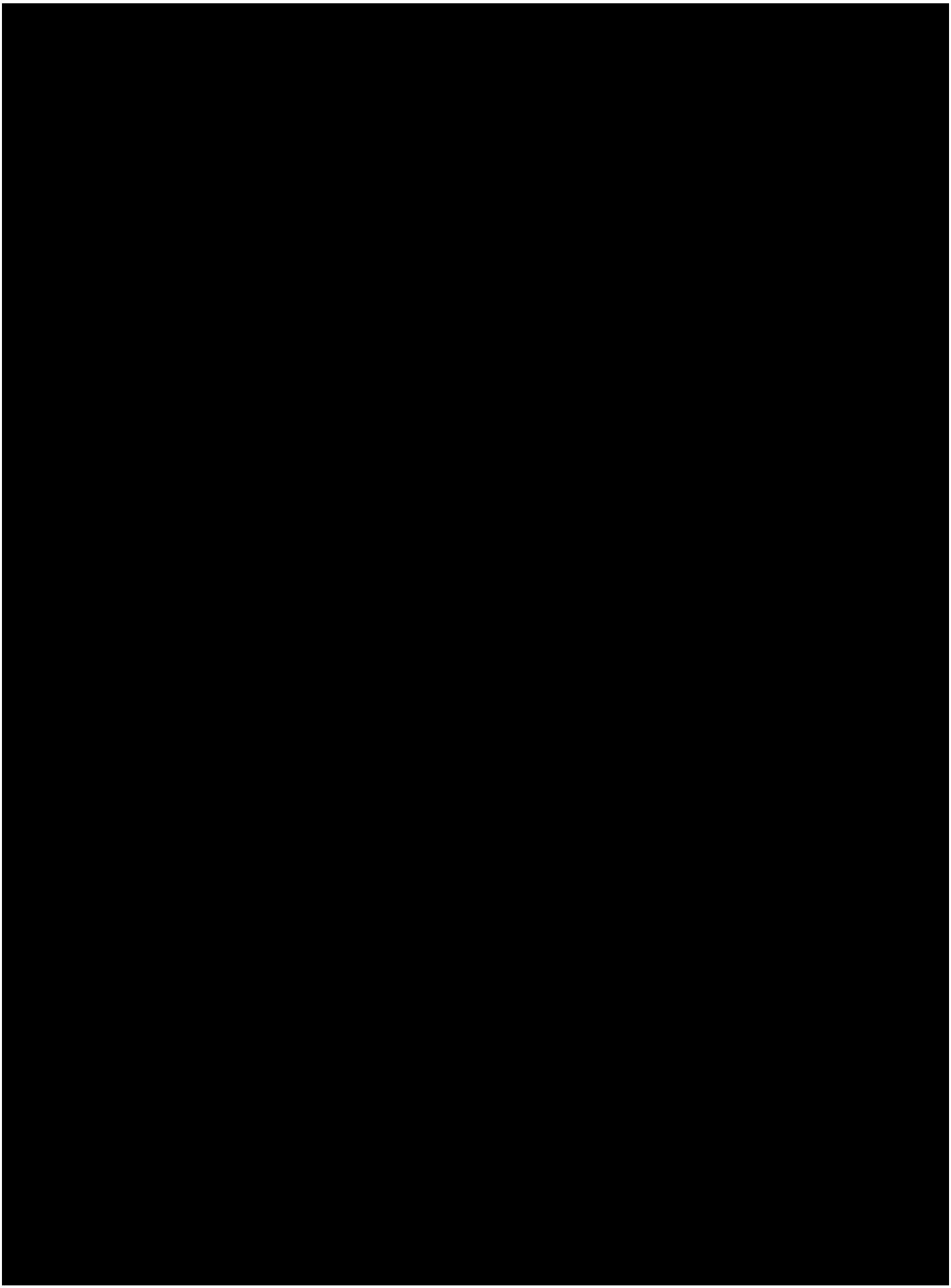


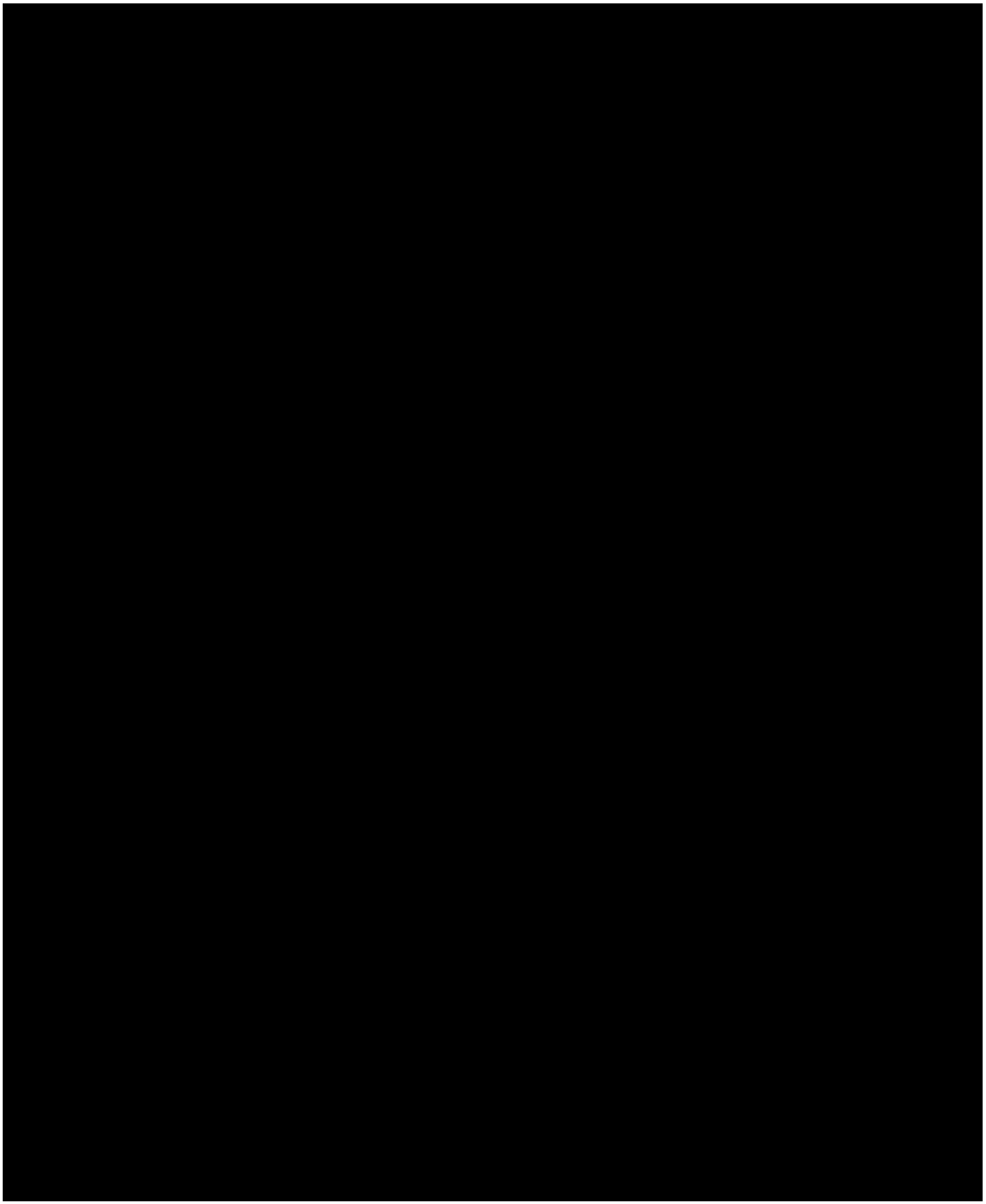


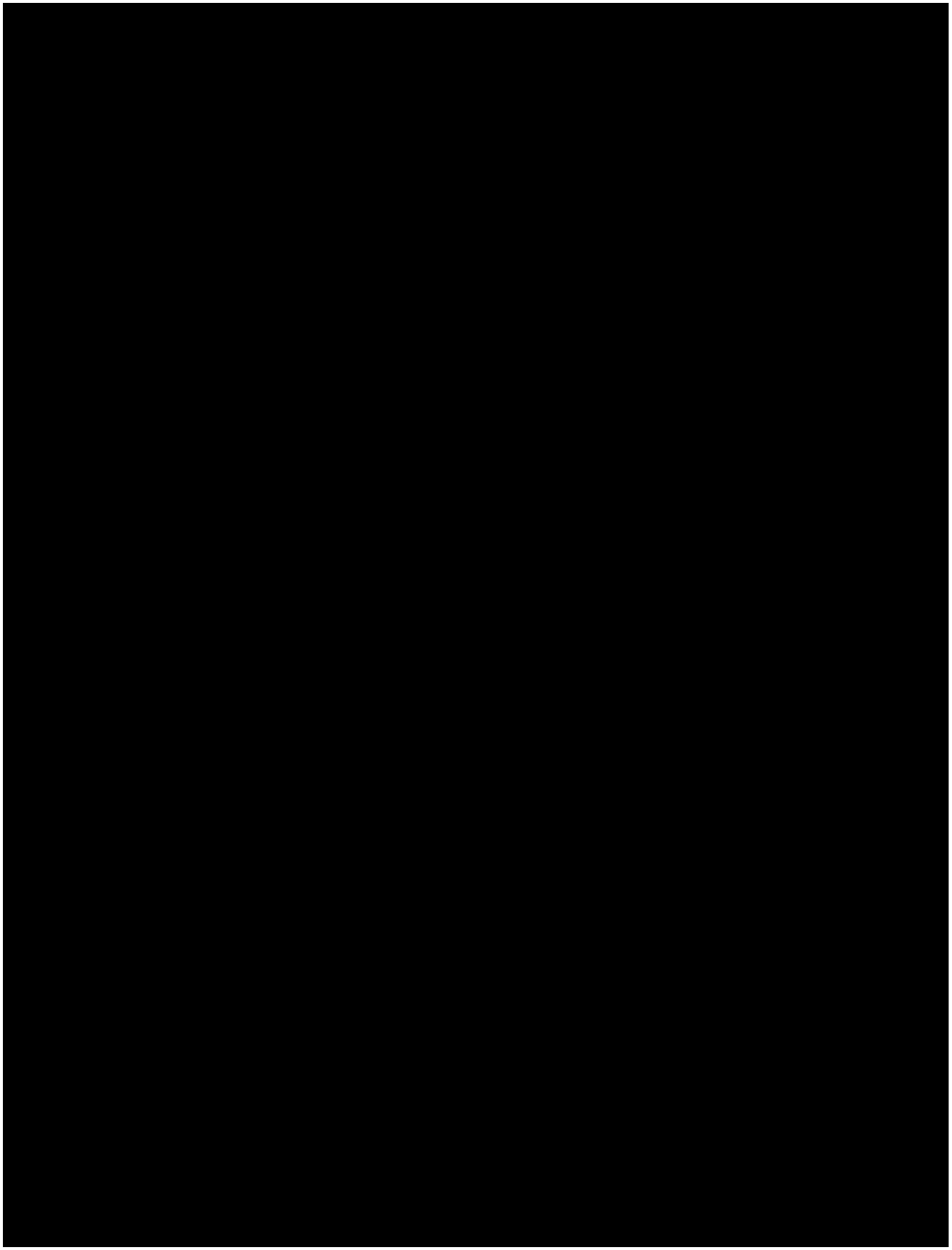


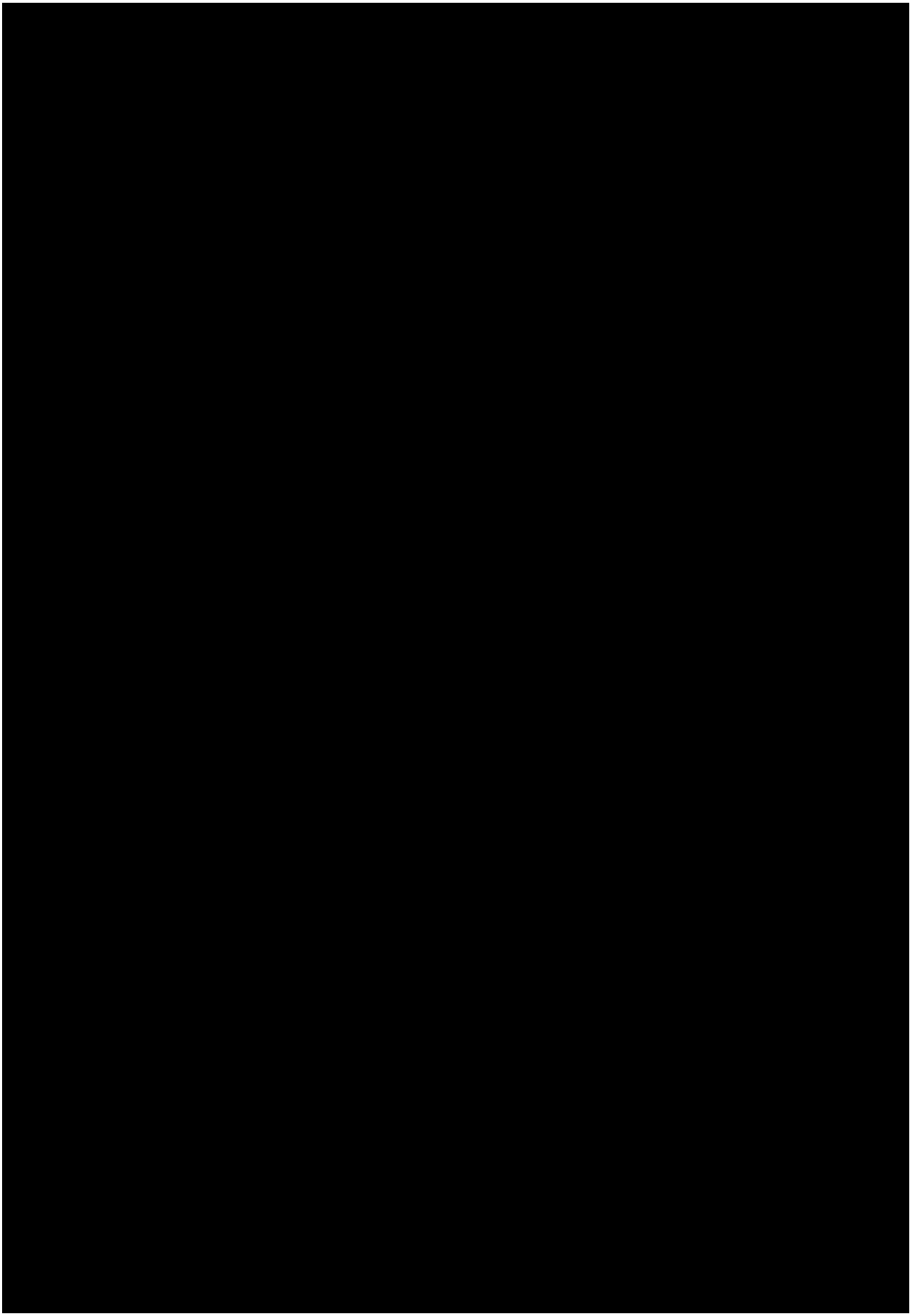


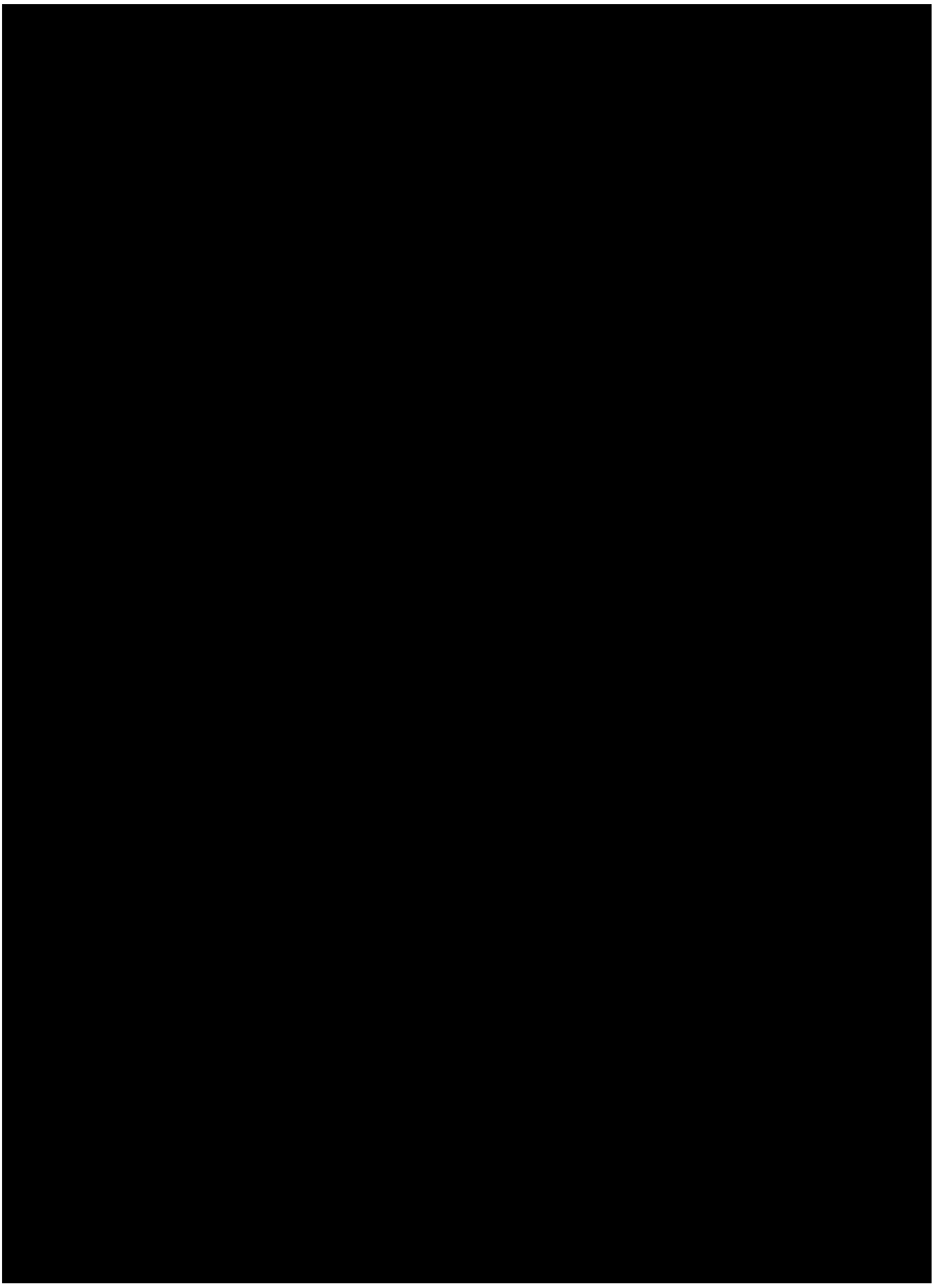


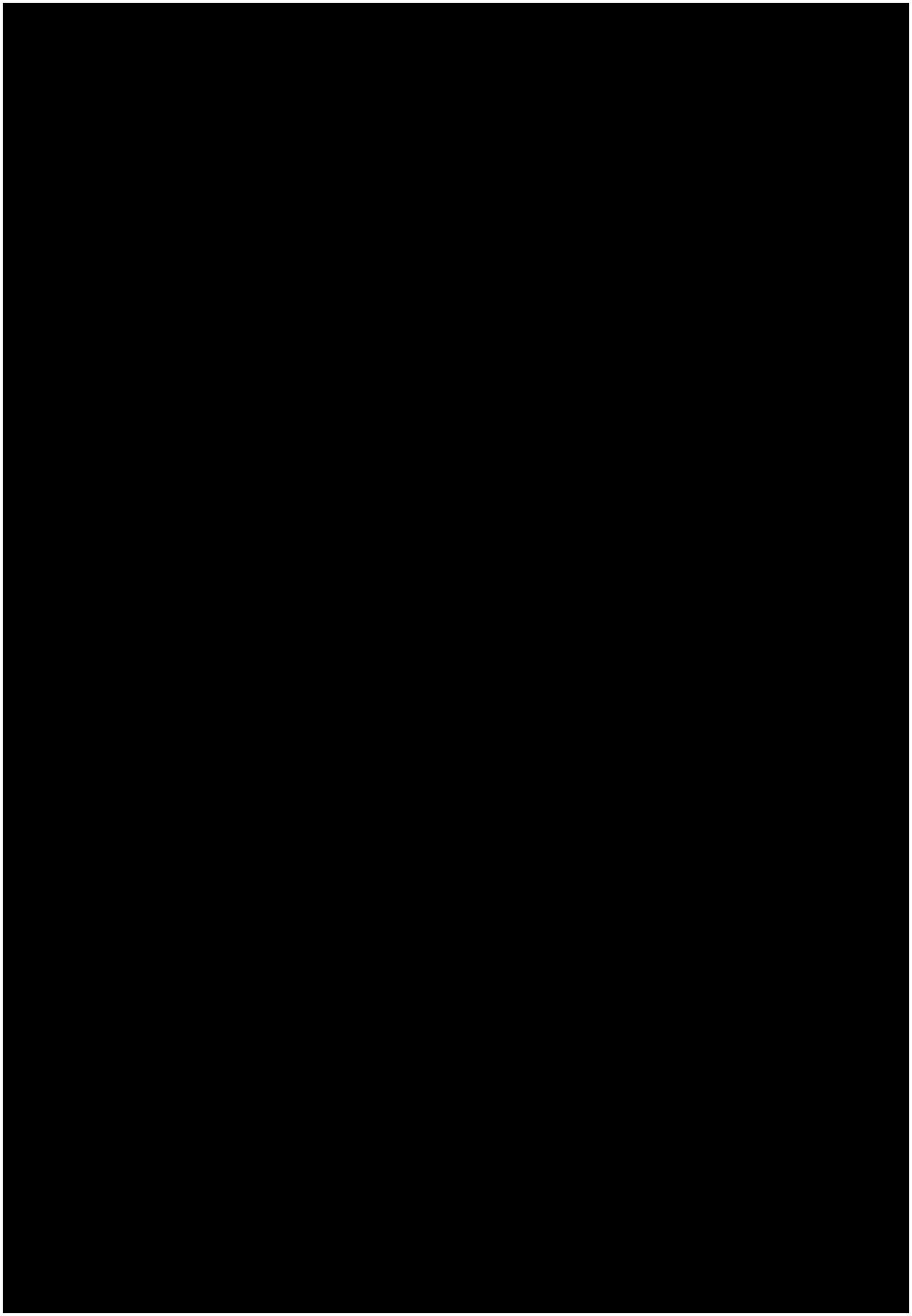


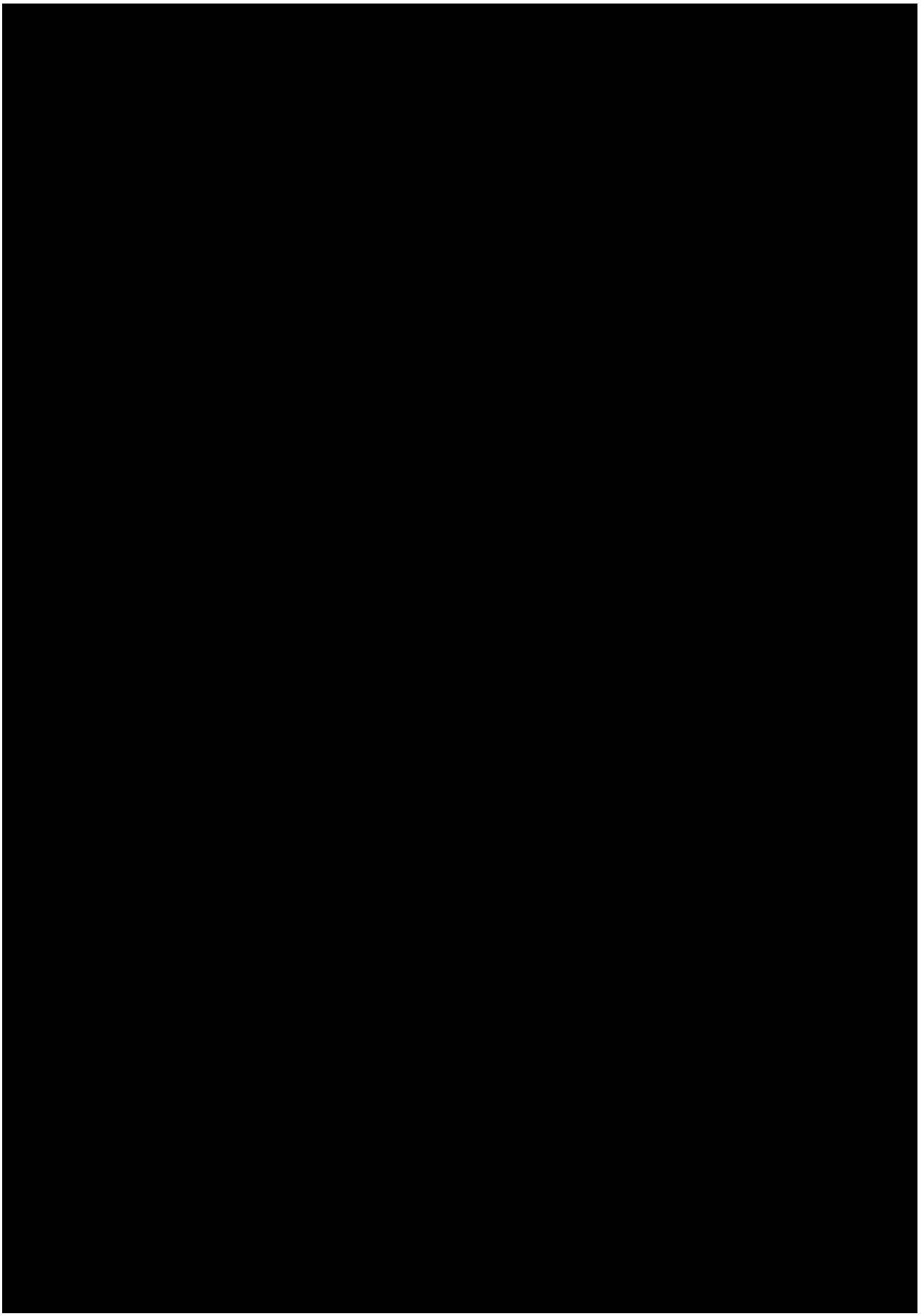


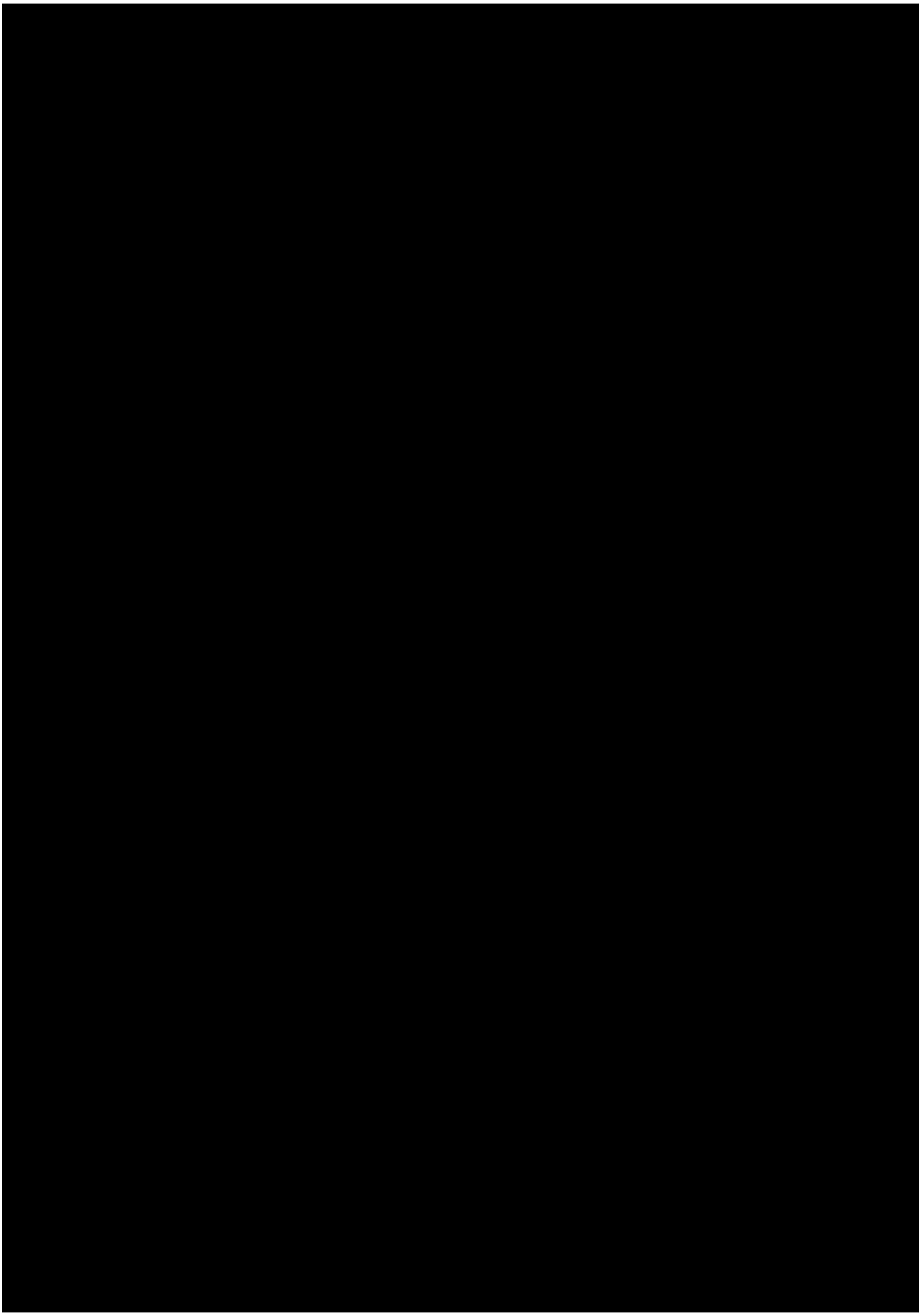


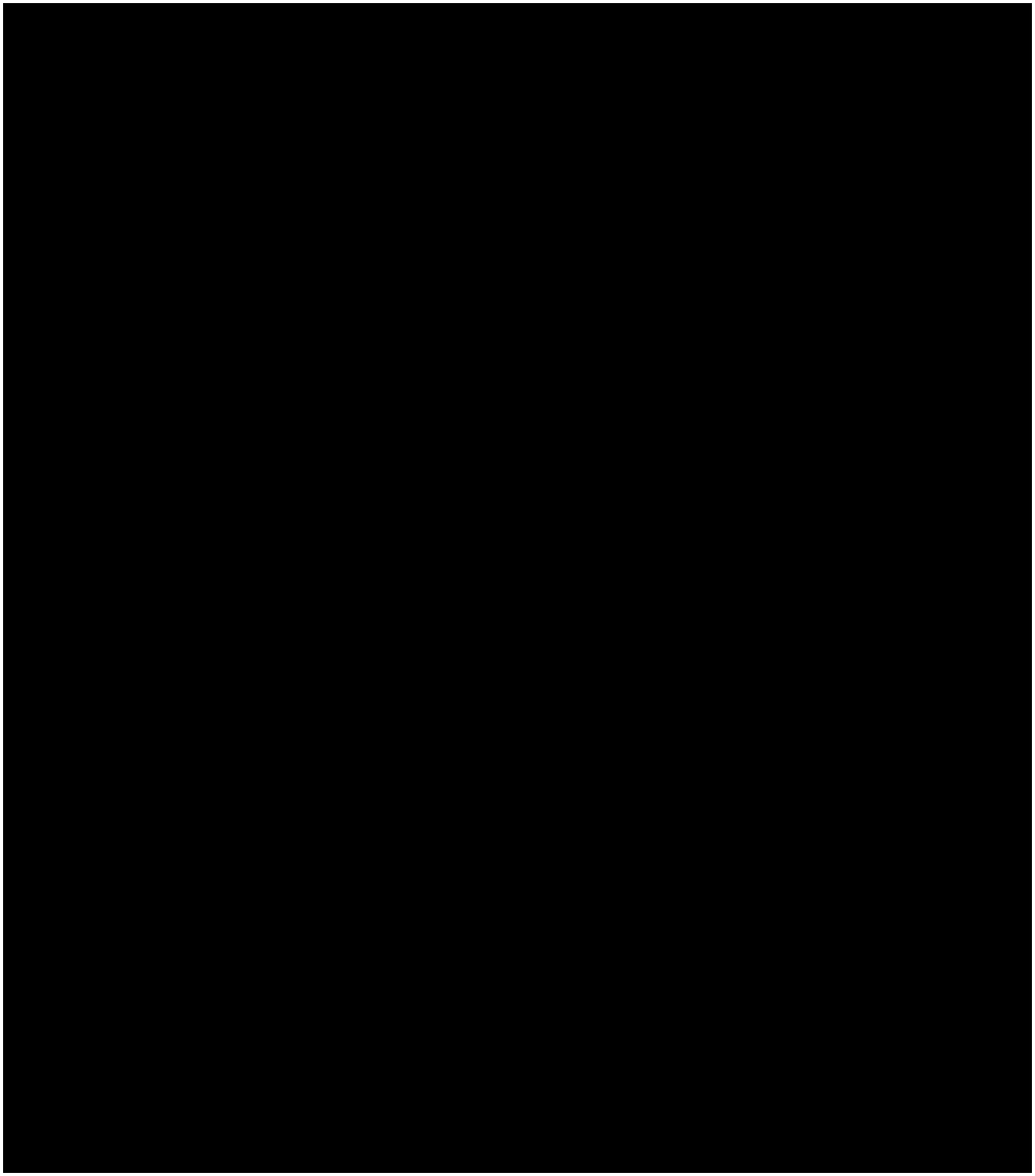












References

Abate, J. and Whitt, W. (1999). Modeling service–time distributions with non–exponential tails: beta mixtures of exponentials. *Stochastic Models*, 15(3):517–546.

Afèche, P., Araghi, M., and Baron, O. (2017). Customer acquisition, retention, and service access quality: Optimal advertising, capacity level, and capacity allocation. *Manufacturing & Service Operations Management*, 19(4):674–691.

Albrecher, H. and Asmussen, S. (2006). Ruin probabilities and aggregrate claims distributions for shot noise Cox processes. *Scandinavian Actuarial Journal*, 2006(2):86–110.

Anderson, T. and Darling, D. (1952). Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212.

Anderson, T. W. (1960). A modification of the sequential probability ratio test to reduce the sample size. *The Annals of Mathematical Statistics*, 31(1):165–197.

Arnold, T. A. and Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2):34–39.

Arslan, H., Graves, S. C., and Roemer, T. A. (2007). A single-product inventory model for multiple demand classes. *Management Science*, 53(9):1486–1500.

Artalejo, J. R. and Gómez-Corral, A. (2003). Channel idle periods in computer and telecommunication systems with customer retrials. *Telecommunication Systems*, 24(1):29–46.

Asmussen, S. and Albrecher, H. (2010). *Ruin probabilities*, volume 14. World scientific Singapore.

Ata, B. (2005). Dynamic power control in a wireless static channel subject to a quality-of-service constraint. *Operations Research*, 53(5):842–851.

Aurzada, F. and Dereich, S. (2013). Universality of the asymptotics of the one-sided exit problem for integrated processes. In *Annales de l'IHP Probabilités et statistiques*, volume 49, pages 236–251.

Avramidis, A. N., Deslauriers, A., and L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908.

Axsäter, S. (2003). A new decision rule for lateral transshipments in inventory systems. *Management Science*, 49(9):1168–1179.

Badila, E. (2015). Queues and risk models. *PhD Thesis*.

Badila, E., Boxma, O., and Resing, J. (2014). Queues and risk processes with dependencies. *Stochastic Models*, 30(3):390–419.

Bandi, C., Bertsimas, D., and Youssef, N. (2015). Robust queueing theory. *Operations Research*, 63(3):676–700.

Bernyk, V., Dalang, R. C., and Peskir, G. (2008). The law of the supremum of a stable Lévy process with no negative jumps. *The Annals of Probability*, 36(5):1777–1789.

Bertoin, J., Doney, R. A., and Maller, R. A. (2008). Passage of Lévy processes across power law boundaries at small times. *The Annals of Probability*, 36(1):160–197.

Bertsimas, D. J. and Nakazato, D. (1992). Transient and busy period analysis of the GI/G/1 queue: The method of stages. *Queueing Systems*, 10(3):153–184.

Bijvank, M. and Johansen, S. G. (2012). Periodic review lost-sales inventory models with compound Poisson demand and constant lead times of any length. *European Journal of Operational Research*, 220(1):106–114.

Bischoff, W., Hashorva, E., Hüsler, J., and Miller, F. (2003). Exact asymptotics for boundary crossings of the Brownian bridge with trend with application to the Kolmogorov test. *Annals of the Institute of Statistical Mathematics*, 55(4):849–864.

Borovkov, A. (1965). On the first passage time for one class of processes with independent increments. *Theory of Probability & Its Applications*, 10(2):331–334.

Borovkov, K. and Novikov, A. (2005). Explicit bounds for approximation rates of boundary crossing probabilities for the Wiener process. *Journal of Applied Probability*, 42(1):82–92.

Borst, S. C., Boxma, O. J., and Combé, M. (1993). An M/G/1 queue with dependence between interarrival and service times. *Stochastic Models*, 9:341–371.

Boxma, O., Essifi, R., and Janssen, A. J. (2016). A queueing/inventory and an insurance risk model. *Advances in Applied Probability*, 48(4):1139–1160.

Boxma, O. J. and Cohen, J. (1998). The M/G/1 queue with heavy-tailed service time distribution. *IEEE journal on selected areas in communications*, 16(5):749–763.

Boxma, O. J. and Cohen, J. W. (1999). Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions. *Queueing systems*, 33(1-3):177–204.

Brigham, G. (1955). On a congestion problem in an aircraft factory. *Journal of the Operations Research Society of America*, 3(4):412–428.

Brown, J. R. and Harvey, M. E. (2007). Rational arithmetic Mathematica functions to evaluate the one-sided one sample K-S cumulative sampling distribution. *Journal of Statistical Software*, 19(6):1–32.

Brown, J. R. and Harvey, M. E. (2008). Rational arithmetic Mathematica functions to evaluate the two-sided one sample K-S cumulative sampling distribution. *Journal of Statistical Software*, 26(2):1–40.

Buonocore, A., Giorno, V., Nobile, A. G., and Ricciardi, L. M. (1990). On the two-boundary first-crossing-time problem for diffusion processes. *Journal of Applied Probability*, 27(1):102–114.

Burke, P. (1975). Delays in single-server queues with batch input. *Operations Research*, 23(4):830–833.

Cai, N., Chen, N., and Wan, X. (2009). Pricing double-barrier options under a flexible jump diffusion model. *Operations Research Letters*, 37(3):163–167.

Calabrese, R. and Zenga, M. (2010). Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking and Finance*, 34(5):903–911.

Carnal, H. (1962). Sur les théorèmes de Kolmogorov et Smirnov dans le cas d'une distribution discontinue. *Commentarii Mathematici Helvetici*, 37(1):19–35.

Carvalho, L. (2015). An improved evaluation of Kolmogorov's distribution. *Journal of Statistical Software*, 65(3):1–7.

Channouf, N. and L'Ecuyer, P. (2012). A normal copula model for the arrival process in a call center. *International Transactions in Operational Research*, 19(6):771–787.

Chen, G. K. C. and Winters, P. R. (1966). Forecasting peak demand for an electric utility with a hybrid exponential model. *Management Science*, 12(12):531–537.

Choudhury, G. L., Lucantoni, D. M., Whitt, W., et al. (1994). Multidimensional transform inversion with applications to the transient M/G/1 queue. *The Annals of Applied Probability*, 4(3):719–740.

CIESIN (2012). National aggregates of geospatial data collection: Population, landscape, and climate estimates, version 3 (place iii).

Cohen, J. W. (1982). *The single server queue*. North-Holland Publishing Company, Amsterdam.

Combé, M. B. and Boxma, O. J. (1998). BMAP modelling of a correlated queue. In Walrand, J., Bagchi, K., and Zobrist, G., editors, *Network performance modeling and simulation*, pages 177–196. Gordon and Breach Science Publishers: Amsterdam.

Conover, W. J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association*, 67(339):591–596.

Cox, D. and Isham, V. (1986). The virtual waiting-time and related processes. *Advances in Applied Probability*, 18(2):558–573.

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, 23(3):289–303.

Crump, K. S. (1975). On point processes having an order statistic structure. *Sankhyā: The Indian Journal of Statistics, Series A*, 37(3):396–404.

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes. Volume II: General theory and structure*. Springer Science & Business Media.

Dassios, A. and Jang, J. W. (2003). Pricing of catastrophe reinsurance and derivatives using the Cox process with shot noise intensity. *Finance and Stochastics*, 7(1):73–95.

Dassios, A., Jang, J. W., and Zhao, H. (2015). A risk model with renewal shot-noise Cox process. *Insurance: Mathematics and Economics*, 65:55–65.

de Kok, T., Grob, C., Laumanns, M., Minner, S., Rambau, J., and Schade, K. (2018). A typology and literature review on stochastic multi-echelon inventory models. *European Journal of Operational Research*, 269(3):955–983.

Dimitrova, D., Ignatov, Z., and Kaishev, V. (2017). On the first crossing of two boundaries by an order statistics risk process. *Risks*, 5(3):43.

Dimitrova, D., Kaishev, V., and Tan, S. (2019a). Computing the Kolmogorov-Smirnov distribution when the underlying cdf is purely discrete, mixed or continuous. *Journal of Statistical Software*, Forthcoming.

Dimitrova, D. S., Ignatov, Z. G., and Kaishev, V. K. (2019b). Ruin and deficit under claim arrivals with the order statistics property. *Methodology and Computing in Applied Probability*, 21:511–530.

Dimitrova, D. S., Ignatov, Z. G., Kaishev, V. K., and Tan, S. (2019c). On double–boundary non–crossing probability for a class of compound processes with applications. *European Journal of Operational Research*, Forthcoming.

Dimitrova, D. S., Ignatov, Z. G., Kaishev, V. K., and Tan, S. (2019d). On double boundary crossing and the overshoot: Applications in queueing, ruin and inventory. *Submitted*.

Dimitrova, D. S., Kaishev, V. K., and Tan, S. (2019e). On a single server queueing model and its double boundary crossing duality. *Submitted*.

Dimitrova, D. S., Kaishev, V. K., and Zhao, S. (2015). Modeling finite-time failure probabilities in risk analysis applications. *Risk Analysis*, 35(10):1919–1939.

Dimitrova, D. S., Kaishev, V. K., and Zhao, S. (2016). On the evaluation of finite-time ruin probabilities in a dependent risk model. *Applied Mathematics and Computation*, 275:268–286.

Doney, R. (1991). Hitting probabilities for spectrally positive Lévy processes. *Journal of the London Mathematical Society*, 2(3):566–576.

Doney, R. A. and Kyprianou, A. E. (2006). Overshoots and undershoots of Lévy processes. *The Annals of Applied Probability*, pages 91–106.

Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, 20(3):393–403.

Durbin, J. (1968). The probability that the sample distribution function lies between two parallel straight lines. *The Annals of Mathematical Statistics*, 39(2):398–411.

Durbin, J. (1971). Boundary-crossing probabilities for the Brownian motion and Poisson processes and techniques for computing the power of the Kolmogorov-Smirnov test. *Journal of Applied Probability*, 8(3):431–453.

Durbin, J. (1973). Distribution theory for tests based on the sample distribution theory. *SIAM, Philadelphia*.

Eddelbuettel, D. and François, R. (2011). **Rcpp**: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Eder, I., Klüppelberg, C., et al. (2009). The first passage event for sums of dependent Lévy processes with applications to insurance risk. *The Annals of Applied Probability*, 19(6):2047–2079.

Embrechts, P., Kaufmann, R., and Samorodnitsky, G. (2004). *Ruin theory revisited: stochastic models for operational risk*, pages 243–261. European Central Bank, Frankfurt.

Epanechnikov, V. (1968). The significance level and power of the two-sided Kolmogorov test in the case of small sample sizes. *Theory of Probability and Its Applications*, 13(4):686–690.

Feigin, P. D. (1979). On the characterization of point processes with the order statistic property. *Journal of Applied Probability*, 16(2):297–304.

Feller, W. (1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions. *The Annals of Mathematical Statistics*, 19(2):177–189.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications: Volume I*, volume 1. John Wiley & Sons.

Fendick, K. W., Saksena, V. R., and Whitt, W. (1989). Dependence in packet queues. *IEEE Transactions on Communications*, 37(11):1173–1183.

Feng, L. and Linetsky, V. (2008). Pricing options in jump-diffusion models: an extrapolation approach. *Operations Research*, 56(2):304–325.

Fischer, W. and Meier-Hellstern, K. (1993). The markov-modulated Poisson process (MMPP) cookbook. *Performance evaluation*, 18(2):149–171.

Frigo, M. and Johnson, S. G. (2005). The design and implementation of fftw3. *Proceedings of the IEEE*, 93(2):216–231.

Frostig, E. (2004). Upper bounds on the expected time to ruin and on the expected recovery time. *Advances in Applied Probability*, 36(2):377–397.

Fu, J. C. and Wu, T. L. (2010). Linear and nonlinear boundary crossing probabilities for Brownian motion and related processes. *Journal of Applied Probability*, 47(4):1058–1071.

Fusai, G., Germano, G., and Marazzina, D. (2016). Spitzer identity, Wiener-Hopf factorization and pricing of discretely monitored exotic options. *European Journal of Operational Research*, 251(1):124–134.

Garrido, J. and Morales, M. (2006). On the expected discounted penalty function for Lévy risk processes. *North American Actuarial Journal*, 10(4):196–216.

Geman, H. and Yor, M. (1996). Pricing and hedging double-barrier options: a probabilistic approach. *Mathematical Finance*, 6(4):365–378.

Gerber, H. U. and Shiu, E. S. (1998). On the time value of ruin. *North American Actuarial Journal*, 2(1):48–72.

Ghobbar, A. A. and Friend, C. H. (2003). Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model. *Computers and Operations Research*, 30(14):2097–2114.

Gleser, L. J. (1985). Exact power of goodness-of-fit tests of Kolmogorov type for discontinuous distributions. *Journal of the American Statistical Association*, 80(392):954–958.

Goffard, P. O. and Lefèvre, C. (2017). Boundary crossing of order statistics point processes. *Journal of Mathematical Analysis and Applications*, 447(2):890–907.

Grandell, J. (1976). *Doubly Stochastic Poisson Processes. Lecture Notes in Mathematics: Vol. 529*. Springer, Berlin, Heidelberg.

Guillaume, T. (2010). Step double barrier options. *Journal of Derivatives*, 18(1):59–80.

Gurvich, I., Luedtke, J., and Tezcan, T. (2010). Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science*, 56(7):1093–1115.

Gutierrez, R. S., Solis, A. O., and Mukhopadhyay, S. (2008). Lumpy demand forecasting using neural networks. *International Journal of Production Economics*, 111(2):409–420.

Heyman, D. P. and Marshall, K. T. (1968). Bounds on the optimal operating policy for a class of single-server queues. *Operations Research*, 16(6):1138–1146.

Huzak, M., Perman, M., Šikić, H., Vondraček, Z., et al. (2004). Ruin probabilities and decompositions for general perturbed risk processes. *The Annals of Applied Probability*, 14(3):1378–1397.

IBM Corp. (2017). *IBM SPSS Statistics for Windows, Version 25.0*. IBM Corp., Armonk, NY.

Ignatov, Z. G. and Kaishev, V. K. (2000). Two-sided bounds for the finite time probability of ruin. *Scandinavian Actuarial Journal*, 2000(1):46–62.

Ignatov, Z. G. and Kaishev, V. K. (2004). A finite-time ruin probability formula for continuous claim severities. *Journal of Applied Probability*, 41(2):570–578.

Ignatov, Z. G. and Kaishev, V. K. (2016). First crossing time, overshoot and Appell–Hessenberg type functions. *Stochastics*, 88(8):1240–1260.

Jongbloed, G. and Koole, G. (2001). Managing uncertainty in call centres using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17(4):307–318.

Kaishev, V. K., Dimitrova, D. S., and Ignatov, Z. G. (2008). Operational risk and insurance: a ruin-probabilistic reserving approach. *The Journal of Operational Risk*, 3(3):39–60.

Kallenberg, O. (1997). *Foundations of Modern Probability*. Springer-Verlag, New York.

Kaplan, M. (1983). A single-server queue with cyclostationary arrivals and arithmetic service. *Operations Research*, 31(1):184–205.

Karr, A. (1991). *Point Processes and Their Statistical Inference. Second Edition. Series in Probability: Pure and Applied*. CRC/Marcel Dekker, New York.

Kerkhoff, U. and Lerche, H. R. (2013). Boundary crossing distributions of random walks related to the law of the iterated logarithm. *Statistica Sinica*, pages 1697–1715.

Khmaladze, E. and Shinjikashvili, E. (2001). Calculation of noncrossing probabilities for Poisson processes and its corollaries. *Advances in Applied Probability*, 33(3):702–716.

Kim, S.-H. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing & Service Operations Management*, 16(3):464–480.

Klüppelberg, C., Kyprianou, A. E., Maller, R. A., et al. (2004). Ruin probabilities and overshoots for general Lévy insurance risk processes. *The Annals of Applied Probability*, 14(4):1766–1801.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91.

Kou, S. G. and Wang, H. (2003). First passage times of a jump diffusion process. *Advances in applied probability*, 35(2):504–531.

Krämer, W., Ploberger, W., and Alt, R. (1988). Testing for structural change in dynamic models. *Econometrica*, 56(6):1355–1369.

Kremer, M., Moritz, B., and Siemsen, E. (2011). Demand forecasting behavior: system neglect and change detection. *Management Science*, 57(10):1827–1843.

Kunitomo, N. and Ikeda, M. (1992). Pricing options with curved boundaries. *Mathematical Finance*, 2(4):275–298.

Kwak, N. K., Garrett, W. A., and Barone, S. (1977). A stochastic model of demand forecasting for technical manpower planning. *Management Science*, 23(10):1089–1098.

Lando, D. (1998). On Cox processes and credit risky securities. *Review of Derivatives Research*, 2(2):99–120.

Last, G. and Brandt, A. (1995). *Marked Point Processes on the Real Line: The Dynamical Approach*. Springer-Verlag, New York.

Lefèvre, C. and Picard, P. (2011). A new look at the homogeneous risk model. *Insurance: Mathematics and Economics*, 49(3):512–519.

Lefèvre, C. and Picard, P. (2014). Ruin probabilities for risk models with ordered claim arrivals. *Methodology and Computing in Applied Probability*, 16(4):885–905.

Lehmann, A. (1998). *Boundary crossing probabilities of Poisson counting processes with general boundaries*, pages 153–166. Birkhäuser Boston, Boston, MA.

Lengu, D., Syntetos, A. A., and Babai, M. Z. (2014). Spare parts management: Linking distributional assumptions to demand classification. *European Journal of Operational Research*, 235(3):624–635.

Lippman, S. A. (1969). Optimal inventory policy with subadditive ordering costs and stochastic demands. *SIAM Journal on Applied Mathematics*, 17(3):543–559.

Lotov, V. I. (1996). On some boundary crossing problems for Gaussian random walks. *The Annals of Probability*, 24(4):2154–2171.

Lucantoni, D. M. (1991). New results on the single server queue with a batch Markovian arrival process. *Communications in Statistics. Stochastic Models*, 7(1):1–46.

Lyberopoulos, D. P. and Macheras, N. D. (2014). Some characterizations of mixed renewal processes. *Unpublished*.

Marsaglia, G., Tsang, W. W., and Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, 8(18):1–4.

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

Miller, L. H. (1956). Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association*, 51(273):111–121.

Moscovich, A. and Nadler, B. (2017). Fast calculation of boundary crossing probabilities for Poisson processes. *Statistics & Probability Letters*, 123:177–182.

Nawrotzki, K. (1962). Ein grenzwertsatz für homogene zufällige punktfolgen. *Mathematische Nachrichten*, 24(4):201–217.

Niederhausen, H. (1981). Sheffer polynomials for computing exact Kolmogorov-Smirnov and Rényi type distributions. *The Annals of Statistics*, 9(5):923–944.

Niu, S.-C. (1980). A single server queueing loss model with heterogeneous arrival and service. *Operations Research*, 28(3-part-i):584–593.

Noé, M. (1972). The calculation of distributions of two-sided Kolmogorov-Smirnov type statistics. *The Annals of Mathematical Statistics*, 43(1):58–64.

Noether, G. E. (1963). Note on the Kolmogorov statistic in the discrete case. *Metrika*, 7(1):115–116.

Novikov, A., Frishling, V., and Kordzakhia, N. (1999). Approximations of boundary crossing probabilities for a Brownian motion. *Journal of Applied Probability*, 36(4):1019–1030.

Oreshkin, B. N., Réegnard, N., and L'Ecuyer, P. (2016). Rate-based daily arrival process models with application to call centers. *Operations Research*, 64(2):510–527.

Ospina, R. and Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, 51(1):111–126.

Pan, Y. and Borovkov, K. A. (2019). The exact asymptotics of the large deviation probabilities in the multivariate boundary crossing problem. *Advances in Applied Probability*, 51(3):835–864.

Panjer, H. H. (2006). *Operational risk: modeling analytics*, volume 620. John Wiley & Sons.

Pelsser, A. (2000). Pricing double barrier options using Laplace transforms. *Finance and Stochastics*, 4(1):95–104.

Pelz, W. and Good, I. (1976). Approximating the lower tail-areas of the Kolmogorov-Smirnov one-sample statistic. *Journal of the Royal Statistical Society B*, 38(2):152–156.

Peskir, G. et al. (2008). The law of the hitting times to points by a stable Lévy process with no negative jumps. *Electronic Communications in Probability*, 13:653–659.

Pettitt, A. N. and Stephens, M. A. (1977). The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, 19(2):205–210.

Pomeranz, J. (1974). Algorithm 487: Exact cumulative distribution of the Kolmorgorov-Smirnov statistic for small samples. *Communications of the ACM*, 17(12):703–704.

Posner, M. (1973). Single-server queues with service time dependent on waiting time. *Operations Research*, 21(2):610–616.

Pötzelberger, K. and Wang, L. (2001). Boundary crossing probability for Brownian motion. *Journal of Applied Probability*, 38(1):152–164.

Prabhu, N. U. (1961). On the ruin problem of collective risk theory. *The Annals of Mathematical Statistics*, 32(3):757–764.

Prabhu, N. U. (1980). *Stochastic storage processes: queues, insurance risk and dams*. Springer Verlag.

Presman, E. and Sethi, S. P. (2006). Inventory models with continuous and Poisson demands and discounted and average costs. *Production and Operations Management*, 15(2):279–293.

Puri, P. S. (1982). On the characterization of point processes with the order statistic property without the moment condition. *Journal of Applied Probability*, 19(1):39–51.

Raaijmakers, Y., Albrecher, H., and Boxma, O. (2018). The single server queue with mixing dependencies. *Methodology and Computing in Applied Probability*, pages 1–22.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Resnick, S. and Samorodnitsky, G. (1997). Performance decay in a single server exponential queueing model with long range dependence. *Operations Research*, 45(2):235–243.

Rolski, T. (1986). Upper bounds for single server queues with doubly stochastic Poisson arrivals. *Mathematics of Operations Research*, 11(3):442–450.

Ross, S. M. (1978). Average delay in queues with non-stationary Poisson arrivals. *Journal of Applied Probability*, 15(3):602–609.

Ruben, H. and Gambino, J. (1982). The exact distribution of Kolmogorov's statistic D_n for $n \leq 10$. *Annals of the Institute of Statistical Mathematics*, 34(1):167–173.

Savov, M. (2009). Small time two-sided LIL behavior for Lévy processes at zero. *Probability theory and related fields*, 144(1-2):79–98.

Scarf, H. E. (1963). *A survey of analytic techniques in inventory theory*, pages 185–225. Stanford University Press.

Schmid, P. (1958). On the Kolmogorov and Smirnov limit theorems for discontinuous distribution functions. *The Annals of Mathematical Statistics*, 29(4):1011–1027.

Serfozo, R. F. (1972). Processes with conditional stationary independent increments. *Journal of Applied Probability*, 9(2):303–315.

Shi, J., Katehakis, M. N., Melamed, B., and Xia, Y. (2014). Production-inventory systems with lost sales and compound Poisson demands. *Operations Research*, 62(5):1048–1063.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. John Wiley, New York.

Siegmund, D. (1986). Boundary crossing probabilities and statistical applications. *The Annals of Statistics*, 14(2):361–404.

Simard, R. and L'Ecuyer, P. (2011). Computing the two-sided Kolmogorov-Smirnov distribution. *Journal of Statistical Software*, 39(11):1–18.

Slakter, M. J. (1965). A comparison of the Pearson chi-square and Kolmogorov goodness-of-fit tests with respect to validity. *Journal of the American Statistical Association*, 60(311):854–858.

Smirnov, N. (1939). Sur les écarts de la courbe de distribution empirique. *Matematicheskii Sbornik*, 48(1):3–26.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281.

Song, J. S. and Zipkin, P. H. (1996). The joint effect of leadtime variance and lot size in a parallel processing environment. *Management Science*, 42(9):1352–1363.

StataCorp. (2017). *Stata Statistical Software: Release 15.* StataCorp LLC., College Station, TX.

Steck, G. P. (1971). Rectangle probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions. *The Annals of Mathematical Statistics*, 42(1):1–11.

Stenius, O., Karaarslan, A. G., Marklund, J., and de Kok, A. G. (2016). Exact analysis of divergent inventory systems with time-based shipment consolidation and compound Poisson demand. *Operations Research*, 64(4):906–921.

Stidham Jr, S. (1970). On the optimality of single-server queuing systems. *Operations Research*, 18(4):708–732.

Teunen, M. and Goovaerts, M. (1994). Double boundary crossing result for the brownian motion. *Scandinavian Actuarial Journal*, 1994(2):139–150.

Teunter, R. H. and Duncan, L. (2009). Forecasting intermittent demand: a comparative study. *Journal of the Operational Research Society*, 60(3):321–329.

Teunter, R. H., Syntetos, A. A., and Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3):606–615.

The MathWorks Inc. (2018). *MATLAB – The Language of Technical Computing, Version R2018a.* The MathWorks Inc., Natick, Massachusetts.

Van Mieghem, J. A. (1995). Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals of Applied Probability*, pages 809–833.

Wald, A. and Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. *The Annals of Mathematical Statistics*, 10(2):105–118.

Walsh, J. E. (1963). Bounded probability properties of Kolmogorov-Smirnov and similar statistics for discrete data. *Annals of the Institute of Statistical Mathematics*, 15(1):153–158.

Wang, L. and Pötzelberger, K. (2007). Crossing probabilities for diffusion processes with piecewise continuous boundaries. *Methodology and Computing in Applied Probability*, 9(1):21–40.

Wang, X., Andradóttir, S., and Ayhan, H. (2019). Optimal pricing for tandem queues with finite buffers. *Queueing Systems*, pages 1–74.

Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management*, 15(1):88–102.

Whitt, W. (2018). Time-varying queues. *Queueing Models and Service Management*, 1(2):79–164.

Whitt, W. and You, W. (2018). Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research*, 66(1):184–199.

Whitt, W. and You, W. (2019). Time-varying robust queueing. *Operations Research, forthcoming*.

Willemain, T. R., Ratti, E. W. L., and Smart, C. N. (1994a). Forecasting intermittent demand using a Cox process model. In *INFORMS Meetings*, pages 1–14, Boston, USA.

Willemain, T. R., Smart, C. N., and Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting*, 20(3):375–387.

Willemain, T. R., Smart, C. N., Shockor, J. H., and DeSautels, P. A. (1994b). Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. *International Journal of Forecasting*, 10(4):529–538.

Willmot, G. E. (2015). On a partial integrodifferential equation of Seal's type. *Insurance: Mathematics and Economics*, 62:54–61.

Wolfram Research, Inc. (2018). *Mathematica, Version 11.3*. Wolfram Research, Inc., Champaign, Illinois.

Wood, C. L. and Altavela, M. M. (1978). Large-sample results for Kolmogorov–Smirnov statistics for discrete distributions. *Biometrika*, 65(1):235–239.

Xu, Y., Scheller-Wolf, A., and Sycara, K. (2015). The benefit of introducing variability in single-server queues with application to quality-based service domains. *Operations Research*, 63(1):233–246.

Yang, H. and Zhang, L. (2001). Spectrally negative Lévy processes with applications in risk theory. *Advances in Applied Probability*, 33(1):281–291.

Ycart, B. and Drouilhet, R. (2016). Computing wedge probabilities. Unpublished Results.

Zan, J., Hasenbein, J. J., and Morton, D. P. (2014). Asymptotically optimal staffing of service systems with joint QoS constraints. *Queueing Systems*, 78(4):359–386.

Zheng, Y. S. (1992). On properties of stochastic inventory systems. *Management Science*, 38(1):87–103.

Zolotarev, V. (1964). The first passage time of a level and the behavior at infinity for a class of processes with independent increments. *Theory of Probability & Its Applications*, 9(4):653–662.